

Profiling, Modelling and Facilitating Online Activism

THÈSE N° 7925 (2017)

PRÉSENTÉE LE 22 SEPTEMBRE 2017
À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE DE SYSTÈMES D'INFORMATION RÉPARTIS
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Julia PROSKURNIA

acceptée sur proposition du jury:

Prof. P. Dillenbourg, président du jury
Prof. K. Aberer, directeur de thèse
Prof. Ph. Cudré-Mauroux, rapporteur
Dr C. Castillo, rapporteur
Prof. R. West, rapporteur



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2017

Better later than never. — Vera Proskurnia

To my grandmother — Vera Proskurnia — for setting the bar high and believing in me...

Abstract

The extensive and successful use of *social media* enhanced and empowered a variety of movements all over the world in a way that would have been hard to achieve through conventional means. This led to numerous studies which leverage online social data to describe, analyze, model or gain insights about online activism, as well as to empower and facilitate information sharing for activists.

Despite the recent success of public movements thanks to the efficiency that online tools brought to the activists, it is still unclear what factors impact success. In addition, the support to empower online activism remains scarce. This thesis investigates the aforementioned issues in respect to, first, analysis and modeling of online activism in several forms, second, algorithmic approaches to producing unstructured texts for individual users and filtering social media streams. To reflect on these issues, we conduct separate studies which enable us to apply consistent methodologies to profile online campaigns, devise systematically interpretable models for online petitions, we propose and assess efficient template induction tools for email composition, and design and compare an efficient and accurate approach for filtering topical short texts.

The main contributions of this thesis are:

- Gained insights into *online public campaigns*. We are comparing over a hundred awareness and mobilization public campaigns on social media regarding online and offline actions that were performed by activists. To this end, we introduced a generic methodology for categorizing online campaigns based on their goals and user engagement, as well as extracted campaigns' actions from their social media traces. We discovered substantial differences between types of campaigns and their corresponding actions.
- Scrutinized and quantified *the effect of external sources* (social media, platform's front page coverage) *on online reinforced phenomena* – over 4,000 e-petitions. We proposed an accurate and interpretable model that dissects the impact of various confounders on the time evolution of petitions' signatures. We showed performance variations of the designed model with various combinations of external factors that outperform not only multiple baselines, but also is interpretable across various petitions. Our findings suggest that external influences shape the popularity of online petitions differently, i.e., effects

Abstract

of social media are prolonged and are stronger for successful petitions, while the direct promotion is the strongest compared to other factors in the absolute terms.

- Assessed *the extent of repetitive content* in targeted email messages. We defined a task of template induction over unstructured email corpora and proposed an efficient and accurate algorithm that, first, identifies repetitive and representative phrases that are usually typed by a user, and second, aligns these phrases into a template. While we found over 1% of email users might benefit from the templatization, we also uncovered the potential of saving up to several dozens of words in email writing effort.
- *Improved document filtering for a particular topic or event* by introducing a method that increases both (1) the accuracy of filtered short texts samples while preserving efficiency and (2) recall for relatively small input training sets. To locate and monitor particular topics or events, this method constructs and applies a filter of automatically generated lists of patterns that represent semantically homogeneous groups of input seed messages.

The overarching goals of this thesis is to contribute a methodological and practical body of research that aims to analyze online activism on social media, to model the popularity and spread of an online reinforced phenomena in an interpretable fashion, and to facilitate collective actions on the web.

Keywords: Collective Actions on Social Media, Online Public Campaigns, Online Petitions Modeling, Template Induction, Semantic Filtering of Text, Pattern Mining.

Résumé

L'utilisation répandue *des médias sociaux* a renforcé et stimulé une variété de mouvements dans le monde entier d'une manière qu'il était impossible d'imaginer auparavant. De nombreuses études tirent parti des données des réseaux sociaux numériques pour décrire, analyser, modéliser ou acquérir des connaissances sur l'activisme en ligne, ainsi que pour renforcer et faciliter le partage d'informations pour les militants.

Malgré le récent succès des mouvements publics en raison de l'efficacité que les outils en ligne apportent aux militants, les facteurs qui influent sur le succès et le soutien de ces mouvements ou sur le dynamisme de l'activisme en ligne sont encore peu étudiés. Cette thèse se penche sur les problèmes susmentionnés et comprend, d'abord, l'analyse et la modélisation de l'activisme en ligne sous plusieurs formes, deuxièmement, des approches algorithmiques pour produire des textes non structurés pour chaque utilisateur et pour filtrer les flux de médias sociaux. Pour étudier ces questions, Nous menons des travaux distincts qui nous permettent d'appliquer des méthodologies cohérentes pour la définition de profil des campagnes en ligne, et de concevoir des modèles systématiquement interprétables pour les pétitions. Nous proposons et évaluons des outils efficaces d'induction de modèles pour la rédaction d'emails, et nous présentons et comparons une approche efficace et précise pour filtrer des textes courts.

Les principales contributions de cette thèse sont :

- *Obtenir des connaissances sur les campagnes publicitaires en ligne.* Nous avons mené une étude comparative qui compare une centaine de campagnes de sensibilisation et de mobilisation sur les réseaux sociaux, concernant des actions en ligne et hors ligne réalisées par les militants. À cette fin, nous avons mis en place une méthodologie générique pour catégoriser les campagnes en ligne en fonction de leurs objectifs, de l'engagement des utilisateurs, ainsi que des actions menées sur les réseaux sociaux en rapport avec cette campagne. Nous avons découvert des corrélations substantielles entre les types de campagnes et leurs actions correspondantes.
- *Examiner et quantifier l'impacte des sources externes, c'est-à-dire les médias sociaux, et de la page d'accueil de la plate-forme, sur les phénomènes numériques renforcés* (plus de 4 000 pétitions virtuelles). Nous avons proposé un modèle précis et interprétable qui examine

Résumé

divers facteurs influençant l'évolution du nombre de signatures au cours du temps. Nous avons étudié les performances du modèle en fonction de diverses combinaisons de facteurs externes. Ces variations surpassent les méthodes de références pour ce type de modèle, mais ce modèle propose également une interprétation des pétitions sans précédent. Nos résultats ont suggéré que les facteurs extérieures façonnent différemment la popularité des pétitions : Les effets des médias sociaux se prolongent et sont plus forts pour les pétitions réussies, tandis que la promotion directe est la plus forte.

- *Evaluer le contenu répétitif dans les messages électroniques ciblés.* Nous avons défini une génération de modèles basée sur des corpus d'*emails* non structuré et nous avons proposé un algorithme efficace et précis qui, d'abord, identifie des phrases répétitives écrit par un utilisateur et, deuxièmement, associe ces phrases à un modèle. Bien que nous ayons constaté que plus de 1% des utilisateurs de courrier électronique pourraient bénéficier de la modélisation, nous avons également découvert le potentiel d'économiser jusqu'à plusieurs dizaines de mots par courrier électronique, ce qui est à son tour un atout essentiel pour les militants.
- *Amélioration du filtrage de documents pour un sujet ou un événement particulier* en introduisant une méthode qui augmente à la fois (1) la précision des échantillons de textes courts, tout en préservant l'efficacité d'exécution et (2) le rappel pour des training sets de petits tailles. Pour localiser et surveiller des sujets ou des événements particuliers, cette méthode construit et applique un filtre de listes de motifs générés automatiquement qui représentent des groupes sémantiquement homogènes.

Les objectifs généraux de cette thèse sont de contribuer à une étude méthodologique et pratique qui vise à analyser l'activisme en ligne sur les réseaux sociaux. Cette analyse permet de modéliser la popularité et la diffusion des phénomènes numériques d'une manière interprétative. Elle permet également de faciliter les actions collectives sur le web.

Mots clés : Actions collectives sur les réseaux sociaux, campagnes publiques en ligne, modélisation d'e-pétitions, génération de template, filtrage sémantique de texte, découverte de motifs de répétition.

Acknowledgements

My first round of gratitude goes to my supervisor Karl Aberer, who gave me enough freedom to be creative, independent, and confident in my research! I would also like to thank my highly-esteemed committee members: Carlos Castillo, Philippe Cudré-Mauroux, Robert West, Pierre Dillenbourg for the time and effort in reviewing on my thesis, as well as their valuable and constructive feedback!

Second, I am so grateful and happy to be a friend of three very important people in my PhD - Guillaume Jean Op 't Veld a.k.a. Giel¹, Adrian Seredinschi a.k.a. A-a-a-dzi, and Victor Ma a.k.a. Victor :). These guys were there all the time except for Adi of course, and Victor since he left us tooooo soon. You are amazing people! Furthermore, I am infinitely grateful to Alevtina and Olga for making me feel smarter than I am and filling all those hours of unproductive work with efficient and pleasant discussion and rehearsals! Special thanks for Panagiotis and Jean-Eudes for proving that fun people also do a PhD, and being there even at lunch, midnight, or later for the rehearsals of my talks!

On a serious note, I am in a great debt to two major figures in my PhD life who made me actually believe in myself and gave me great support and mentorship! Thank you Philippe Cudré-Mauroux for taking me under your wing a few years ago and showing me what a PhD is about and how to do research. Any 1:1 that we had resulted in at least 2 weeks of highly motivated and productive work! You are amazing! Great thanks to Carlos Castillo for listening to me, making me believe in my skills, ability, and research! You showed me the “nobel” and honest side of the research community and taught me a lot!

Of course, I am very happy to be part of my lab and I want to thank all my lab mates and associates: Alexandra, Berker, Julien, Tian, Nataliya, Thanh Tam, Rameez, Matteo, Hao, Hamza, Remi, Jean-Eudes, Panagiotis, Amit, Mehdi, Jean-Paul, Martin, Alex, Michele, Tri-Kurniawan, Hung! Chantal François deserves a separate mention for her endless administrative, logistic support, and also for being always there, always positive and ready to listen to my rather funny stories. I would also like to thank my EPFL friends for coffee breaks, discussions, complain

¹Giel is proven to be a better spell checker than Microsoft Word or Grammarly. True story!

Acknowledgements

sessions and rejection-story sharing. I thank Renata, Artem, Ksenia, Amer, Damian, Katarina, Matej, Catherine, Onur, Valentina, Lyudmila, Jonas, Judith, Florence, Camille, and many others!

A great round of appreciation goes to the Exascale Lab for their support and great lunch discussions in Fribourg: Roman, Ruslan, Michael, Alberto, Djellel, Mourad, Dingqi, Artem, Alisa, Laura, Paolo, Victor - thanks a lot!!!

During my PhD, I have made about a hundred trips to Zurich and some outside of Switzerland. I am glad to have met so many people that contributed to the discussions about the PhD, playing piano, skiing, travelling, bouldering, etc. I thank Zurich associates Andrius, Alexej, Luca and Maria, Werner and his girls, Christian and Elena, Noëmi, Merve, as well as Kiev associates Ira, Oksana, Nina, Bogdan, and many others! In particular, I would like to acknowledge the best team in Google Zurich! I felt trully honoured to work with all of you during the 6 month of my internship: Ivo, Balint, Tobias, Laszlo, Tom, Karol, Marina, Felix, Alex, Lluís, James, Marc-Allen!

I enjoyed every conference I attended and I was happy to meet many great people. Thanks a lot for the inspiration - Manuel Gomez Rodriguez, Muhammad Anis Uddin Nasir, Luca Maria Aiello, Ricardo Baeza-Yates, Claudia Wagner, Cody Buntain, Daniel Gatica-Perez, Yelena Mejova, Gianluca Demartini, Gianmarco De Francisci Morales, Pablo Aragón, Bogdan State, Ryota Kobayashi, Przemyslaw A. Grabowicz, and many more! Honourable mention: Mikey, Chandan, Denis for the calmness they showed enduring my driving in Western Australia (on the left side).

During the last few years of my PhD I did a lot of sports. I have particularly warm memories of the best pole dance schools ever - Pink Attitude and Pole Emotion. My biggest thanks go to Aude, Aina, Lionelle, Frédérique, Virginie, Morgan, Gaëlle, Luca, Naomi, Kristina, Sonia, Laura, Noémi, Christina, Christelle, and many others. Aude - you are so positive and motivating; without your trust and encouragement I am sure I would not have achieved as much as I did!

Clearly, all of this would not have been possible were it not for the support of my family! My dearest parents and grandparents I love you so much and I am very proud of you! Thanks a lot for all your support, stress, endless love, and trust in me! I also would like to thank Elvyra for our nice conversations, understanding and believing in me!

Žygi! Aš tave labai myliu! Tu tai žinai! Thanks a lot for being there all the time! You saw all the shades of the PhD life and managed to handle it! You gave me the chance to improve a huge range of skills: explaining material, listening to opposing opinions, working independently, configuring settings, playing PS, etc. Now I know how to do risotto and scrambled eggs. Such a great achievement!

Lausanne, 2017

Dr. Julia Pro.

Contents

| | |
|--|-----------|
| Abstract (English/Français) | i |
| Acknowledgements | v |
| List of figures | xi |
| List of tables | xv |
| 1 Introduction | 1 |
| 1.1 Research Problems and Contributions | 3 |
| 1.2 Thesis Outline | 8 |
| Background | 11 |
| 2 Online Activism on Social Media and Beyond | 13 |
| 2.1 Digital collective actions and its analysis | 14 |
| 2.2 On modelling and predicting the popularity of online content | 18 |
| 2.3 Facilitation of collective actions | 20 |
| 2.4 Positioning | 26 |
| Profiling | 29 |
| 3 Profiling Large-Scale Public Campaigns on Twitter | 31 |
| 3.1 Introduction | 31 |
| 3.2 Related Work | 33 |
| 3.3 Data Collection and Cleansing | 36 |
| 3.3.1 Twitter data collection | 36 |
| 3.3.2 Unique tweets identification and retweets count | 38 |
| 3.3.3 URL usage statistics | 38 |
| 3.4 Campaign analysis | 38 |
| 3.4.1 Types of campaigns | 39 |
| 3.4.2 User engagement patterns | 39 |
| 3.5 Tweet Type Identification and Classification | 44 |

Contents

| | | |
|----------|---|-----------|
| 3.5.1 | Types of tweets | 44 |
| 3.5.2 | Data analysis | 48 |
| 3.5.3 | Retweets | 51 |
| 3.6 | Discussion | 52 |
| 3.7 | Conclusions | 53 |
| 4 | Online Environmental Petitions in Public Campaigns | 55 |
| 4.1 | Introduction | 55 |
| 4.2 | Data Collection, Cleansing and Insights | 57 |
| 4.3 | Petition Analysis | 58 |
| 4.3.1 | Petitions and tweets stats | 58 |
| 4.3.2 | Petitions in public campaigns on Twitter | 59 |
| 4.3.3 | Campaigns' petitions on Twitter | 60 |
| 4.4 | Conclusions | 61 |
| | Modelling | 63 |
| 5 | Predicting the Success of Online Petitions | 65 |
| 5.1 | Introduction | 65 |
| 5.2 | Related Work | 67 |
| 5.2.1 | Popularity prediction on the web | 67 |
| 5.2.2 | Analyzing the dynamics of online petitions | 67 |
| 5.3 | Data Collection and Insights | 69 |
| 5.3.1 | Data collection | 69 |
| 5.3.2 | Data insights | 71 |
| 5.3.3 | Circadian cycles and external influence | 77 |
| 5.3.4 | Matching twitter users and signers | 78 |
| 5.3.5 | Front page effect | 80 |
| 5.4 | Petitions Modelling | 80 |
| 5.4.1 | Circadian rhythm and aging | 81 |
| 5.4.2 | Self-excitation and external influence | 81 |
| 5.5 | Experimental results | 82 |
| 5.5.1 | Metrics | 82 |
| 5.5.2 | Baselines | 83 |
| 5.5.3 | Prediction | 83 |
| 5.5.4 | Analysis of estimated parameters | 85 |
| 5.6 | Conclusions | 87 |

| | |
|---|------------|
| Facilitating | 89 |
| 6 Efficient Document Filtering Using Vector Space Topic Expansion and Pattern-Mining: <i>The Case of Event Detection in Microposts</i> | 91 |
| 6.1 Introduction | 91 |
| 6.2 Related Work | 93 |
| 6.3 Data Collection and Seed Extraction | 93 |
| 6.3.1 Data collection | 94 |
| 6.3.2 Seed extraction | 94 |
| 6.4 Method description | 97 |
| 6.4.1 Overview | 97 |
| 6.4.2 Text similarity metric | 99 |
| 6.4.3 Pattern extraction | 101 |
| 6.4.4 Patterns vs clustering: a case of coverage | 103 |
| 6.5 Experimental Results | 103 |
| 6.5.1 Baselines | 104 |
| 6.5.2 Metrics and their estimation | 105 |
| 6.5.3 Results | 106 |
| 6.5.4 Discussion | 106 |
| 6.6 Conclusions | 109 |
| 7 Template Induction over Unstructured Email Corpora | 111 |
| 7.1 Introduction | 112 |
| 7.2 Related Work | 114 |
| 7.2.1 Email content mining | 114 |
| 7.2.2 Template induction | 115 |
| 7.2.3 Autocompletion | 116 |
| 7.2.4 Assisted email composition | 116 |
| 7.3 Methodology | 117 |
| 7.3.1 Preprocessing | 117 |
| 7.3.2 Clustering | 117 |
| 7.3.3 Baseline phrase extraction | 119 |
| 7.3.4 Suffix array based approach | 121 |
| 7.3.5 Constructing a template from phrases | 123 |
| 7.4 Experiments | 124 |
| 7.4.1 Synthetic corpus | 125 |
| 7.4.2 Enron corpus | 126 |
| 7.4.3 Scalability analysis | 130 |
| 7.5 Conclusions and Future Work | 131 |
| 8 Conclusions | 133 |
| Conclusions | 133 |

Contents

| | | |
|-----|--------------------------------------|------------|
| 8.1 | Conclusions and Discussion | 133 |
| 8.2 | Future Work | 135 |
| | Bibliography | 139 |

List of Figures

| | | |
|-----|--|----|
| 1.1 | Overview of the structure and the conceptual flow of this thesis. | 9 |
| 3.1 | Data collection pipeline for the profiling of the public campaigns. | 36 |
| 3.2 | Top-20 domains used in the climate change campaigns for the original tweets. . . | 37 |
| 3.3 | Different user engagement patterns observed in campaigns. | 41 |
| 3.4 | Distribution of user engagement patterns for the different types of campaigns. . . | 42 |
| 3.5 | Example of the active and inactive kernel involvement. #standforforest has gini 0.3, while #leardblockade - 0.80. | 43 |
| 3.6 | Comparison of top domain name usage across campaigns | 46 |
| 3.7 | Comparison of the distributions of actions for the types of climate change cam- paigns defined in Section 3.4.1. Dark indicate greater importance of a particular actions in the campaign type. | 49 |
| 3.8 | Comparison of the distributions of actions for the two main categories of cli- mate change campaigns: awareness and mobilization. Dark indicates a greater importance of a particular actions in the campaign type. | 50 |
| 3.9 | Comparison of duplicate content between campaigns | 51 |
| 4.1 | The final number of signatures received by each petition. The red line indicates the required number of signatures. A change in the slope of the zipf distribution occurs at 1K signatures, which represents a threshold for a petition to make a potential impact. | 59 |
| 4.2 | <i>SignatureRate</i> against number of unique users posting about a petition on Twitter. | 61 |
| 5.1 | Collection pipeline for the petitions dataset. | 70 |
| 5.2 | Cumulative Distribution Function (CDF) of the signatures collected by successful and failed petitions during their entire history (left) and during their first three hours (right). | 73 |
| 5.3 | Average of normalized Cumulative Distribution Function (CDF) in four clusters of petitions. Clustering was performed using Dynamic Time Warping (DTW). Numbers in parenthesis represent the size of each cluster. | 74 |
| 5.4 | Petition signature daily cumulative distribution function. Straight line corre- sponds to the signature goal set by the owner. First two distributions belong to the failed petitions while the last two belong to the successful ones. | 75 |

List of Figures

5.5 Average number of signatures per petition that are contributed by the top active countries. 76

5.6 The number of the distinct petitions (left) and distinct petition subcategories (right) signed by the users, e.g., (left) 101 users (x axes) signed 324 petitions (y axes); (right) 136 users signed petitions from all environmental categories. A change in the slope of the distribution occurs for the users with less than 300 petitions signed which correspond to the average number of petitions signed by the owners. 76

5.7 Top petition creators are USA, Great Britain and Canada. On the x axes there are country codes of the petitions owners. Left y axes (bars) shows the average rank of the owners country among top 10 countries that contribute to the petitions. Right y axes (scatter) shows absolute number of petitions owners from a given country (country code). 77

5.8 Daily pattern of the signature (left) and tweet activity (right) with 10 minutes time intervals. Both activities can be fitted by using a sinusoidal function: $a + b\sin(2\pi(t + t_0)/24)$ 78

5.9 Error in the prediction of signatures after 3 days (top) and 7 days (bottom), in terms of SMAPE (left) and cumulative SMAPE (right). For each timestamp t_s (x axis) a predictor was trained using $\{s^p(i)\}, i < t_s$ for all petitions p in the data set. The shaded area depicts the 20th and 80th percentile of the performance of the best model (CRD and social media). 82

5.10 Hourly average social media exposure for a petition during its first week. CRD model has the following parameters: $a = 102.14, b = 16.67, \phi = 8.90, k = 0.25, \tau = 37.86$ 84

5.11 Example showing the prediction of the number of signatures of one petition, after 3 days of observation, by a sample of methods including the best performing ones. 85

5.12 Distributions of parameter estimations for failed petitions (top) and successful petitions (middle). We also consider separately petitions promoted on the front page, all of them successful (bottom). Details about each parameter are provided in Section 5.4.1. 86

5.13 Influence function estimation for self-excitation, the influence of social media, and the front-page effect. A value of i on the X axis refers to the median influence of this aspect i hours in the past. On each plot, the Y axis presents an absolute scale for successful and failed petitions and shows the multiplicative effect on the number of signatures. Petitions promoted on the home page are all successful. 86

6.1 Pipeline overview. 98

6.2 Relatedness of documents as function of their pairwise distance. 101

6.3 Estimated part of topically-related document pairs that can be covered by patterns. The graph is generated for the training size of 5,000 examples. 103

7.1 An overview of the clustering and template induction. 117

7.2 Average template coverage and entropy for two synthetically generated corpuses, i.e., equally and normally distributed cluster sizes. We present the coverage by varying clustering threshold θ and fixed phrase frequency γ . Suffix Array based approaches consistently show higher coverage (portion of the email marked as fixed) with lower Template Entropy (number of fixed-phrases per template). . . . 126

7.3 The average ED as cluster size increases for $\theta = 0.8$ 128

7.4 Correlation between average body edit distance and average subject edit distance within cluster and fraction of distinct receivers (1 - all receivers are distinct) respectively. 129

7.5 Average template coverage and entropy for $\gamma = 0.8$ and $\theta = 0.8$ over Enron corpus all and high quality clusters. 130

7.6 A comparison of the increase in extraction size as the average cluster size increases. 131

List of Tables

| | | |
|-----|---|----|
| 2.1 | Various analysis of <i>digital activism</i> , e.g., campaigns, e-petitions, protests and movements. | 15 |
| 2.2 | Activism profiling through social network analytics, communities and influencers detection as well as pitfalls and biases of social media analysis. | 17 |
| 2.3 | Overview of the most prominent approaches and applications of the popularity predictions on social media. | 19 |
| 2.4 | Overview of the most prominent approaches and applications to facilitate content generation. In particular, we focus on template induction, email mining and auto-completion. | 21 |
| 2.5 | Overview of the most prominent approaches for document filtering and summarization. | 23 |
| 2.6 | Overview of the most prominent approaches and applications of event detection, summarization and tracking. | 25 |
| 3.1 | The list of the campaigns annotated by both their goal and user engagement pattern. “G” corresponds to the categorization of campaigns by their goal (awareness or mobilization). “U” stands for the categorization of campaigns by the user engagement pattern. Gini - corresponds to the gini coefficient computed as explained in Section 3.4.2 Clarifications: <i>a</i> - awareness, <i>m</i> - mobilization, <i>succ</i> - ever-growing, <i>multi</i> - multi-burst, <i>non</i> - inactive, <i>r</i> - annual, <i>one</i> - one-day campaign respectively. | 40 |
| 3.2 | Gini coefficients. Lower values correspond to almost equal user contribution, higher values represent campaigns where only a small fraction of users contribute. | 43 |
| 3.3 | Sample tweets for each type of action considered. | 45 |
| 3.4 | Examples of rules and number of tweets for each type of action. | 46 |
| 3.5 | Precision, Recall and F1-score values for classification of different types of actions with different sets of features. | 47 |
| 4.1 | Global statistics of the petition dataset of environmental campaigns. We show the data for the successful and failed petitions, as well as total numbers. Users are unique individuals who tweeted the petition URLs at least once. $S(p)$ and $C(p)$ for successful and failed petitions are highlighted in the table. Additionally, we show statistics of the petition tweets that do not have a campaign hashtag. | 59 |

List of Tables

| | | |
|-----|--|-----|
| 5.1 | Selected works on popularity predictions in social media. Typical tasks in this context are to classify as successful/unsuccessful (top), to predict the overall popularity (middle), and to forecast the popularity time series (bottom). | 68 |
| 5.2 | Number of petitions, main signature goal and mean number of collected signatures (in thousands) that are directed to a particular category of the petition target. | 71 |
| 5.3 | Petition categories labeled by <i>thepetitionsite.com</i> | 72 |
| 5.4 | Petition categories' success rates. | 72 |
| 5.5 | Dataset characteristics. Each characteristic in this table shows a significant difference at $p \ll 0.001$ | 72 |
| 5.6 | Characteristics of the user profiles that were unambiguously matched between <i>The Petition Site</i> and Twitter. | 79 |
| 5.7 | Comparison of petitions that were promoted to the front page (FP) against similar petitions that were not promoted (\neg FP). A significant difference at $p < 0.01$ is denoted by **. | 80 |
| 6.1 | Wikipedia dataset characteristics. We list the top 25 countries and the top 4 attack types as described on Wikipedia. The column "Event" corresponds to the total number of events for a country, while "Tweets" contains the number of matching microposts for each attack description. | 95 |
| 6.2 | Examples of seed microposts related to the attacks. | 96 |
| 6.3 | Examples of extracted synsets. | 102 |
| 6.4 | Top features extracted by PMI using unigrams only (top) or unigrams and bigrams (bottom). Results are obtained taking the entire training data for all events (terrorist attacks) as the positive class. | 105 |
| 6.5 | Evaluation results for the micropost extraction task of the four baseline methods against our method. The average size of a synset pattern was 204, 373, 465, 439, 451, 462 attributes for 100 - 10,000 training examples respectively. | 107 |
| 6.6 | Examples of patterns and associated documents generated by our approach. Mean WDM refers to mean pair-wise WDM document distance. Pattern is presented as a combination of stemmed words. | 108 |
| 7.1 | LCP, SA and actual suffixes ordered lexicographically. | 123 |
| 7.2 | Examples of the templates created based on SA. _____ correspond to the non-fixed regions of the templates. | 124 |
| 7.3 | Average characteristics of the generated synthetic corpus. U and N are uniform and normal of the cluster sizes. | 125 |
| 7.4 | Description of the Enron sent mail corpus. | 127 |

1 Introduction

“... there is no reason to believe street protests necessarily have more power than online acts... Besides, most street protesters today organize with digital tools, and publicize their efforts on social media.”

– Prof. Zeynep Tufekci, 2017, quoted from “Twitter and Tear Gas” p. 131 [291].

Digital dualism, the separation between “the real world” and “cyber space”, disperse at an incredible speed. Yet, it is still sometimes mistakenly imposed on collective actions without understanding and considering the mechanisms by which such movements operate [28, 127, 291]. The specifics of digital tools and technologies, their spectrum of features, affordances and limitations, and the way layers of influence interact and intermix matters to understand “networked movements and protests” [291]. Favourably, digitalization and connectivity opened unprecedented opportunities to provide insights and to answer relevant questions about public campaigns, digitally organized revolutions, uprising, and movements all over the world, among many other collective actions [81, 103, 104, 174, 181, 183, 233, 235, 290, 313].

Collective Actions. Social movements, collective actions, protests, and revolutions are engraved into human history. They have been widely researched, since they have direct impact on our lives. Such movements can vary from a simple legislation reform, as the civil right movement in The United States in the mid 1950s, to social revolutions, as it was done in France in 1789, Russia in 1917, or China in 1966, or to regime shift, as in Tunisia in 2011, or Ukraine in 2013. Without hesitation, these activities will continue to exist. However, digital connectivity reshapes how these movements connect, organize, maintain, and evolve during their lifespan.

Different types of movements might not develop in collaboration or at the same rate. Thus, it is not always true that digital technologies empower them in numerous ways, i.e., the effect of these technologies for the movements is rather non-uniform [291]. Therefore, when studying large collections of data, it is crucial to (1) carefully generalize the findings into various phenomena types, and (2) accurately predict the effects of each technology on the different movements. Moreover, to study collective and user-driven nature of the movements, campaigns, or any other types of the collective actions, *social media traces* are often used [44, 214]. These traces

might include any type of user generated content, such as, social media data (Facebook, Twitter, Instagram, etc.), crowd-sourced data (Mechanical Turk, CrowdFlower, etc.), activity tracking data (FitBit, Jawbone, Foursquare, Yelp, etc.), user generated content (GMail, YahooMail, WordPress, Medium, YouTube.). The ability to obtain these types of data provides a way to understanding of both individual behaviour and collective actions, as well as, offers activists services (personalized or general-use) tailored to the need of engaging more people.

On a Personal Note. I had the chance to personally experience what impact can digitalization have. In 2004, Ukraine experienced a first wave of protests that later developed into a widespread movement named “the Orange revolution”. The protests erupted after the 2004 Ukrainian presidential election, since citizens suspected the election to be massively affected by corruption and electoral fraud. While social media was not yet well developed at that time (note that Facebook was founded in 2004, and VKontakte, a more popular in Ukraine russian counterpart of Facebook, was founded in 2006), it was almost impossible for citizens and activists to easily reach a wide audience, and lead or participate in a public debate. Many relied on information provided through other channels, such as television, radio, newspapers, which normally can be used for spreading the information only by few social groups. It is therefore believed that the political party, which afterwards won the election, had great influence on aforementioned information sources and used them to organize activities, protests, demonstrations. Less than ten years later, in 2013, another wave of demonstrations and civil unrest, later named “Euromaidan”, developed on Maidan Square, after the government’s decision to suspend the signing of an association agreement with the European Union. However this time, many people were using social media and other digital platforms; thus, more groups, such as non-government affiliated activists, had a chance to reach a wide audience. Social media therefore became the major channel for organizing movements and demonstrations, spreading news, video footages, and ideas; and it was almost impossible for government institutions to stop or control the information flow they hence needed to rely on other measures, e.g., physical violence, which not always yielded desirable results. Specifically, during the early development of the protests, government institutions tried to stop a demonstration of university students by granting a permission to Berkut (special police forces) to use physical force. This backfired, as a video footage from the event spread over the social media at a rapid speed, which in turn lead to a greater number of demonstrations involving many more participants.

It is clear that digitalization empowered activists to spread awareness, but it also brought a variety of content specific issues and challenges, such as filtering of relevant information, efficient content creation, etc.; and platform specific issues and challenges, such as censorship, hate speech, etc. In this thesis, platform specific issues are surveyed in Chapter 2, while content specific issues are analysed in greater detail in Chapters 7 and 6.

Opportunity Space. Regardless of how efficient and successful online activism became, there is still a lack of analysis of the traces that are left by online activists on social media; understanding the underlying motivations and influences, which guide and engage users to participate, and empowering the organizational structures with efficient tools and methodologies to distribute and

consume user-generated content on the Web [291]. There is a limited number of studies covering the exploration of online and offline actions performed by activists; therefore, we have made an extensive study of online petitions, which is one of the particular tools identified to be used by the campaigns [233], which quantified how social media and other promotions shape user engagement. As a result of these studies, we have identified numerous examples of repetitive human generated content, and thus, proposed an efficient tool to extract the repetitive content, and use it further for the creation of the textual templates. Finally, due to the volumes of semantically similar information on social media, revealed by campaigns and petitions analysis, we proposed an algorithm that creates a representation of semantically similar messages of particular topics.

Specific Platforms and Domains. The overarching goals of this thesis are (1) to explore various aspects of digital activism and their challenges on social media, (2) to gain insights into designing dedicated tools that empower activists to generate and consume more content online. We cover these goals through four studies, each focusing on a well-defined domain – public environmental campaigns on Twitter, e-petitions from petitions’ aggregator *ThePetition Site*, email outboxes, and events on Twitter – for which the required data was collected in a way to represent a domain of interest, allowing us to explore various types of challenges when analysing collective actions.

1.1 Research Problems and Contributions

The primary purpose of this thesis is two-fold. We *first* broaden and deepen the scientific understanding of the types, approaches, success, and failures of the online activism, and *second*, build models and tools for supporting activists, as they navigate through the overwhelming amount of the created information, as well as, facilitate the content creation.

There is a growing body of work [95, 191, 235, 289, 290, 308, 313] that raises concerns about the ways online activism is framed and organized to answer a variety of complex questions about the nature of digital activism, as well as, to offer support and optimization of the movements’ influence and user engagement (information summarization [22, 228], prediction of user involvement into a campaign [235, 237], autocompletion [147, 234]). The research problems we formulate here are grounded in the literature, which can be categorized into three broad classes [73, 81, 156]: (1) profiling of online activism – highlighting various approaches to categorizing online campaigns, such as, goal and user engagement, as well as, defining general categorization of the online/offline actions performed by the activists; (2) modelling of online activism – stressing on issues of prediction accuracy and interpretability of content popularity; (3) facilitating online activism – focusing on challenges around content compression and filtering based two different corpora, user emails, and short texts respectively. The first two points are framed around the analysis of public online campaigns, the use of e-petitions to engage people in a campaign, and the modelling of factors that shape user’s participation in online petitions. We reflect on the last point as algorithmic contributions, in particular, we propose novel models and algorithms to predict content popularity and further suggest a methodology to efficiently summarize the content produced by and for the activists.

In this section, we formulate a set of broad research questions (RQs) concerning various challenges faced when analyzing traces generated by online activists. For each of them, we highlight the specific circumstances (defined by domains and applications) in which we study them. In the following, we detail our contributions and list the associated conference papers that were published during our research.

RQ1 (Profiling): Analysis of Public Campaigns. Applying a consistent methodology to collect the information about a large collection of public campaigns and their presence on social media is a good practice but can be highly prone to errors and requires substantial annotation and supervision. Nevertheless, to be able to gain insights about campaign's popularity, its main influence factors, and dissect online and offline actions performed, one should analyze a large collection of user activity that is annotated with high agreement and accuracy.

Context of the question. Referring to research of the social media traces for online public campaigns, e.g., Earth Hour 2015, COP21 [81], United for Global Change [103], which usually examines one or several online campaigns and tries to make generalizable conclusions to other campaigns in the same or different domain. This is particularly problematic as the outcomes of such research are usually intended to be reused for future campaigns by online activists. Yet, our research shows that user engagements and campaign's actions differ from one campaign to another even in the same domain of environmental and animal welfare activism [237].

The issues that we are looking at are: *What are the characteristics of public campaigns on social media? Can we differentiate public campaigns by their goals and user engagement? Are campaign's actions generalizable from campaigns with similar goals to campaigns with similar user engagement? What are the similarities and differences between campaigns regarding actions and which actions result in greater user participation? Do online petitions increase the popularity of public campaigns?*

To study these questions, we analyze the use of social media for over 100 environmental and animal welfare public campaigns that are associated with hashtags. Our systematic examination of a diverse set of attractive or less engaged campaigns uncovers substantial variability in the presence of various actions (calls for actions, duplicate content to promote the hashtag to trending) to engage more people in the movement. Moreover, we observe that despite the actions performed by the campaigners, e.g., created petitions, protests, conferences, publications, the number of highly active users on social media is directly correlated with the user engagement into the campaign. This work is covered in Chapter 3, and the results are published in:

[237] [Julia Proskurnia](#), Ruslan Mavlyutov, Roman Prokofyev, Karl Aberer, Philippe Cudré-Mauroux (2016) *Analyzing Large-Scale Public Campaigns on Twitter*. In: Spiro E., Ahn YY. (eds) Social Informatics. SocInfo 2016. Lecture Notes in Computer Science, vol 10047. Springer.

[233] [Julia Proskurnia](#), Karl Aberer, and Philippe Cudré-Mauroux. *Please sign to save...: How online environmental petitions succeed*. In EcoMo'2016 ICWSM, Cologne, Germany.

RQ2 (Modelling): Popularity Prediction of Online Petitions. The prevalence of popularity modelling and prediction studies concentrated on a single signal, e.g., a popularity of a post on social media [120, 324], new media [171], or attention to some domain specific events [205, 213, 229], has led to substantial concerns about the interpretability and completeness of the proposed models. As Tufekci [289] remarks, the focus on certain platforms is appropriate. However, the effort put into understanding of the influences between various platforms is necessary to understand better the effects they make on each other. Recently, an attempt was made to study the effect of several signals within a platform on full [156, 324] or partial prediction task [245, 246]. Yet, it is not well studied, *how content popularity changes under the explicit promotion*.

Context of the question. The data that we obtain from the web is almost invariably noisy and incomplete and often comes from several heterogeneous sources. In particular, reverse engineering the ways to promote a campaign or online petitions is rather challenging due to the data availability. Empirical evidence shows that e-petitions are promoted on multiple platforms with some being more popular [78, 235], e.g., Facebook vs. Twitter. Moreover, activists utilize several sources to reinforce the campaign [291] and those are often heterogeneous. For online petitions, there is a growing need to model these combined effects from multiple platforms, and thus, enable a consistent decision-making framework for the activists.

Following these observations, we focus on online petitions covering various topics and promoted both on social media and on the petition's platform itself, and we are particularly interested in the following questions: *How prediction of the various correlated phenomena can be achieved with high accuracy? How can the influence of the external effects be estimated? Which signals have the greatest effect on the signature rate? How causal influences can be traced between studied phenomena and confounding components?*

To tackle these questions, we propose new forecasting models for online content dissemination that can capture several heterogeneous elements: self-excitation, seasonality, web platform artifacts, social media, and is further enhancement of the interpretability of the results. We confirm the strength of the influence of the external signals by performing multiple Granger causality tests, as well as tracing platform-to-platform user behaviour on an individual level by linking user accounts to them. To evaluate our model we used over 4,000 e-petitions aggregated on *The Petitions Site* from a variety of suggested topics (environmental, human right, LGBTQ, and others). We use our model to predict the time evolution of the signature acquisition by each petition and show that We show that our model outperforms by a significant margin multiple state-of-the-art methods on both short- and long-term prediction. Additionally, we have quantified how external factors shape the user engagement, e.g., social media has prolonged effect while explicit promotion on the front page shows the highest gain in number of people signing. This work is discussed in Chapter 5 and has been published in:

[235] [Julia Proskurnia](#), Przemyslaw Grabowicz, Ryota Kobayashi, Carlos Castillo, Philippe Cudré-Mauroux, and Karl Aberer. *Predicting the Success of Online Petitions Leveraging Multidimensional Time-Series*. In Proceedings WWW '17, pages 755-764, Perth, Australia, 2017.

RQ3 (Facilitating): Content Filtering Considering the vast variety and volume of the information that is produced on the web, and particularly, on social media (both human and machine generated), the representativeness, the completeness, and the accuracy of content filtering is a challenge. It should be noted, several works make emphasizes on the keyword content filtering [189, 217, 289], e.g. words, phrases for the documents, or hashtags and keywords for the social media postings. However, usually, those approaches rely on either ranked list of keywords or sets of keyphrases preselected by an expert, which can result in a decrease in precision and recall of the content filtering and usually require a lot of adaptation to the domain. In other words, an incorrect list of keywords could result in missing relevant information or detecting too much of the irrelevant information [105]. On the other hand, similarity based techniques or classification approaches are either too computationally intensive or imprecise respectively.

Context of the question. A problem of efficient and precise filtering of the relevant content was raised in multiple contexts on the social media, such as, events [215], disasters [68, 214], protests [289, 291], elections [93], terrorist attacks [16]. Clearly, in the context of filtering of sensitive content mistakes are costly, as they could result in serious consequences, such as, over- or under-estimation of the presidential election, over- or under-reaction to the occurring disaster or an attack.

We focus on the filtering of the information relevant to terrorist attacks and answer the following questions: *Can we filter relevant content from the high volume stream of data efficiently and accurately? Can we minimize the amount of the prior knowledge and training data used for a particular topic? Is it possible to generalize the method for unseen topic detection?*

To this end, our strategy is to take advantage of the (1) semantic representation of the set of given query documents, e.g., event description, related social media short texts and so on, as well as (2) effective pattern mining techniques to get the minimal representation of given topic that results in further extraction of the homogeneous documents. In particular, we propose an algorithm to generate minimal patterns, e.g., sets of unigrams and semantic word clusters, that represent the set of given training examples and are further used to extract unseen examples similar to the training set. In other words, our method constructs a representation of a topic that (1) guarantees to create a filter for extracting relevant content accurately and with a consistently increasing recall as the training input size grows, and (2) is robust to the small amount of noise in the training sample, (3) produces consistently good results for various tuning parameters (similarity threshold and support). We evaluate our approach and compare it to multiple baselines, and show that it leads to better F1 measure on even small training sizes. Chapter 6 details this study:

[236] Julia Proskurnia, Ruslan Mavlyutov, Carlos Castillo, Karl Aberer, Philippe Cudré-Mauroux. 2017. *Efficient Document Filtering Using Vector Space Topic Expansion and Pattern-Mining : The Case of Event Detection in Microposts*. 26th International Conference on Information and Knowledge Management (CIKM '17) Singapore, 2017.

RQ4 (Facilitating): Content Generation. Another area of concern is how to efficiently empower activists to produce the vast variety of content that is disseminated through various channels, such as, social media, emails, to increase awareness and mobilize people [81]. Multiple tools are created and open sourced that summarize and filter relevant information about particular events and campaigns, which are further used to optimize activists’ workflow [32, 62, 66, 155, 197]. However, apart from filtering valuable information, it is also important to analyze writing patterns of the users. Several studies are advocating for automation and facilitation of the question answering and autocompletion [63, 135] – which helps with replying and extending upon a given query, e.g., answer a particular request or question, reply to a message or email. Nevertheless, a substantial part of the user generated content is newly produced without an explicit request for it.

Context of the question. One of the most attractive and efficient ways to distribute information is either through social media channels, such as postings on Twitter or Facebook pages, or dissemination of the information via email and email groups. While the former usually consists of shorter messages, it is computationally less expensive to process and summarize such information [172]. At the other side of the spectrum, the latter (email groups) might consist of long messages, thus making it impossible to apply complex methodologies at scale. Moreover, several studies showed that human generated content is rather repetitive [164, 234, 237] and is usually specific for each user [63]. These observations motivate the need to (1) extract these repetitive phrases and (2) generate high-quality message/document templates that are frequently used by the individuals, thus saving typing time. This particular problem could benefit activists to save their time on generating content that is to be distributed among the public.

As a result, the questions that we want to emphasize are the following: *Are the greedy methods perform similarly to the optimal methods in the context of fixed phrase extraction? Can we generate accurate and usable templates without scarifying the quality and complexity? How many users might benefit from the created messages and how much time activists could save?*

To this end, our strategy is to take advantage of already existing efficient data structures and show, that by modifying and adapting *the suffix array* to the document phrase extraction, we can improve the quality (with emphasis on longest repetitive phrases) of the extracted frequently used phrases as well as further construct of the message templates. We describe a novel, generic and linear in complexity algorithm to extract the fixed phrases that works on any textual corpus and can be generalized to any language, as well as a simple yet effective solution to find a compressed representation of the message – template. We evaluate the proposed solution on synthetic and real-world datasets and show that we can maintain a better quality of the template no matter the quality of the input.

[234] [Julia Proskurnia](#), Marc-Allen Cartright, Lluís Garcia-Pueyo, Ivo Krka, James B. Wendt, Tobias Kaufmann, and Balint Miklos. *Template Induction over Unstructured Email Corpora*. In Proceedings of the 26th International Conference on World Wide Web (WWW ’17), pages 1521-1530, Perth, Australia, 2017.

Methodological commonality. Two last chapters of the thesis, while being different, share some commonalities at both high and low level. Both of the proposed methods are aiming at finding patterns in short and long documents respectively. In particular, both methods can be well suited for the compression of the data. In the first case, the messages of a particular topic can be summarized using patterns that further will be stored to represent similar messages. In the second case, template of a particular cluster of emails can be stored only once and only variable parts will have to be stored for each particular document or email.

We note that all the research questions are contingent on online activism, e.g. the understanding of the collective actions can guide its further modelling and forecasting, which in turn can help to design and develop tools and methodologies to empower and facilitate the activists to start, maintain, and follow up on the issue they are addressing.

Finally, we emphasize that, in this thesis we seek to study the research problems described above in the context of social media or email communication (which is also considered as a social sharing tool [214]), highlighting the improvements and benefits of the proposed methodologies and algorithms. We do not attempt to provide an ultimate solution to the question of how to create and maintain a successful campaign, petition, movement, protest, YouTube channel, blog. We do not claim the generalizability of the finding to other types of online activities, such as electoral campaigns, crowdfunding campaigns, blogging. Moreover, not all issues and challenges of online collective actions are addressed by this work, e.g., we do not cover the detection and spread of the misinformation or disinformation, influence maximization, information diffusion, a priority ranking of the social media content, harassment, hate speech, or offensive campaigns. We discuss in more details and depth environmental campaigns, e-petitions, and algorithmic challenges of questions above about facilitation of collective actions. This thesis contributes *a practical perspective of a research body that aims to understand digital activism by its online traces, to design more accurate and interpretable popularity models, and to devise and evaluate efficient algorithms to facilitate online content consumption and production.*

1.2 Thesis Outline

We now give an overview of the thesis structure and describe how the Chapters are grouped and how they relate to the research problems that are formulated in Section 1.1. The conceptual flow of the thesis is highlighted in Figure 1.1. The thesis is organized into four parts. Chapter 2 is covering the broad context of the research problems we address in this thesis, including online and digital activism, its primary challenges, description of various types and methods applied to the modelling of the user engagement, and finally, an overview of the most common tools, approaches, and methods in the context of facilitation of the activism in the digital space. Chapter 3 and 4 are describing profiling of online activism in the form of public campaigns on Social Media; it comprises of a detailed methodology for identification and categorization of the campaigns by their goals and user engagement, as well as the role and importance of online petitions to perform actionable changes. Chapter 5 is focusing on the modelling of the

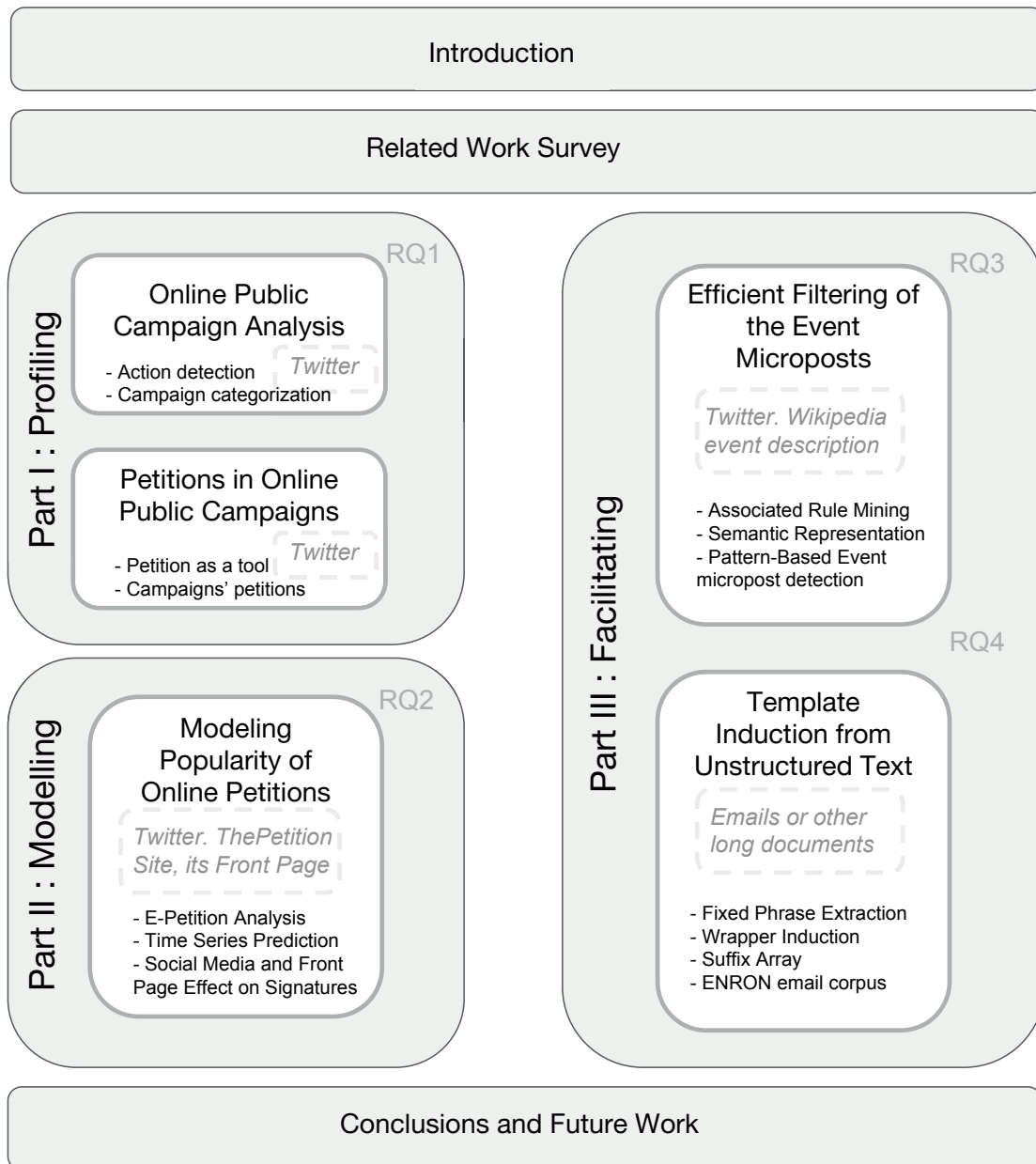


Figure 1.1 – Overview of the structure and the conceptual flow of this thesis.

user engagement and campaign popularity. Chapter 7 is studying the ways to facilitate online activism, both concerning content consumption and production.

Part I: Background

Chapter 2 lays out the broad context in which online activism is developed, modeled and optimized. By surveying relevant prior work that, first, raises awareness on online campaigns, movements, online petitions, protests, etc., second, explores the limits of modelling reinforced

phenomenon as a function of multiple confounders, third, and scrutinizes the tools, methods, and approaches to both facilitate content consumption and production.

Part II: Profiling

Chapter 3 looks at public campaigns and their presence on social media, exploring how different campaigns acquire their users and which actions are the most efficient to enhance user engagement. For this, it introduces the methodology to categorize public campaigns by their traces on Social Media, as well as identify most common campaign actions.

Chapter 4 studies to what extent public campaigns utilize online petitions as a tool to reach the goal and how these petitions are actively promoted on social media by the campaigns.

Part III: Modelling

Chapter 5 investigates how we can model the time evolution of the reinforced, promoted phenomenon, e.g., e-petition. It introduces a novel model that accurately mimic hourly attention of the online petitions by using social media and petition's platform signals. Moreover, it shows how our model can be interpreted for further utilization by the activists.

Part IV: Facilitating

Chapter 6 explores how to filter documents that are semantically related to a query defined by a small sample of the reference documents, e.g., paragraphs, microposts, keywords, phrases. The chapter introduces a novel approach that, first, allows to construct a compact and interpretable representation of the topic, and, second, efficiently extract relevant document.

Chapter 7 tests unstructured text autocompletion on the example of templates generation using frequently composed user messages, i.e., text documents, emails, short texts. The chapter introduces a novel approach that allows composing user's template in linear time with high accuracy.

Chapter 8 concludes the thesis by summarizing the main contributions and outlining possible directions for the future work.

Background Part

2 Online Activism on Social Media and Beyond

In this chapter, we provide a broad context in which we frame the research problems that we tackle in this thesis by surveying prior work. In particular, we focus on the analysis and methodologies of tracking, maintaining and facilitating collective actions.

As development and use of the digital tools and platforms flourish in multiple domains, including political uprising, public campaigns, news media generation, etc., the community has also started to explore the ways to analyze the traces left by the collective actions on those platforms as well as facilitate their usage. Core research areas include analysis of the online activism through the online traces left by activists on social media platforms [81, 237], content popularity with respect to the content produced by the activists [230, 235], complexity and volumes of the content as a result of consumption or production of the user generated content [147, 234], to name a few.

Here we focus on three main aspects of the collective actions:

- (1) profiling, describing, identifying, categorizing and analysing the collective actions (Section 2.1),
- (2) modelling collective actions as a content popularity or user engagement into a particular action (Section 2.2), and finally
- (3) facilitating collective actions on social media platforms by surveying typical applications, methods and tools (Section 2.3).

While some of the prior and related works study the aspects related to the collective action within a general framework [44, 291, 313], most of the studies cover them in the context of a specific topic or platform [73, 126, 215, 289]. We note that studies such as those related to privacy, legislation and organizational depth of the collective actions are outside of the scope of this Chapter. Moreover, the categorization in the following chapter is not mutually exclusive and can overlap. Finally, methodologies and algorithms surveyed in the following sections are not complete by any means; however, they provide an excellent overview of the most representative works in the field.

2.1 Digital collective actions and its analysis

The chapter explores and highlights the research that raises concerns about the collective action during the digital era [290, 291] (Table 2.1), as well as research that tries to analyse, structure and understand collective actions [44, 52, 78, 316] and their main channels and patterns of communication (Table 2.2). I recommend anyone interested on online activism (both from the technical and sociological perspective) to read the book of Tufekci [291]. In particular, we *first* list variations of the digital activism, such as, non-governmental, political, crowdfunding campaigns, petitions, movements, uprising, etc., and *second* their main characteristics in terms of the involved stakeholders and their interactions online. Finally, we point out some pitfalls and biases in respect to the online data analysis. See Table 2.1, 2.2 for a comprehensive overview of the relevant prior work.

Digital Activism. The digital era and the growth of internet connectivity brought a new wave of collective actions (Table 2.1), in particular, in Turkey, Egypt, Mexico [104, 174, 290]. Networked protests [72, 220, 231, 263], campaigns [81, 103, 104, 191, 237] and movements [77, 129, 284, 308] in XXI century differs in many significant ways from the movements of the past [291]. Digital technologies are so integral to the social and public campaigns and movements that many of them are referred to by their hashtags [289] rather than official names – #jan25, #direngezi, #occupywallstreet, #actonclimate, #helpcovedolphins, etc. Majority of works in these domains are focused on the prediction of the hashtag adoption and user participation evolution [89, 181, 219, 237, 249].

Online Petitions and Crowdfunding. Sometimes, collective actions online are associated with “clicktivism” as known as “slacktivism” [118] (an act of easy actions, effort or commitment). However, an assumption that people who connect online are involved only online are rather controversial [291]. Some examples of the online actions that make an impact on the current society are online petitions and crowdfunding campaigns, where a so-called “click” corresponds to either ideological or financial support respectively. The former is widely researched for particular platforms, topics or impacts [132, 145, 183, 233, 235, 313]. In particular, I consider papers of Yasserli *et al.*[313] and Proskurnia *et al.*[235] the most prominent for the analysis of the online petitions. The latter is explored in the context of the *Kickstarter* platform that encourages and empowers local manufacturers, creators, and inventors to spread their products to the public. Most of such research is directed towards exploring the ways to increase user participations, and thus, recommending the means to reach the target audience in an efficient manner [17, 78, 193, 204]. For those interested in crowdfunding, a paper by Etter *et al.*[78] should be considered as the starting point.

Political and environmental activism. Digital tools provide an ability to rapidly amass a large number of stakeholders to empower social, political or environmental movements. In this context, on one side, several stances can be supported by the media, people or activists. On the other hand, the main problems of the social movements are usually “lack of experience and organizational support of the tools used, or even culture of collective decision-making and

| <i>Application</i> | <i>Short Description</i> |
|--|---|
| ... Digital Activism | |
| Digital movements | [95] summarizes social media practices among popular movements. In particular, media role in campaigns such as, “shabab-al-Facebook”, “Twitter pashas”, “indignados” [104], “Occupy Wall Street”, and their differences. [290] describes potential weakness and opportunities of the protest organization and maintenance through digital infrastructure. [174] digs into Save Darfur “Cause” massive online social movement and studies how it failed to convert clicks to donations. [289] emphasizes the opportunities and pitfalls that big data brings to analysis of the phenomenon reflected by social media. |
| ... Campaigns and movements | |
| Multiple political, environmental and social campaign analysis | [81] describes and analysis communication, engagement and behavioural change for Earth Hour 2015 campaign and United Nations Climate Change Conference 2015. [237] analyse and classify over a hundred online public campaigns that appear on Twitter and explore campaigns’ user engagement patterns. [104] identify user roles in the social network on the example of “indignados” movement [19, 103]. Mobilizing members for political activism (<i>MoveOn.org</i>) is well observed and presented as manufactured communities in [77]. [191, 308] describes “Anonymous” movement that emphasizes the importance of free speech and support of Wikileaks. Several works detail description analysis and information dissemination during 2011 Tunisian and Egyptian Revolutions [186, 307, 314], “Arab Spring” [129], “London riots” [72, 99, 231, 284, 296]. |
| ... Hashtag activism | |
| Spam Campaigns | [58] characterizes spam campaigns represented by a hashtag. |
| Campaigns as hashtags | [181] describes social dynamics of emerging hashtags. [237] highlights over a hundred public campaigns associated with a hashtag on Twitter. Following are the studies for highly trending hashtags, such as, “#blacklivesmatter” [219], “#AllLivesMatter” [91], “#Ferguson” [89], political hashtags [249] etc. |
| ... Online petitions analysis | |
| E-petitions platforms analysis | Several works explore the dynamics in e-petition support and usage patterns of the various platforms, such as, German Bundestag petition platform [145, 183], petitions associated with environmental public campaigns [233], UK government [313], UK No. 10 Downing Street website [117], <i>Change.org</i> [132], <i>thepetitionsite.com</i> [235]. |
| ... Crowdfunding campaigns | |
| Kickstarter success prediction and user recommendation | [78] predicts the success of the Kickstarter campaigns by utilizing campaigns metadata, social media signal and language used in the descriptions [204], gender dynamics [193]; [17, 33] recommend various products to particular social media users. |
| ... Political and environmental activism | |
| Protests | [263] quantifies the importance of social media and its ability to represent protests and campaigns over time. [220] presents a large scale analysis of the protests from over seven years. [19, 103] study mobilization of people through social networks as well as structural holes and bridges in those networks during protests (“indignados”). Particular protests are studied in terms of habits, opinions and behaviour on Twitter, e.g. London riots [72, 231], Occupy Wall Street [287]. |
| Politics | Political elections and opinion mining from social media gained an extensive attention during the last decade, e.g., election prediction in USA [292], Holland [258], collective actions [10, 57], candidate approval ratings [60], political opinions [203, 271], supporters interaction [4]. Hundreds of emerging hashtags during the 2012 election in USA were analysed by [181]. [50, 241] study political movements originated from <i>MoveOn.org</i> as well as highlights some evidences of first-, second-level agenda setting. |
| Environment | Multiple research scientists explore the effects and implications of changing climate [12, 128]. Thus, social media is a crucial instrument to convey information and mobilize people to act [237]. [215, 260] study the extend to which social and news media [122] are aligned when reflecting climate related events. [154] explores the discourse about the climate change on social media. |

Table 2.1 – Various analysis of *digital activism*, e.g., campaigns, e-petitions, protests and movements.

long-term actions” [291]. Especially, it remains unclear how such movements would develop if traditional censorship and conspiracy could have been evaded. Despite possible discrepancies in activism, research on this topic focuses on single or several instances of the phenomena [10, 12, 57, 122, 128, 258, 292] (election, uprising) rather than generalizing the finding to multiple events.

Social network analysis. Analysis of the collective actions on social media (Table 2.2) has several aspects that are usually explored by the research community, such as, information diffusion, community analysis, and influence propagation [106, 316]. A survey of Guille *et al.*[112] provides an extensive overview of the information diffusion in online social networks, thus is recommended for those starting to get involved into this domain. Multiple studies focus on the information cascades [101, 276] about a particular phenomena [107] (event, online content, micropost, video). Cascade predictions are usually studied after observing its developments over time [55, 162], rather than before a cascade to happen [44]. Several works [15, 235] go beyond an online interaction and show how offline actions shape online ones and vice versa.

Online communities. The development and growth of the online communities on social media simplified and amplified information propagation on the digital platforms. More information on community detection on social media can be found in the survey by Papadopoulos *et al.*[224]. Research of the network structure is prevalently based on the graph representation of the users and links between them on one or several digital platforms. There exist several approaches to identify structural, topical or topological types of communities [75, 88, 153, 209]. Interestingly, weak ties [108] between network participants are shown to be the most efficient for the information spreading within and between like-minded communities [323]. As a result, these types of ties are often lead to the field evangelists (influentials) that are described as part of the structural community or entire network.

Influence online. Online activities heavily rely on the online platforms and digital tools for communication, organization, and publicity of any collective action. However, the affordance of the large group participation of people does not always mean uniform participation; thus, numerous works focus on the influence propagation in the online social networks and further suggest the ways to maximize information propagation in such networks [52, 53]. Kempe *et al.*[150] is the most prominent and inspiration work for the influence and information spread maximization. Gonzalez-Bailon *et al.*[104] describes a number of roles that participants of the online interaction might take, among them are influentials, hidden influentials, broadcasters and others. Other researchers [49] define the influence propagation to be initiated from the mass media, “evangelists” or ordinary people. Another area of research studies whether there is a connection between the online influence to the offline world [108] and reveals that new information is more likely to spread over weak or intermediary connection.

Pitfalls and biases. Variety of the analysis and research of networked collective actions are attributed to the actions performed online on social networks. Despite its popularity, important questions about the limitations and “proper” ways to use social media data are gaining traction in

| <i>Application</i> | <i>Short Description</i> |
|--|--|
| ... Network analysis, communities and influencers | |
| Information propagation | [52, 316] summarize main principles and ideas of the social media mining, in particular, information diffusion, influence, community analysis. [101] explores information cascades on a variety of platforms. [107, 249] emphasize the importance of topicality on information diffusion. [207] studies not only the patterns and content shared by the users of social media but also their network interaction, i.e., creating new and destroying existing links. In particular, [276] models the retweet time on Twitter. [15] goes beyond online setting and explored how online actions shape offline ones and vice versa. [136] studies conflicts formation and collaboration on Wikipedia. |
| ... Online communities | |
| Community detection | Commonly in Social Media communities are detected using Louvain modularity based [88, 209] community detection algorithm. [153] surveys various community detection techniques in multi-layer graphs. [75] compares topological and topical communities and their differences. |
| Community analysis | [42] proposes a scalable joint community profiling and detection model that characterises the community through content and diffusion profile. [304] explores information diffusion across communities as complex contagion and finds that some viral information might spread across communities, like diseases. [131] models topics and communities in a unified framework while [278] infers latent topical communities among users using probabilistic generative model. [82] studies how collective trend emerges from individuals' topical interests, i.e., network structure, dynamics of content production, user's behaviour. [27] recommends and explains the link prediction in social networks, e.g. topical and social link. [238] studies group behaviour of content generation. [167] explores how a post's title, community, timing affect its further popularity. |
| ... Influence online | |
| Influence Propagation | [52] presents an extensive overview of information diffusion and models as well as its application in the form of influence maximization. [49] compares information diffusion patterns from three user types, e.g., mass media, evangelists, ordinary users. [252] proposes a joint model to identify communities and roles simultaneously. [285] explores how to identify important information, how to promote and motivate those who spread these information and finally how influential individuals can be identified. [104, 121] use structure of the networks and activity levels to identify user roles and their position in the discourse. [108] establishes a parallelism between online and offline social networks. [106] explores users popularity as a temporal evolution of one's audience. |
| ... Pitfalls and biases | |
| Social media analysis biases | Despite an enormous effort in social media data analysis, it is important to account for hidden biases of data collection, analysis, methodology and algorithms [65, 254]. [214, 216] go deeper into social media biases and pitfalls by summarizing and surveying major issues. In particular, [289] describes the recent biases towards platforms with open data to study human phenomena. On the other hand, [40] shows both challenges and opportunities of the "Big Data" era, thus importance of eliminating biases to enable better decision making. |
| Algorithmic fairness | Apart from emphasizing the importance of proper accounting for various confoundings, it is also crucial to design algorithms that are able to provide interpretable and fair solution [80]. Some recent works search the balance in fairness and diversity [34, 48]. The main issues that are studied in regards to the fairness of the algorithms is concentrated around gender and race or individual and group discrimination [116, 280] and a set of methods to rule out the discrimination are rule mining, similarity based approaches, bayesian and probabilistic causation. |

Table 2.2 – Activism profiling through social network analytics, communities and influencers detection as well as pitfalls and biases of social media analysis.

the research community [40, 44, 214]. In particular, a work of Olteanu *et al.*[216] surveys most prominent issues and their corresponding solutions. Limitations of the social media analysis can be split into, (1) biases related to the data extracted from the social media that is usually noisy, incomplete, poorly aligned, and, (2) biases incorporated to the algorithms and methods applied to the social media data. Since not all the activity online, current research also emphasizes the importance of exploring offline actions performed by the users of social networks. However, a general problem of attempting to answer questions about collective actions using online data is that such data can be big and have no guarantees of representativeness of the general population. Finally, some of the relevant information might be even missing from the online public sphere due to the sensitivity, misunderstanding and individual capacity to communicate effectively [44], e.g., issues on women's or LGBTQ rights, unpopular or opposing opinions.

2.2 On modelling and predicting the popularity of online content

Activists might be considered as the bridges for a broader public who can be mobilized. However, to make a noticeable impact, large social movements and campaigns require the participation of a vast number of people [82, 235, 245, 283]. Most of the time, those people do not have prior experience or even knowledge of the tools and platforms that facilitate mobilization and knowledge spread. Due to the wide spread of social networks, such as Facebook and Twitter, users could support (like, click, sign) for the first time an digital invitation to a social movement, public campaign, online petition, political uprising.

Predicting the popularity of an online item is an active field of research. Many different types of online items have been studied such as YouTube videos, online news, social networking campaigns, crowdfunding campaigns, online petitions. On social media, a user participation or interest can be registered as an activity (tweet, retweet, "like", follow action). Most works in the area of popularity prediction focus on answering the following questions:

- (i) *Can an online item become successful?* This question involves the predicting if the total popularity of the online item will be larger than a threshold [55, 69, 126, 138, 187]. We recommend reading the work of Hong *et al.*[126] if you start working on this direction.
- (ii) *How would popularity evolve over time?* This question relates to time-series forecasting, i.e modelling the popularity dynamics over time [6, 92, 156, 182, 240, 246, 266]. Works of Gao *et al.*[92] and Rizoii *et al.*[246] are recommended to get an overview of the best approaches for this problem.
- (iii) *Can we predict the final popularity of an item?* This question correspond to the regression task where the final number of attention shall be predicted [26, 120, 126, 162, 163]. We recommend Kupavskii *et al.*[162] to gain some initial insights into the problem.

Regardless of the task and a particular phenomenon to be modelled, two types of methods have been developed to solve these problems. First, machine learning techniques rely on an exhaustive list of potential features extracted from the phenomenon's online traces, including structural and

2.2. On modelling and predicting the popularity of online content

| <i>Approach</i> | <i>Domain</i> | <i>Short Description</i> |
|--|------------------------------|---|
| ... Classification | | |
| Logistic regression | Twitter | [126] studies tweet cascades classification, where the most prominent features include the history of retweets and the number of users' followers. |
| Feature based | Twitter | [187] uses various features for the classification—such as the number of tweets containing a certain hashtag during a certain time period and the number of unique users that post messages with a certain hashtag for the same time period. [138] compares Generalized Linear Model and Naive Bayes and uses number of followers, tweet length, sentiment, URLs, number of hashtags in a tweet as features. [55] focuses on the prediction of the structure of the reshare cascades using temporal and structural features. [69] constructs a temporal analysis of hashtags in order to discover breaking events in real-time and tried to distinguish the hashtags of social events from hashtags of virtual topics or memes. |
| Social transfer Model based | Multiple Google Trends | [251] utilizes external information to model the video popularity. [67] studies the effects of the external shocks on the time series evolution and thus classify the content as one of the three burst patterns. |
| ... Final Number | | |
| Feature based | Twitter | [163] compares the prediction of the retweet cascades as well as shows cascades using multiple social, content and infection features. [162] examines a number of retweets a tweet might obtain using the flow of the retweet cascade and PageRank score on retweet graph. [26] predicts if a tweet is retweeted more than a certain threshold based on the structural characteristics of the networks spanned by early retweeters. |
| SVM, KNN. Feature based. | News | [25] focuses on the prediction prior the release of the item of interest. |
| Logistic Regr. Bipartite graph | Twitter | [120, 126] study prediction of the absolute content popularity based on the single source of information. |
| Social dynamics | Digg | [171] utilizes social influence and Digg web site layout to predict content popularity (being promoted in the friends page). |
| Model Based | YouTube | [230] proposes adaptive model selection based on the similarity to the previously seen examples. |
| Model based | Twitter | [324] uses self-excitation component (point process) that allows to predict whether a post will become popular and what will be its final number of reshares. |
| Model based | Earthquake, neurons, crimes. | [205, 213, 229] utilize a point process models that predict space-time earthquake patters, activity of neurons and crime rate respectively. |
| Model based | Multiple | [56] employ the retweet data that used to perform timely query expansion based on temporal information, i.e., retweets of documents are used to boost documents' relevance over a period of time. |
| ... Time evolution | | |
| Time series clustering | YouTube, Digg, Vimeo | [6] focuses on the content clustering based on the evolution of its popularity and prediction the popularity of the content based on its transitions between various evolution patterns. |
| Model based | Twitter | In some cases, retweets are modelled as point process due to the instantaneous nature of the tweets [92, 266]. The model assumes the multiplicative nature of the diffusion as a tweet tend to trigger another ones. [156] also incorporate the circadian nature of the underlying phenomenon into the model. |
| Model based | Multiple | [182, 245, 246, 310, 311] analyse the cross platform effect on the content popularity. For example, the structures of the influence networks between various processes as a result of Granger causality [109] or the effect of the breaking news, posts from social friends and user's intrinsic interests on content popularity. |
| Model based | Search queries | [240] introduces Dynamics Model Learners algorithm that incorporate an internal trend and periodicity of the time series. |
| Temporal clustering. K-Spectral Centroid | Twitter | [312] proposes a new metric that is invariant for scaling and hifting of the time series. |

Table 2.3 – Overview of the most prominent approaches and applications of the popularity predictions on social media.

temporal characteristics and features from other sources that affect the cascade. Then learning methods are applied for the purpose of classification or regression. These kinds of methods have drawbacks, including a high dependence on the quality of the features, the requirement of computing power due to the requirement of an exhaustive training, and in some cases the model's interpretability is limited. Second, model-based techniques aim at calibrating a specific parametric model that we assume that drives the phenomenon. The main drawback is that in some cases they are hard to formulate; however, they are more interpretable.

In the following we focus on predicting, modelling and describing various online and offline, real-world phenomena with data sets of online digital traces of human behaviour collected from various sources, such as videos [6, 230, 246], posts [25, 26, 55, 126, 324], blogs [6, 171], Google trends [67], search queries [240], memes [172], online petitions [233, 235, 313], campaigns [81, 237], natural phenomena [205, 213, 229]. A detailed description of the works on online and offline content popularity prediction on the web and social media is shown in Table 2.3.

2.3 Facilitation of collective actions

Online activism aims at attracting millions of Internet users. On one side, to be able to reach such a vast amount of people, an activist must seek the ways to produce high-quality content and communicate it, e.g., posts on social media, email to the individual users, posts on forums, email groups, or write articles and blog posts. On the other side, large number of the users are willing to follow the lead of the activists and take actions. However, the large flow of the information from a variety of sources is often impossible to effectively consume and filter. In this chapter, we bridge the gap between producers and consumers of the information on the web. We present an extensive list of approaches and methods to facilitate a writing effort, i.e., autocompletion, summarization, templatization (Table 2.4), as well as, filtering of the information on a particular topic, irrelevant or spam messages (Table 2.5), events (Table 2.6). We recognize that the categorization defined in the following tables is not complete and some categories might overlap.

Auto-Completion: A great variety of the activities online can benefit the users and activists in two most important aspects: first, reduce the possible typos and errors made while generating the content and, second, minimize the time used to type the content. Autocompletion of text was studied in several contexts, such as, search queries [30], microposts [21, 265, 274, 295], emails [63, 147, 242], documents [135, 294]. SmartReply is the first email-specific machine-learned tool designed to assist the user in composing an email. The tool is built on recurrent neural networks (one to encode the incoming email and the other to predict possible responses) that automatically suggests replies to email messages [147]. As a result, it provides canned responses meant to satisfy as many users as possible. Finally, another area of study is the generation of the chat bot [83] replies similarly to the automated responses to one's emails. Little research was made to profile individual users and extract their specific phrasings that were written in the past [234]. Overall, a work of Kannan *et al.*[147] and Hyvonen and Eutu [135] is one of the prominent and the most extensive description of the auto-completion and response generation

from the unstructured text.

| <i>Approach</i> | <i>Short Description</i> |
|--|---|
| ... Template Induction | |
| Web Page Templates | |
| HTML tree structure analysis of the web pages. Clustering. Rule and probabilistic modelling. | [20] suggests the use of nesting of equivalence classes to find the most general structure of the web page to generate required pages. [115] uses structured classification and clustering based on the document HTML structure which is transformed to the document fingerprint. [164, 166, 259] summarize web data extraction tools, in particular, authors focus on the parsing, analysis and merging html tree structures of the web pages. |
| Email Mining and Templates | |
| HTML emails representation as DOM trees. | [73] analyses HTML emails [13], that are parsed into DOM trees and further merge by its branches. [321] proposes to use conditional random field to represent the HTML email and formulate the templating as a prediction task of the next email segment. [8] describes the threading of emails that belong to the same conversation. [302] utilizes templates to assign email labels in a more accurate manner. |
| Generalization of the suffix array on text emails. | [234] proposes the use of suffix array to identify fixed segments of the text document that is not marked up with HTML tags. |
| ... Auto Completion | |
| Indexing. Categorization | [30] compares multiple IR techniques to suggest next word, and proposes a novel pre-computed inverted lists for unions of words. [135] extends the idea of word auto-completion with the semantic information obtained from the domain ontologies. |
| Deep Learning. Recurrent Neural Networks. | [294] shows the use of the RNN (seq2seq) on the next word generation task. [295] describes how LSTM [124] can be used to generate the image caption (text) based on the image itself (picture). [274] shows how RNN could be used to generate responses with some additional signals, e.g., context. [264] uses bootstrapped word embedding representation of the input. and bidirectional Recurrent Encoder-Decoder. [265] proposes Neural Responding Machine for the generation of the short-text conversations. [63, 147] describe the latest framework to generate human-like responses on the fly in response to an email. |
| LSTM, LSH, Graph Learning. | [242] describes a state-of-the-art techniques to map a given message to a set of possible replies. Interestingly, an inference of the replies is done locally on the device, thus making the flow more private. |
| Feature based. Statistical Machine Translation. | [21] suggests a simple feature based model to predict if a particular microblog (tweet) will be replied to. [24] proposes a model to predict the length of the response (feature based) as well as user's participation (self-excitation and bimodal distribution of participation). [243] uses statistical machine translation model to generate a response to a tweet. [222] uses a simple N-gram language model with topic modelling to predict the next probably word of a query. |

Table 2.4 – Overview of the most prominent approaches and applications to facilitate content generation. In particular, we focus on template induction, email mining and auto-completion.

Template Induction: Web data is generally formatted in a human-readable format which a machine renders but might not necessarily understand. One such example could be a web page of an event whose body contains images and text organized either using HTML tables or division tags and CSS. When rendered, the information will often be presented in an appealing way to the user. However, the rendering machine will not know what rendered information is most pertinent to the page's purpose (e.g. event title, location, date, start and ending time). Web extraction techniques have been successfully proposed to solve this issue of extracting information structured for human comprehension from the structured web pages [20, 115, 164, 166, 259]. In particular,

works of [164] and Proskurnia *et al.*[234] can be considered as a starting points to get more insights about template induction for HTML and plain text documents respectively.

Much research rests upon the assumption that online content is rather structured; nevertheless, the majority of human generated content is not structured and usually represented in a plain text format[188] (microposts, 10% of emails, blog posts). [172] uses an expensive edit distance computation for the microposts and suggests a methodology that summarizes the short messages. On the other hand, [234] proposes a method to produce a template of plain text corpora (regardless of being short or long texts) that is highly scalable and accurate.

On Elimination of the Irrelevant: Information on the web and social media differs from the conventional news media since it lacks traditional editorial and censorship board that could filter and spot contradictory and false information and prevent it from spreading, and vice versa. Tambuscio *et al.*[282], Kumar *et al.*[161] and, in general, news media[79, 83, 279] urge both academics and activists to scrutinize their use of social media data against a variety of possible misinformation, disinformation [161], impersonation [83]. In particular, it is visible in the movements and campaigns with highly contrasted stances about the issue being discussed, such as, Tahrir uprising in Egypt, Brazil “come to the streets” protests, Gezy park protests, Maydan Square revolution in Kiev. Multiple works strive to profile and identify incredible information rely on meticulous feature engineering [38, 46, 47, 134, 141, 161, 218] or crowdsourcing [198]. On the other hand, it has also been shown that there is a correlation between the content and the source credibility [275, 305], thereby providing a more accurate profile for the misinformation.

Many challenges exist with respect to the filtering of the spam information, including emails [18, 35, 43, 223, 319] and social media posts [83, 114, 277, 293, 325] - yet some type of spam is easier to detect than other. Elimination of the spam information is often considered to be similar to one of the bot sources as described in the book of Castillo [44]; however, information coming from the latter might have more social, political and economic value. In the works of Ferrara *et al.*[83] as well as overall in the field, the problem of separating non-human from human accounts has gained interest recently, especially, after the recent elections in the US.

Document Categorization and Compression: Noise and content redundancy (lexical and semantic) often account for a significant fraction of the web content [41]. The vast amount of data that platforms collect about users are monitored for its further summarization, categorization and content recommendation. Data is usually the only currency that the platform possesses and therefore, platform’s size and its content base are essential for both activist (to reach high volumes of people) and platform providers (to gain as much profit as possible on ads), and the produced content has to be analysed, categorized and summarized. Most works on the summarization and compression of the large document collections focus on syntactic [84, 85, 86, 87, 142, 172, 228] or semantic [11, 36, 37] representations that allow to balance between high precision and recall respectively. Another area of content representation develops around knowledge base use for data mining [61, 130, 257]. We note that analysis, methods and challenges of knowledge base construction [232] are outside the scope of this Chapter. Li *et al.*[177] and Li *et al.*[176] present

| <i>Approach</i> | <i>Short Description</i> |
|---|--|
| ... Relevance Filtering | |
| Spam Detection | |
| Statistical spam filtering. SVM, KNN. | [223] describes initial probes to detect spam using Naive Bayes and rule induction. [18] further compares various techniques to pre-process spam related keywords and their use with Naive Bayes classifier. [319] extends upon statistical approaches and compares various machine learning algorithms to detect whether an e-mail is a spam or not. |
| SMTP, Ontology, Feature engineering. | [35] goes beyond keyword based spam detection by adding features related to the source of the message, HTML structure, user feedback, [43] summarizes both SMTP, ontology based, machine learning techniques in combination with security and privacy concerns. |
| Social Media Spam User Classification | [293] identifies various types of users on social media platform, among them are spammers. [325] describes a supervised approach towards spammers detection using message and social behaviour features. [277] explores feature based spam users, messages and campaigns detection on various social media. [114] tackles the problem of credibility of the information in tweets using message and source feature extraction, pseudo relevance feedback and SVM regression. Development and public exposure to the chat and spam bots in recent years lead to discourse on its detection and elimination [83]. |
| Probabilistic deduplication. | [248] uses LSH to resolve similar tweets into the same cluster/machine and further enhance the detection by the time locality of the messages. |
| Hoax and Fake Information | |
| Feature based credibility analysis. | [46, 47] explore message, user, topic and information propagation features verified by crowdsourcing. [134, 218] statistically analyse the importance of the features used for the credibility analysis. [141] presents a study of 2016 US election rumours diffusion on social media. |
| Multiple | [38] explores the combinations of various tweet and user features on the fake messages (with image) classification task. [282] models hoaxes as viruses using epidemiological framework with three user states and four model parameters: spreading rate, gullibility, probability to verify the hoax and forget users current believe. [198] proposes a taxonomy of rumour tweets and verifies it by the crowd. [143] uses discriminative modelling to fuse and resolve the information from various data sources. [161] analyses the nature of Wikipedia hoaxes as well as the characteristics of the successful ones. |
| ... Document categorization | |
| Document Summarization | |
| LDA | [11, 36, 37] describe usages and advances of the Latent Dirichlet allocation. |
| Syntactic summarization | [142] proposes a text filtering based on the contextual, syntactic and statistical features. [84, 85, 86, 87] explore the use of NLP techniques (dependency trees), to summarize various types of content, e.g., sentences, video, financial news. [228] summarizes existing heuristic, compression and memory based approaches and advertises the latest (parse tree based). |
| Edit distance. Knowledge bases. | [257] describes the documents as connected graphs which are further used to summarize their contents. [61] replies on the knowledge bases to measure the semantic similarity between texts. [130] identifies features that are to be described and extracts user's opinion about these predefined features. [172] describes how edit distance and graph pruning are used for a meme tracking. |
| Email categorization | |
| Regression. Templatization. Clustering. | [71] emphasizes the role of email reading/filtering automation to avoid information overload. [3, 32, 110, 155, 160, 179, 302] describe how to identify important and topical messages in the user's email inbox by applying various methods. e.g, templatization, regression and statistical analysis. |
| Graph construction. Text and template matching. | [173, 299] describe how text matching can be applied to identify message threads. [8] presents a novel problem of recovery of the causal threads, i.e., emails that belong to the same event or template (purchase history). |

Table 2.5 – Overview of the most prominent approaches for document filtering and summarization.

some of the most prominent paper on efficient documents labelling and topic modelling of the short texts respectively.

Each tool and platform that is used by the activists contains its own package of actions that can be performed and are encouraged, and, consequently, some actions are harder to perform and might not be even supported. For example, emails, spreadsheets, and other productivity tools are still widely used among activists due to their flexibility. Thereby, another line of research points to the importance of properly organizing and structuring inbox and outbox email collections of the users [3, 8, 32, 71, 110, 155, 160, 179, 299, 302]. In particular, some works highlight the approaches to emphasize important information in the email context [32, 155, 179], i.e., important messages, important snippets of the message body. Since a considerable fraction of the produced emails is very similar to each other, templating and on the fly suggestions of the messages [147, 234] have started to gain traction in the literature.

Detecting an event or a topic on the web is challenging and might require domain adaptation. In particular, this task is essential for the activists to retrieve a particular information (location or presence of a protest, attack) from a large stream of generated content in an efficient manner.

Event Detection: Variety of methods cover general and specific topic or event detection and usually focus on the following methods: some form of clustering [2, 14, 74, 76, 98, 148, 159, 199, 221, 297, 320], burst detection [119, 170, 200, 215, 303, 322], similarity rankings [139, 152, 165, 318], feature engineering or lexicons [29, 159, 180, 247, 256, 298]. Most prominent works in the domain are the following: survey by Weiler *et al.*[300] and guides through the experimental analysis of the major event detection approaches. Moreover, works of He *et al.*[119], Schubert *et al.*[261], Guille and Favre [111], and Xie *et al.*[309] are recommended for those starting the exploration of the domain of event detection. In particular, on-demand extraction of online content based on a seed document or query is challenging and typically requires large amounts of annotated data to build supervised models [9, 29, 70, 144, 247, 298]. In some cases, the query is not known a priori and is only implicitly represented through a set of documents that are relevant to a topic of interest [152, 165, 177]. Similarity-based approaches tend to be inefficient [66] and difficult to scale. Another method to tackle topical document detection is to rely on content clustering and topic modelling (see Table 2.6). However, these approaches work best for document extraction relating to past events (thus, specific details are known and can be used for the extraction) and are hard to adapt to a stream processing context (where neither particular details nor dates are known ahead of time). A number of techniques leverage a lexicon that can efficiently and accurately represent a given topic, yielding a high precision but a rather low recall. For instance, [217] uses pseudo-relevance feedback to improve recall for the lexicon-based methods, which however hampers their capacity to detect new events [244]. Finally, a range of new deep learning architectures have been recently proposed to both represent the document in a semantic space as well as classify documents by topics based on their vector space representation [152, 165, 177]. Such methods are supervised and require a large corpus of annotated data.

| Approach | Short Description |
|---|--|
| ... Event detection | |
| ... Retrospective Event Detection | |
| Feature engineering. Geo features. | [29, 247, 256, 298] represent both the input stream messages or their clusters through a variety of features, e.g., terms frequencies and weights, topicality, skewness, timeliness, periodicity, keyword position, context. [211] further stratifies the events into sub-events based on four main features, e.g., content, temporal, diffusion degree and sensitivity. [7] discovers that NLP based constructed lexicons work best for the specific topics. [180] estimates the importance of classified tweets for a particular event. [159] adds information about the geographical activity within each voronoi diagram space. [255] iteratively selects phrases to track a particular topic and, thus, improves the extraction over time. |
| Content clustering | [221, 300] survey various cluster based techniques to identify events on the web and on social media. [148, 320] cluster co-occurring words to identify the event. [14, 297], first, cluster semantically close tweets and then extract event specific features from the clusters. [2, 76, 159] use geo clustering of the bursty terms which are further scored. [199] describes a production valid system based on the message clustering for the event detection. [119] cluster keywords based on their spectral representation using Kullback–Leibler divergence. [98] uses LSH to make document clustering mode efficient, further, each bucket is checked on the manually selected burstiness threshold, while keywords are used to identify the type of the event. [74] presents another class of tweet clustering using LDA based on tweet proximity and source of the message. |
| Similarity-based ranking | [165] compares various similarity metrics based on document vector representations and shows that averaging the word embeddings in a document leads to underestimating the similarity between documents. A better measure of similarity is defined as Word Mover’s Distance that is explored further in [152]. [318] explores various embedding estimations of the queries for a specific topic extraction. [139] utilizes Web-click graphs to rank documents for a given query. |
| Burst Analysis | [170, 200, 215] describes main categories of the event patterns on social media represented by a set of keywords that burst over a particular threshold. [119, 303] use spectral and wavelet transformation of the keywords frequencies to further cluster and filter possible events. Similarly, [261] traces exponentially weighted moving average of the bucketed terms. [322] treats burstiness of a keyword as a probability of generating related document in a time slot. [215] composes event specific lexicons based on the specificity and relevance of the n-grams. [175, 200] track the burst of the expanded queries (or segments) using pseudo-relevance feedback. |
| BiNets | [94] uses clusters of the interconnected bursty elements/features to identify the event. |
| ... Incremental Update summarization | |
| Incremental Update Summarization | [195] extracts keyword contexts to analyse the development of a particular event. The following is the list of method that assume that the stream of relevant documents is given or extracted with a set of specified keywords. [22] describes the general task of temporal summarization with a particular interest in novelty, prevalence and timeliness. [197] relies on the set of prevalence, novelty and quality features of the ranked summary updates to determine the most optimal cut off for the ranked list of summaries. [197] models an expected and incremental precision of the summaries as relevance and novelty respectively and selects sentences with the better characteristics. [149] uses basic, query, language model, geographic and temporal relevance features to predict the salience of the update integrated with the affinity propagation clustering [90]. [253] proposes to use integer linear programming techniques to optimize the summary coverage of the content words. |
| ... New Event Detection | |
| Clustering, anomalies | [244] proposes an accurate open domain event extraction pipeline that gathers named entity, event phrase (CRF), date, and type (LinkLDA). [192] presents an algorithm for automatic peak detection and annotation that are further to be examined by humans. [151] studies geotagged volume of tweets and hashtags using burst detection. [286] uses semantic analyses and ontologies to detect complex events with a high precision. [227] proposes an efficient LSH based heuristic to detect first story. [210] uses 3 main features, e.g. occurrence, diffusion, sensitivity, represented with FFT and evaluated on two conditions, presence and decay. [133] explores users profiles and interests to trace specific events. [176] uses auxiliary word embeddings to model topic distributions in short texts. [250] relies on non-parametric distributional clustering to infer topical infection of the users in information cascades. [226] uses LDA to infer a central topic model that is further enhanced with a two-phase random walk, thus allowing to accurately model even-specific topics. |
| Classification, Poisson event model | [9] presents the event detection problem as multi-task learning and proposes an optimization model that utilizes tweet content and event category relation. [70] relies on non-parametric topic modelling within time epochs to track semantically consistent topics and models an event arrival as a Poisson process with (non) bursty periods. [144] explores linear models with a rank constraint and a fast loss approximation and shows that they perform on par with deep learning classifiers. |
| Dataless Text Classification | [177] learns to extract relevant documents based on a small seed of related keywords by exploiting explicit word co-occurrence patterns between the seed words and regular words. Similar extraction techniques leveraging lexicon expansion are described in [217]. [54, 184] analyze the extent to which query words can be used to represent a topic of interest for further extraction. [273] shows how a semantic representation of a query and a document allow to measure the similarity between the two. |
| Anomaly detection | Information summarization can be represented as an anomaly detection tasks that identifies not common/new patterns in any information sources, e.g., time series, texts, geo location, graphs [51, 111, 185, 309, 315]. |

Table 2.6 – Overview of the most prominent approaches and applications of event detection, summarization and tracking.

New Event Detection: Extracting unseen event instances on a particular topic usually poses multiple obstacles to the methods used for the retrospective topic detection. In particular, feature engineering, geo or lexicon-based approaches are hard to generalize for the unseen events, while clustering and similarity based technique are highly dependent on the similarity measure and might not generalize well across other topics. Such similarity measures are usually rather expensive to compute on large corpora and thus impractical [152]. Anomaly [51, 185, 315] or burst [151, 192, 210, 226, 227, 244, 250] detection is often used to detect new events as the method's sensitivity can be adjusted to capture new trending topics or events. However, such techniques might not be accurate and flexible enough. Despite the fact that classification usually requires an enormous amount of training data, it is used for retrieving the relevant information [9, 70, 144]. Finally, some studies explored the ways to reduce amount of training data needed to retrieve new relevant events [54, 177, 184, 273] – so-called dataless classification.

Event Summarization and Tracking: Event summarization and tracking differ from the previous tasks since those take a stream of pre-filtered topically related texts (Table 2.5) as their input. The general task of temporal summarization is described in Aslam *et al.*[22]. Further, a variety of techniques were proposed to tackle this problem, such as keyword tracking, feature extraction, linear optimization [113, 197, 253].

2.4 Positioning

In summary, this thesis provides a context of online activism from the two main perspectives: (1) exploration and analysis of the user engagement in online collective actions and (2) methods to facilitate content filtering and repetitive content creation.

For the former, recent studies on the coverage of debates, protests, and campaigns do not reveal how they develop both in the online and offline setting. In particular, some works [58, 81, 158] examine a campaign from the traces it leaves online, while we also try to differentiate offline actions that are performed by the activists. We further explore types of campaigns and how popular they are based on the messages they contain. As users increasingly tend to rely on their social entourage to filter information [122], we examine how different message types and techniques engage people in a variety of ways throughout the campaign. Thus, we focus on the classification and analysis of the campaign content with further insights on user involvement.

Since user engagement in a campaign might differ significantly, depending on the raised issue, locality, organization, etc., we narrow down our study to the context of the online petitions. In this context, the activist precisely specifies the preferred number of involved users and thus, a petition succeeds when the number of collected signatures is higher than a particular goal. Unlike other works that study online petitions, we focus on signature rate dynamics using co-evolving time series information. Our approach is based on modelling the conditional mean of a Hawkes process, where we extend the model with a more flexible aging, i.e., raise and decay, and include both internal dynamics (self-excitation) and external factors (social network, front page effect).

Moreover, each external factor is modeled as a continuous effect on the signature dynamics, rather than a series of single external shocks. To the best of our knowledge, we are the first to present a model that captures the interaction between multiple platforms in a model-based framework and with easily interpretable parameters.

For the latter, the work on repetitive content templating that is presented here aims to provide assistance that is learned correctly from an individual email sender, whereas other solutions provide “canned” responses meant to satisfy as many users as possible based on a “global” model. Our work also differs from the conventional auto-completion studies that tried to infer the intent where we emphasize on the time-efficient generation of a template based on the content produced by the same sender. Importantly, the proposed method obtains deterministic and accurate results with a linear complexity.

Considering the vast amount of information produced on social media and the web, filtering of the relevant information, such as texts about the events or other complex heterogeneous topics, is a challenging and crucial endeavor. Contrary to the variety of methods to do this task described in Table 2.6, such as classification methods that require a substantial amount of annotated data, or methods are based on query similarity, that require pairwise similarity comparison between each query text and input data, we propose a method that is more efficient and accurate, as well as require a small training dataset.

Profiling Part

3 Profiling Large-Scale Public Campaigns on Twitter

Social media has become an important instrument for running various types of public campaigns and mobilizing people. Yet, the dynamics of public campaigns on social networking platforms still remain largely unexplored. In this thesis, we present an in-depth analysis of over one hundred large-scale campaigns on social media platforms covering more than 6 years. In particular, we focus on campaigns related to *climate change and animal welfare*¹ on Twitter, which promote online activism to encourage, educate, and motivate people to react to the various issues raised by climate change. We propose a generic framework to identify both the type of a given campaign as well as the various actions undertaken throughout its lifespan: official meetings, physical actions, calls for action, publications on climate related research, etc. We study whether the type of a campaign is correlated to the actions undertaken and how these actions influence the flow of the campaign. Leveraging more than one hundred different campaigns, we build a model capable of accurately predicting the presence of individual actions in tweets. Finally, we explore the influence of active users on the overall campaign flow.

3.1 Introduction

Social media have become central to our digital lives, as they allow individuals to share news, photos, or opinions, as well as to have online discussions in real-time. One particularly interesting phenomenon is social media marketing, which can be defined as the process of drawing attention to some specific issue or product via social media platforms. Such endeavors often take the form of extensive campaigns, whose aim is to raise the awareness of the public on a particular topic and potentially to engage it into concrete actions.

Social media platforms provide tools to effectively conduct these campaigns; On Twitter, for example, people use so-called *hashtags* to associate their messages to a certain topic. Many campaigns, therefore, have their own hashtags that uniquely identify them. Moreover, many tweets associated with a campaign convey some specific messages to the audience, such as

¹From now on we will refer to the topic of climate change and animal welfare as *climate change*.

requests for signing a petition, asking for a certain action, attending a demonstration, etc. These messages can be considered as certain *actions*, and their effect on the dynamics of the campaigns remains largely unexplored in the scientific literature. Identifying and categorizing such messages within the context of a campaign would enable us to answer questions such as what drives attention to a particular topic or how to reach a certain target audience. In this work, we propose a number of categories to classify the actions from the perspective of the goals of the campaigns as well as a methodology to identify them. We build a classifier for the action types based on the tweets content and study the distribution of these action types for different types of campaigns.

In the second part of this work, we analyze the resulting user involvement patterns in order to explore the dynamics of the campaigns. Analyzing such patterns is key to understand how attractive the campaigns are and who are the main contributors to the information dissemination. We perform a comparative analysis of the campaigns and their contents, through which we identify noticeable differences between the various types of campaigns. We observe that campaigns where only a tiny fraction of users create the major part of the content are less likely to attract users on social media. Finally, we cluster the involvement patterns and study their correlations with the types of campaigns. For instance, campaigns with a precisely defined goal use more calls for actions and mobilization messages than official meetings, e.g. animal welfare campaign #helpcovedolphins was trying to involve as many people as possible during the initial period to save dolphins in Japan.

This work focuses on campaigns related to climate change and animal welfare (referred as climate change later on). Those two topics recently gained increased attention and have the advantage of gathering a high number of users for relatively long periods of time, thus are well suited for our study. Moreover, these topics are mainly of interest for non-profit and governmental organizations, and this work might help them to better understand the impact of their actions on the audience.

We consider a campaign in its online form on Twitter as messages associated with a certain hashtag that is mainly promoted by a few users or activists [95], [306]. Our analysis is based on the current understanding of the activism directed towards making impact on changing climate. For the purpose of this study, by the *activism on climate change* we mean the active involvement of certain people or organizations in promoting ideas, actions, information on Twitter about the climate change ².

In summary, the main contribution of this chapter is a large-scale study of the dynamics of campaigns on social media.

This study focuses on the following research questions:

- How to identify and compare various types of public campaigns and their corresponding

²Climate change is a complex problem... It either impacts on - or is impacted by - global issues, including poverty, economic development, population growth, sustainable development and resource management. http://unfccc.int/essential_background/items/6031.php accessed April 2014

actions? (Section 3.4)

- How is the initial goal and content of a campaign correlated to the user engagement pattern of a campaign? (Section 3.4.2);
- Is there a relationship between the type of a campaign and the actions undertaken through the campaign lifespan? (Section 3.5);

Our key insights and findings:

- Various campaigns differ by the actions, i.e., some of the user engagement campaigns utilize more physical actions while other mainly post factual information. Later showed the correlation with more long term campaigns with ever-growing user engagement.
- Popular as well as mobilization campaigns contain a lot of duplicate or near-duplicate content.
- First degree neighbours of the users that post about the campaign are essential for getting retweets, while duplicate content attracts less retweets.
- The less diverse the main contributors of the campaign, the less likely it is to gain bigger audience.

The rest of this chapter is structured as follows. We start with an overview of related work in the areas of Twitter analytics and social media analysis below in Section 3.2. Section 3.3 describes our data collection, aggregation, and cleansing processes. We analyze the collected data in Section 3.4 by extracting different types of campaigns and clustering them by their user engagement patterns. Section 3.5 extends our analysis by focusing on various types of tweets and on their distributions in campaigns, as well as building a classifier that is able to predict the type of a tweet. Finally, we discuss the obtained results in Section 3.6 and draw conclusions in Section 3.6.

3.2 Related Work

Social media platforms quickly came to the attention of the research community, since they allow to conduct large-scale studies on various aspects of social network dynamics, such as popularity prediction. Many studies have recently focused on the data from micro-blogging platforms such as Twitter³, which provides an access to a small (compared to the overall data) sample of its data based on keyword queries.

In this work, we study the communication patterns and message type preeminence for various campaigns on climate change. A number of studies have focused on Twitter communication patterns, including studies on hashtag life-cycles [169, 212], event detection and their analysis [137, 215], food consumption patterns [1], and usage across different languages [125].

Climate change discourse. Climate change issues receive increased attention as they lead to a

³<https://twitter.com>

number of global challenges [12, 128]. Many studies recently examined how the climate change debate is covered on social media channels [215, 260]. However, coverage of debates does not reveal how campaigns develop, and how popular they are based on the messages they contain. As users tend to increasingly rely on their social entourage to filter information [122], we examine in this chapter how different message types and techniques engage people in different ways throughout the campaign.

Campaign analytics Social media is a very influential tool for widening public awareness on various issues as noted by [270]. Previous work on campaigns on social media mostly focus on political and protest campaigns. Tumasjan *et al.*[292], for example, tackled the problem of predicting elections based on sentiment analysis of large sets of tweets, and [181] studied the dynamics of emerging hashtags during 2012 US presidential elections. Jin *et al.*[140] used a bispase model based on a Poisson process to capture the propagation of information in both Twitter and non-Twitter environments. Additionally, Gonzalez *et al.*[103] explores how social networks are used to spread the protest information. In our work we focus not on the information dissemination but rather how the campaigns were conducted and what are the main actions that were taken to reach the goal. Finally, one of the most recent works on campaign analysis [81] focuses on the behavioural stage sequences of the users during the COP21 and EH2015 forums and proposes a framework to identify a user stage by her tweets. On the contrary, we focus on the campaign actions and the corresponding users' engagement rather than user behavioural stages. Moreover, our analysis is carried out on over a hundred public campaigns.

Tweet topic identification In the context of topic identification, recent works focused on classifying and clustering tweets based on their topics [31, 168, 239]. Those techniques produce different sets of topics for different datasets. In our case, however, such approaches did not result in valid clusters of message types. To the best of our knowledge, this work is the first on tweet type classification in campaigns. An extensive study on the theoretical principles underpinning public communication campaigns is described by [23]. The work of Segerberg and Bennett [263] was an important motivation for the definition of further types of tweets, such as official meetings, calls for action and physical actions. Given our objective of comparing campaign agendas, we look into certain types of campaigns and actions in this chapter in order to identify the correlation between the types of campaigns and the different actions.

Tweet and hashtag popularity By far the most widely researched topic on Twitter is predicting the popularity of both messages (tweets) and hashtags (trends), which link different messages to a single theme. A number of works in this domain tackled the popularity prediction problem using regression models [26, 158], classification [126, 187] and time series modelling [158]. However, in this work we focus on the classification and analysis of the campaign content with further insights on user involvement. For regression and classification models, previous solutions mainly focused on identifying and exploring effective features for popularity prediction.

Hong *et al.* [126] modeled the problem of popularity prediction as a classification task with several classes specifying the number of retweets a message will receive. The most effective

predictive features they studied included the history of retweets and the number of followers for a given user. Kupavskii *et al.* [162] addressed the subproblem of predicting the number of retweets within a fixed period of time, and explored additional features based on the flow of the retweet cascade and the PageRank score derived from the retweet graph. Further works on this topic explored features based on structural characteristics of the networks created by early retweeters [26] and features based on sentiments extracted from the tweets [138]. Xu *et al.* [311] studied the reasons why users post messages on Twitter, and suggested a set of incentives in that context including breaking news, posts from friends and the intrinsic interests of the users, and proposed a mixture latent topic model to combine all these factors. Finally, Choi *et al.* [56] used retweet data to perform query expansion based on temporal information, in the sense that retweets of documents were used to boost the documents' relevance over a period of time.

Most of the pieces of work in this domain, however, try to predict the absolute tweet popularity, defined as the number of retweets [126], or a boolean specifying whether the number of retweets will be higher than a certain threshold [138]. We focus on a different issues in this piece of work, as we predict unusually high number of retweets for users participating to a campaign, which actually is a fairly complex task since most tweets receive a retweets number that is highly correlated with a number of followers. [92] - next generation work that tries to model process of retweeting via extended reinforced Poisson process model with time mapping process. Jenders *et al.* [138] addressed a tweet popularity from an angle of finding reasons why users retweet a tweet, predicting virality (whether it will be retweeted more than a certain threshold T), compared Generalized Linear Model and Naive Bayes, and used features such as the number of followers, tweet length, tweet sentiment, URLs, number of hashtags in a tweet. Li *et al.* [178] predicts popularity not only for tweets, but also for other forms of content on social media, such as videos.

The second well-researched topic in this context is hashtag popularity. As introduce above, hashtags identify topics that can be used to virtually regroup sets of messages. Cui *et al.* [69] proposed a temporal analysis of hashtags in order to discover breaking events in real-time and tried to distinguish the hashtags of social events from hashtags of virtual topics or memes. Ma *et al.* [187] modeled the popularity problem as a classification problem and used various content and context features—such as the number of tweets containing a certain hashtag during a time period and the number of unique users that post messages with a certain hashtag for the same time period—to make better predictions. However, not every hashtag on Twitter represents a valid campaign, while in this work we are particularly focusing on campaign analytics and apply a crowdsourcing pipeline to filter out non-campaign hashtags.

One recent trend is the analysis of hashtags with respect to geo-locational data. Glasgow *et al.* [100], for example, studied the lifespan of various hashtags during the 2011 riots in London while Kamath *et al.* [146] addressed the problem of predicting the popularity of hashtags in specific geographical locations using geo-spatial reinforcement learning models.

3.3 Data Collection and Cleansing

In this section, we first describe the process through which we collected tweets related to the domains of climate change and animal welfare (Section 3.3.1). We then introduce the strategy we took for identifying campaigns in those domains. Finally, we describe our process to identify the retweets and duplicated tweets in Section 3.3.2. The resulting dataset, consisting of more than 8.5M tweets, is available online for the future research.⁴

3.3.1 Twitter data collection

We developed a data collection pipeline (see Figure 3.1) to gather a broad range of Twitter campaigns related to climate change and animal welfare. Those two domains are usually tightly connected [268]. For example, there are multiple examples highlighted by Olteanu *et al.*[157] on the connection between climate change and animal welfare, i.e. the link between the number of farm animals and the amount of methane released to the atmosphere and thus causing climate change.

In order to achieve such a broad coverage, we started from a generic corpus comprising tweets that are highly related to climate change and that were downloaded using a set of key-phrases described in detail in [215] from Topsy⁵. Topsy was a primary partner of Twitter delivering search and analytic services and claiming to index all public tweets. The timespan of our corpus ranges from the beginning of 2009 to the beginning of 2015 and resulted in over 10Gb data containing more than 55M tweets.

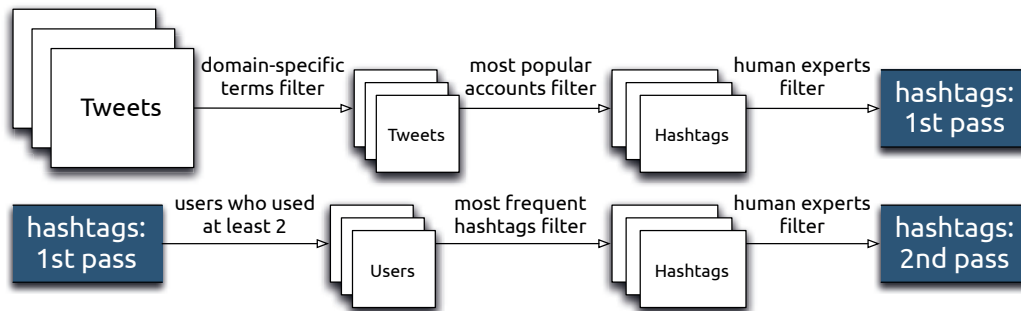


Figure 3.1 – Data collection pipeline for the profiling of the public campaigns.

First pass We proceeded in two phases in order to identify the campaign. In the first pass, we extracted all available tweets from Topsy⁶ for two very prominent accounts that are related to climate change and animal welfare-related: @AlGore and @GreenPeace (2.77M and 1.33M

⁴<https://github.com/toluoll/CampaignsDataRelease>

⁵<http://topsy.com/>

⁶Topsy (<http://topsy.com/>) is a primary partner of Twitter delivering search and analytic services and claiming to index all public tweets.

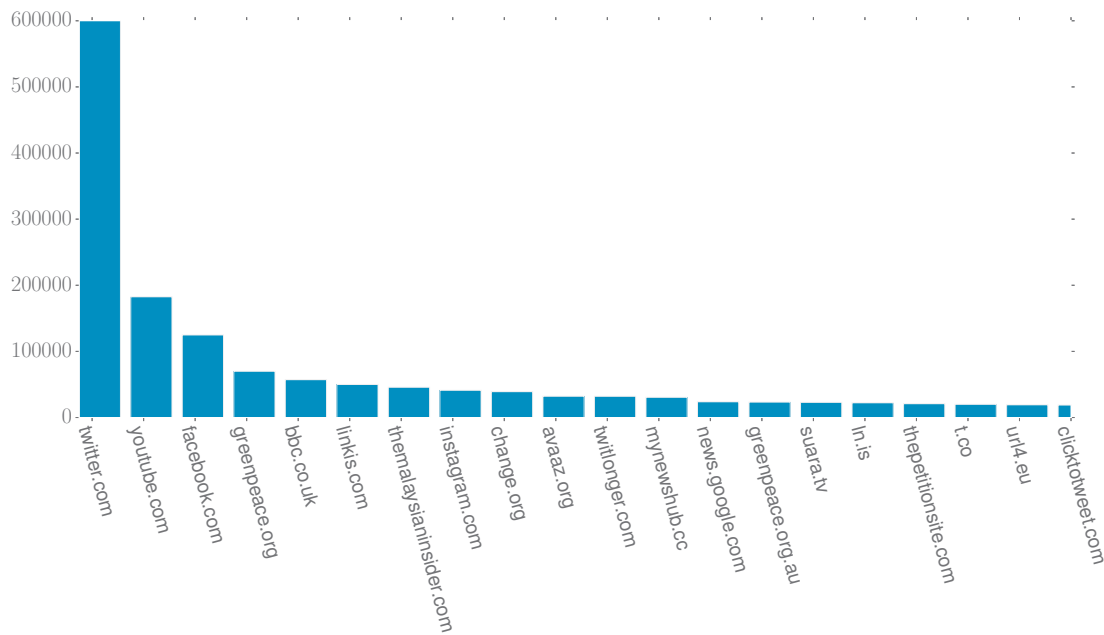


Figure 3.2 – Top-20 domains used in the climate change campaigns for the original tweets.

followers respectively). This first pass resulted in 27K tweets comprising 1250 unique hashtags. Over 50% of these hashtags either occurred in less than 50 tweets or had a single narrow peak over the whole timespan. Single peak means that the hashtag was spanning for less than a week. Cleaning of the infrequent or time-short hashtags resulted in the set of around 612 *candidate campaigns*, including multiple false positives that required to be resolved manually. To select valid campaign hashtags out of these 612 hashtags, we decided to rely on the annotations made by the author of this thesis and two of her colleagues. A hashtag was considered to be related to a campaign *iff* it was selected by all of the annotators. This manual annotation produced a set of 52 campaign hashtags.

Second pass To increase the recall of our process, we run a second pass. We identified further accounts (users) that mentioned at least two campaign tags (out of the first 52) in their messages. In that way, we identified 80 additional accounts for a total of 34K unique hashtags. We filtered out hashtags that appeared in less than 50 tweets, which accounted for 75% of the tweets. Similarly to the first pass, the author of this thesis and two of her colleagues annotated each resulting hashtag and 56 additional hashtags were identified as relevant. Overall, our process resulted in a dataset of 108 climate and animal welfare-related campaigns⁷, each represented by a distinct Twitter hashtag. The total number of unique tweets (without retweets) in the resulting dataset amounts to 4M.

⁷It is worth noticing that many of the hashtags (around 20 each) in our campaign dataset are created using the morphological filters. For example, we collected hashtags that contain words such as save, protect, call, lead, act, 4, forthe, etc. (e.g #savethedolphins, #call4action).

3.3.2 Unique tweets identification and retweets count

One of the main issues with the data collected from Topsy is that the tool does not provide information about retweets. Therefore, we had to create a heuristic to make sure that we could properly identify all retweeted messages. Taking into account that all tweets returned by the tool are sorted by timestamp, we can easily figure out the origin of all the tweets using a simple regex pattern `((RT|MT) @author tweet_prefix)`. First, it does not identify complex retweet structures, such as where a tweet text is cited using quotes. We found that such cases are quite rare on Twitter and amount for $\sim 0.5\%$ of all tweets.⁸ However, certain retweets could be missing when a hashtag does not fit into the message due to the tweet length limit. To solve this problem, we leveraged the Topsy API, by returning and analyzing related tweets for each requested tweet in order to identify all further retweets. Finally, we note that we apply this process recursively—searching for retweets of retweets iteratively—in order to capture potentially complex retweet patterns.

When no new retweets can be identified, we identify content that was not retweeted but duplicated. The practice of duplicating tweets gained traction on the platform as it can help promote topics into *Twitter Trends*. We consider a tweet to be a duplicate whenever at least 80% of its contents exactly matches an original tweet excluding punctuation. As for the retweets, we sort the original tweets by date, analyze their contents and cluster similar tweets that share a high degree of overlap.

3.3.3 URL usage statistics

We unshortened and clustered by domain names all URLs appearing in the original tweets, i.e., the tweets that were not identified as retweets. In total, over 3M distinct URLs were identified. Figure 3.2 gives the 66th percentile of the most frequent domain names. Not surprisingly, the most frequently used domain is `twitter.com`, and accounts for about 20% of all URLs, out of which 97% were images or photos. Social media, such as YouTube and Facebook, account for around 5% each. Over 2.5% of the URLs belong to “petition” websites, such as `avaaz.com`, `thepetitionsite.com` and `change.org`. The rest of the URLs cover specific sites that correspond to some of the campaigns or to news aggregators.

3.4 Campaign analysis

Given that our research question connects two domains—climate change / animal welfare campaigns and social media content analysis—the framework we propose for campaign analysis is composed of two parts. First, we annotate the campaigns according to their primary goals. Next, we cluster them by examining user engagement patterns and by mining active users for

⁸In order to compute the complex retweet cases we aggregated the obtained tweets collection with at most 5 characters edit distance. Further, we have discarded explicit retweets `((RT|MT) @author)` and exact duplicates which resulted in 0.5% tweets on average to be cited.

the campaign (i.e., users who tweet most often for a particular campaign). When organizing our data and constructing the annotation process, we turned to dimensions considered in the theory of public communication campaigns [23, 154, 263]. For each campaign, we consider the major goal of the campaign (increase awareness, mobilize people), user engagement over time (ever-growing, regular, one-day, inactive), as well as user activity.

3.4.1 Types of campaigns

Following the theoretical analysis of public communication campaigns by [23], we separate the campaigns into two classes based on their primary goals:

- *Mobilization campaigns* refer to the campaigns whose primary goal is to engage and motivate a wide range of partners, allies and individual at the national and local levels, towards a particular problem or issue (e.g., #protectthearticrefuge, #endsharkcull, #freetheartic30, #savebucky, #freelolita, #takeaction, etc.).
- *Awareness campaigns* refer to the campaigns whose primary goal is to raise people's awareness regarding a particular subject, issue, or situation. As discussed in Section 3.2, environmental awareness campaigns usually make a large use of mass media, and in particular, of Twitter (e.g., #lifentextinction, #noseaworldq102, #leadonenergy, #blackle, #tweet4elephants, etc.).

These campaign types represent very different endeavors, which affects both the type of contents used in such campaigns as well as their user involvement pattern over time, which we analyze further.

Table 3.1 shown a full list of campaigns with corresponding types.

The author of this thesis and two of her colleagues manually annotated the campaigns as either mobilization or awareness campaigns. The category was considered as valid only when all experts agreed on it. This way, 50 awareness and 58 mobilization campaigns were identified. A few sample hashtags are #savesolar, #climateaction for **mobilization** and #cleanair4kids, #worldfoodday for **awareness** campaigns.

3.4.2 User engagement patterns

In the following, we present an analysis of user engagement in Twitter campaigns. We identify two main axis for analyzing user engagement: the first one focuses on user engagement patterns over time, while the second one analyzes the behaviour of the most active users throughout the campaign.

| Hashtag | G | U | Gini | Hashtag | G | U | Gini |
|---------------------|---|-------|------|--------------------------|---|-------|------|
| #action2015 | a | succ | 0.49 | #protectparadise | m | one | 0.40 |
| #action4climate | a | multi | 0.52 | #protecttheartcticrefuge | m | one | 0.26 |
| #action4dolphins | m | multi | 0.81 | #rescuesaama | m | one | 0.82 |
| #actonclimate | m | succ | 0.59 | #saveafricananimals | m | one | 0.88 |
| #animalwelfare | a | succ | 0.68 | #saveanimals | m | succ | 0.52 |
| #askdrh | m | multi | 0.46 | #savebucky | m | one | 0.78 |
| #backclimateaction | m | one | 0.50 | #saveenergy | m | succ | 0.52 |
| #banfoiegras | m | multi | 0.84 | #savefaroewhales | m | multi | 0.81 |
| #blackle | a | non | 0.42 | #savefukuchildren | m | non | 0.82 |
| #c4climate | a | non | 0.75 | #saveourwater | m | one | 0.50 |
| #captivitykills | a | succ | 0.70 | #saverareelephants | m | one | 0.75 |
| #changetheworld | a | succ | 0.37 | #savescotlandseals | m | one | 0.53 |
| #cleanair4kids | a | multi | 0.66 | #savesharks | m | succ | 0.66 |
| #climateaction | m | succ | 0.57 | #savesolar | m | one | 0.53 |
| #climatemarch | a | one | 0.46 | #savethearctic | m | multi | 0.54 |
| #climateweek | a | r | 0.48 | #savethebees | m | succ | 0.39 |
| #connect4climate | a | non | 0.33 | #savetheplanet | m | succ | 0.28 |
| #consciouscollege | a | multi | 0.73 | #savethereef | m | multi | 0.40 |
| #divestment | m | succ | 0.57 | #savetigers | m | one | 0.51 |
| #endsharkcull | m | one | 0.64 | #savewater | m | succ | 0.42 |
| #energy4all | a | multi | 0.64 | #soscovedolphins | a | one | 0.75 |
| #fastfortheclimate | m | r | 0.67 | #sport4climate | a | non | 0.49 |
| #fightforthereef | m | r | 0.51 | #standforforests | m | one | 0.30 |
| #forwardonclimate | a | one | 0.55 | #standupforthePacific | m | one | 0.53 |
| #fossilfree | m | succ | 0.56 | #stopicelandicwhaling | m | one | 0.75 |
| #freelolita | m | succ | 0.69 | #stopkillingdugongs | m | non | 0.77 |
| #freethearctic30 | a | one | 0.65 | #stopwildlifecrime | m | multi | 0.55 |
| #freetonytiger | m | r | 0.81 | #stopyulin2015 | m | one | 0.82 |
| #furfreefriday | a | succ | 0.78 | #storm4arturo | m | one | 0.87 |
| #gofossilfree | m | multi | 0.43 | #takeaction | m | succ | 0.41 |
| #greenu | a | non | 0.78 | #talkenergy | a | r | 0.77 |
| #grindstop | m | one | 0.78 | #talkfracking | a | multi | 0.61 |
| #helpcovedolphins | m | one | 0.79 | #talkpoverty | a | succ | 0.71 |
| #jpolyboycott | m | one | 0.90 | #tcktck | a | multi | 0.59 |
| #leardblockade | a | multi | 0.80 | #thinkeatsave | a | one | 0.47 |
| #lifenotextinction | a | non | 0.73 | #tweet4dolphins | m | succ | 0.85 |
| #marchforlolita | m | one | 0.84 | #tweet4elephants | a | multi | 0.48 |
| #noarcticoil | m | one | 0.40 | #tweet4taij | a | multi | 0.71 |
| #noseaworldq102 | a | one | 0.81 | #up4climate | a | one | 0.57 |
| #nowasharkcull | m | multi | 0.62 | #voices4climate | a | non | 0.61 |
| #oilspill | a | one | 0.75 | #vote4cleanpower | m | one | 0.48 |
| #opinfinitepatience | a | one | 0.68 | #votegreen2015 | m | succ | 0.65 |
| #opkillingbay | a | multi | 0.88 | #worldaid4dolphins | m | one | 0.64 |
| #opseaworld | a | succ | 0.83 | #worldaidfaroeislands | m | multi | 0.78 |
| #opstormfreearturo | a | one | 0.76 | #worlddelephantday | a | r | 0.39 |
| #peoplescoal | a | one | 0.33 | #worldenvironmentday | a | r | 0.41 |
| #protectcleanwater | m | multi | 0.58 | #worldfoodday | a | r | 0.35 |
| #protectgloucester | m | multi | 0.79 | #worldlovefordolphinsday | a | r | 0.76 |

Table 3.1 – The list of the campaigns annotated by both their goal and user engagement pattern. “G” corresponds to the categorization of campaigns by their goal (awareness or mobilization). “U” stands for the categorization of campaigns by the user engagement pattern. Gini - corresponds to the gini coefficient computed as explained in Section 3.4.2 Clarifications: *a* - awareness, *m* - mobilization, *succ* - ever-growing, *multi* - multi-burst, *non* - inactive, *r* - annual, *one* - one-day campaign respectively.

User engagement patterns over time

Subsequently, we cluster the campaigns by engagement patterns of their users to detect whether the engagement correlates with the campaign types. In order to do this, we first extract the number

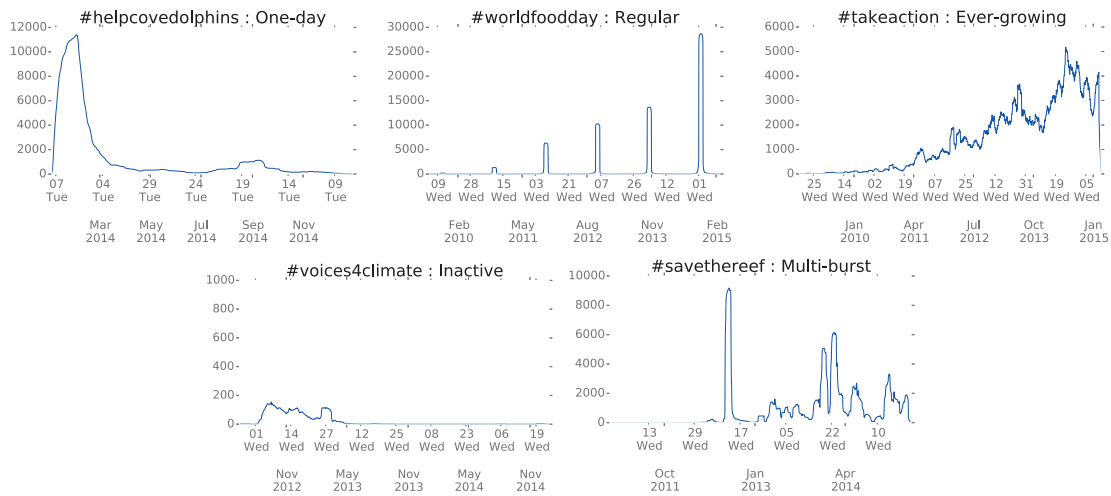


Figure 3.3 – Different user engagement patterns observed in campaigns.

of unique daily users for each campaign hashtag and aggregate these numbers with a 30-day time window. Then, we compute the similarities between the resulting time series using Dynamic Time Warping [96] and cluster them using K-means by varying the K and chose the setup with the smallest in-cluster distance. This resulted in five major clearly distinguishable clusters.

From our data collection through this process we have identified several major types of user involvement patterns. Following their overarching distribution, we name them:

- *one-day campaign*, a campaign that is organized over a short period of time to tackle some urgent issue;
- *regular campaign*, a campaign that happens on a regular basis, e.g., annually;
- *ever-growing campaign*, a campaign that gains traction over time;
- *multi-burst campaign*, a campaign that have multiple peaks of activity;
- *inactive campaign*, a campaign that shows a constantly low user engagement throughout its timespan⁹

Sample campaigns with the above described types are shown on Figure 3.3.

Finally, we compare the representations of two major classes of campaigns with their user involvement patterns. The campaigns are distributed across aforementioned engagement groups as follows: 36%, 10%, 21% 22%, 11%. As can be observed on Figure 3.4, most of “regular” and “inactive” campaigns fall in the awareness category, while both “one-day” and “ever-growing” campaigns are dominated by the mobilization category. The main reason for the dominance of mobilization campaigns for the “one-day” type is the urgency of their issues and the need for immediate action. On the other hand, “regular” campaigns, that are organized on a periodic basis and pursue long-term goals consist of awareness campaigns mostly. “Ever-growing” campaigns

⁹“Inactive” category might be orthogonal to the other ones, however, it gives valuable insights regarding campaigns that have less traction on Twitter

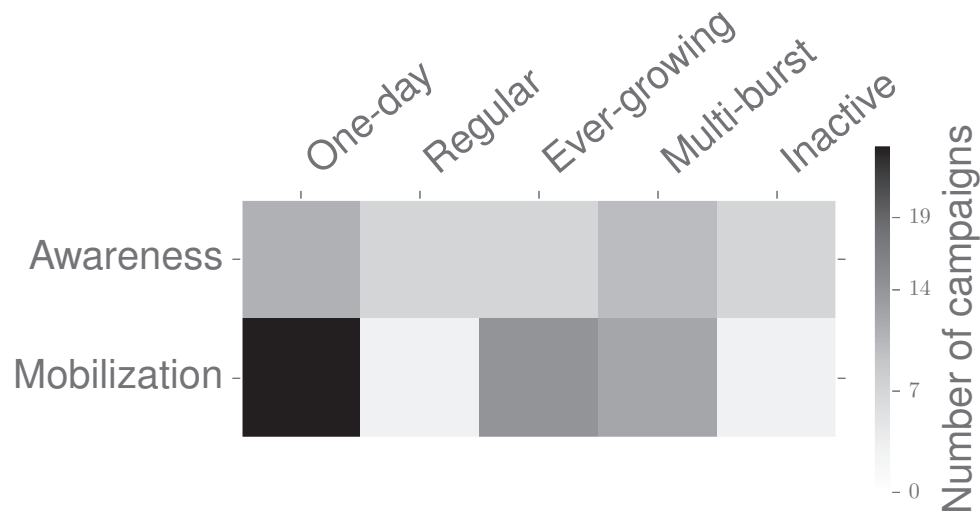


Figure 3.4 – Distribution of user engagement patterns for the different types of campaigns.

also dominated by mobilization campaigns and focus on global issues and challenges, e.g., #saveanimals, #animalwelfare etc. “Multi-burst” campaigns are almost equally represented by the both types.

User engagement patterns by volume

We observe that in many campaigns, there is a distinct subset of users who are authors of the majority of the campaign tweets. Interestingly, we have identified that apart from the overall user involvement in the campaign, it is important to explore main content creators of the campaigns and is there any distinction between campaigns in terms of this factor. We call such a set of users a *campaign kernel*. A *kernel* identifies users with the most tweets and retweets in the campaign. In order to study the influence of a kernel, we propose the following technique: (1) for each user, compute the total number of original tweets and retweets posted in the campaign; (2) rank all users relatively to the volume of content produced for the campaign; (3) compute the Gini coefficient¹⁰ based on the normalized per-user impact relative to the volume of messages in the campaign.

Figure 3.5 shows sample distributions of the relative amount of content generated by users participating in campaigns. As a result, we have discovered noticeable difference between type of the campaign and corresponding user kernel efficiency. We observe a clear distinction between campaigns where users are contributing the content almost equally (blue curve) and campaigns where only a tiny fraction of users create the major part of the content (red curve). By activism we mean the number of tweets posted and retweeted by the users normalized by the overall tweets in the campaign. Table 3.2 shows campaigns with the lowest and the highest Gini coefficient values.

¹⁰ http://wikipedia.org/wiki/Gini_coefficient

| Hashtag | Gini |
|--------------------------|------|
| #protecttheartcticrefuge | 0.26 |
| #savetheplanet | 0.28 |
| #standforforests | 0.30 |
| #connect4climate | 0.33 |
| #peoplescoal | 0.33 |
| ... | |
| #storm4arturo | 0.88 |
| #saveafricananimals | 0.88 |
| #opkillingbay | 0.88 |
| #jpolyboycott | 0.91 |
| #unity4malaysia | 0.99 |

Table 3.2 – Gini coefficients. Lower values correspond to almost equal user contribution, higher values represent campaigns where only a small fraction of users contribute.

Such values denote campaigns where the majority of the contents is created by few users only. Values that are close to zero, on the other hand, characterize campaigns where users contribute almost equally.

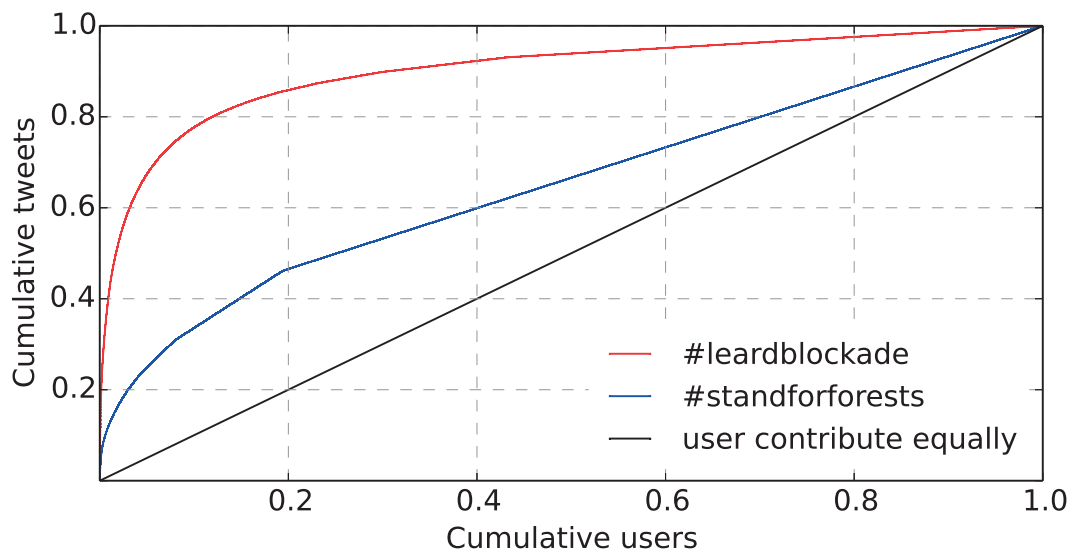



Figure 3.5 – Example of the active and inactive kernel involvement. #standforforest has gini 0.3, while #leardblockade - 0.80.

Interestingly, we found a direct correlation between the total number of followers of the kernel users and the total amount of users participating in a campaign. The value of Pearson correlation coefficient for these variables is more than 0.85. This behaviour is observed for various kernel sizes (that correspond to various proportions of total tweets they generated). The distribution of Pearson correlation coefficient between the number of kernel followers and the number of users engaged in the campaign for different kernel sizes has the shape of a bell curve ,

where thresholds for proportions of tweets are displayed on the x axes (50% - 99%), and Pearson correlation is on the y axes (0.70 - 0.858). The maximum correlation is reached for the kernels that produced 75% of the content and on average this corresponds to the 2.5% of the campaign users. Thus, we use this percentage of users further as a kernel of each campaign. Interestingly, we found no clear distinction between awareness and mobilization campaigns with regard to their kernels. However, the activity of kernel users differs with respect to the user engagement patterns described in 3.4.2. First, we observe that the majority of content in inactive campaigns is produced by a tiny fraction of users, while regular and ever-growing campaigns accumulate tweets from a much larger subset of users. Similarly, we observe that the kernels of inactive campaigns are 10% smaller than one-day campaigns, while ever-growing campaigns have both small kernels and high participation of the involved users.

3.5 Tweet Type Identification and Classification

This section presents an in-depth analysis of the tweets from our dataset, focusing on the types of actions they contain (Section 3.5.1), their correlations with the campaign types (Section 3.5.2). We collected additional information about the tweets through a large-scale crowdsourcing experiment (Section 3.5.1), in order to collect enough annotations to build a supervised model capable of accurately predicting the type of action contained in a given tweet (Section 3.5.1).

3.5.1 Types of tweets

An action in our context correspond to generic activity that is integral part of the campaigns agenda (meetings, protests, advertisings, events and so on) and intended to help a campaign to reach its goals. In this Chapter, we are aiming to identify high level categorization of the typical campaigns' actions. As discussed in Section 3.2, the campaigns that are the most effective at influencing users are typically related to either promoting some positive behaviour or preventing some negative actions [23]. In our context, prevention campaigns typically focus their attention on negative consequences rather than on positive alternatives. This introduces our first class of protest-related actions: physical actions [263]. Next, awareness campaigns that promote positive behaviours try to actively connect with either informational or instructional resources [154, 260]. This motivates the definitions of two further types of actions: publications and calls for actions. Since most campaigns have some sort of supporters or base community, when running a campaign it is important to focus not only on the general public but also on specific stakeholders, e.g., to empower important communities, activate voluntary associations, or collaborate with governmental agencies. This often prompts the campaigns to organize official meetings, conferences, and debates [154, 260, 263]. Taking the above information into account, we consider five different classes of Tweets for our study.

- *Calls for action* correspond to tweets that contain a clear message calling for action, including actions to sign a petition, prevent events from happening, etc.

- *Publications* correspond to tweets that contain a reference to publication, news or some information related to the campaign, including videos, articles or background information on the campaign.
- *Official meetings* correspond to tweets that contain information about an official meeting, a conference, a convention or a debate related to the campaign.
- *Physical actions* correspond to tweets that contain information about past, current or upcoming actions organized by an individual, a group of people, or an organization that is related to the campaign. This includes proposals to participate to challenges, contests or to dedicate some time to a specific issue, e.g., cleaning streets or repairing homes.
- *Others*, finally, correspond to tweets that do not belong to the four categories above, such as content that is indirectly related to climate change or animal welfare domains, as well as personal opinions and experiences, or tweets in other languages.

| Type of action | Sample tweets |
|------------------|--|
| Official meeting | Monday Dec 1, U.N. COP climate talks begin Lima Peru @YebSano Just witnessed a sign of hope at the climate talks in #Cancun - ... #UNFCCC #tckctck |
| Physical action | #WorldEnvironmentDay #treeplanting is taking place around 09:00 at Tsarogaphoka in #Soshanguve We came. We swooped. We're camping!!! #climatecamp |
| Call for action | The #GreatBarrierReef is not a dump! Protect our World Heritage. #UNESCO #FightfortheReef Take Action: Stand with me and support clean #energy and a safer #climate future! #CleanAir4Kids |
| Publication | 660 million Indians could lose 2.1 billion years as a result of air pollution... #gofossilfree Water Fact: Fact: At 1 drip per second, a faucet can leak 3,000 gallons per year. #savewater |

Table 3.3 – Sample tweets for each type of action considered.

Tweet filtering and annotation

Next, we explain how we classified the tweets from our dataset based on the classes introduced above. Since manually annotating our whole dataset is unrealistic, given the high number of tweets involved, we introduce a two-step process, where we first use micro-task crowdsourcing to annotate parts of the dataset and then leverage the resulting annotations in order to build an effective classifier.

The aim of the first step, i.e., crowdsourcing, is to collect as many high-quality annotations as possible pertaining to the types of tweets while limiting the amount of the annotation from the crowd of the positive examples (tweets about each particular type of action). In order to do this, we first design a set of rules to preselect the tweets given our types. Those rules were created using simple regular expressions based on the analysis of a sample of the tweets, and are presented in Table 3.4.

In total, we created approximately 40 rules for each message type¹¹. These rules were geared towards high recall based on the message types, rather than high precision. Nonetheless, they allowed us to significantly narrow down the number of tweets that would be presented to the

¹¹<https://github.com/toluoll/CampaignsDataRelease>

| Type of action | Sample rules | N# of tweets |
|------------------|-------------------------------------|--------------|
| Official meeting | <i>speaking at (demo the)</i> | 113989 |
| Physical action | <i>action at (the)?park</i> | 154874 |
| Call for action | <i>tell (the)?to (keep protect)</i> | 328603 |
| Publication | <i>great news</i> | 2559063 |

Table 3.4 – Examples of rules and number of tweets for each type of action.

crowd by focusing on subcategories early in the process. The resulting counts of tweets obtained from this process are given in Table 3.4.

We then crowdsource the action type annotation using the CrowdFlower platform¹². The author of this thesis and two of her colleagues manually labeled 5% of the tweets beforehand to create a set of test questions for the crowd. Crowd workers could only work on our tasks if they correctly answered at least 7 out of 10 test questions. We additionally selected workers from English-speaking countries only and collected three independent judgments for every tweet. Agreement was obtained through the majority voting. We also made sure to identify and block malicious crowdworkers by leveraging a series of unambiguous test questions, following standard recommendations from CrowdFlower.

For each type of action, we considered a random sample of 2100 tweets. For more exploration, only half of these tweets is randomly selected from the collection complying with the regular expressions, while the other half is randomly selected from the rest of the campaign tweets. The results obtained through this process were consistent, with an agreement rate of 87.5%. In general, human annotators applied our definitions for the types of actions very strictly. However, this sometimes narrowed the results; For instance, human annotators did not always correctly annotate the tweets related to the attendance of a conference or a meeting when obvious keywords or the acronym of the event were missing, e.g., “conference”. As before, the annotated tweet collection is available online.

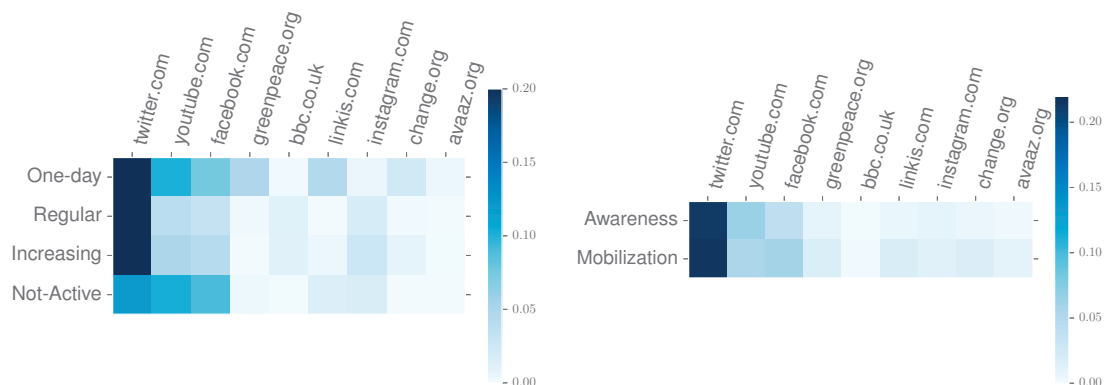


Figure 3.6 – Comparison of top domain name usage across campaigns

¹²<http://www.crowdfLOWER.com/>

Figure 3.6 shows the distributions of the urls by each particular campaign type.

Action classification

At this stage, we use the results of the crowdsourced annotation campaign as a training set to create an effective type classifier for the tweets. For this task, we consider the following features:

- *Semantic features.* Having a large textual corpus of 10Gb, we trained a Word2Vec model [201] using the implementation from the Gensim library¹³ with 200 word vector dimensions. To train the model, we preprocessed each tweet as follows: (a) deleted all punctuation excluding hashtag(#) and handler(@), (b) lowercased the tweets, (c) tokenized the tweets into words. Furthermore, we interpreted each tweet as a bag of word vectors and calculated an averaged vector for every tweet. The main motivation behind the choice of semantic features is their ability to capture the semantic similarity between words and phrases using contextual information [201].
- *Syntactic features.* In addition to the above features, we added manual rules based on the regular expressions from Section 3.5.1. This resulted in 46, 42, 38, 20 additional features for meetings, actions, calls for action and publications respectively.
- *Contextual features.* Finally, we added features based on the URLs inside tweets. We selected the most frequent domain names and used them as binary features for the classifier. The frequency threshold was chosen at one sigma.

| | Meetings | | | Actions | | | Call for actions | | | Publications | | |
|--------------|--------------|--------------|--------------|--------------|--------------|--------------|------------------|--------------|--------------|--------------|--------------|--------------|
| | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| All features | 0.896 | 0.616 | 0.730 | 0.771 | 0.577 | 0.660 | 0.902 | 0.664 | 0.765 | 0.897 | 0.528 | 0.665 |
| Sem | 0.723 | 0.605 | 0.659 | 0.707 | 0.510 | 0.592 | 0.751 | 0.441 | 0.556 | 0.857 | 0.461 | 0.599 |
| Sem + Cont | 0.788 | 0.531 | 0.635 | 0.703 | 0.518 | 0.597 | 0.792 | 0.439 | 0.565 | 0.865 | 0.487 | 0.623 |
| Sem + Synt | 0.912 | 0.590 | 0.717 | 0.754 | 0.480 | 0.587 | 0.920 | 0.563 | 0.699 | 0.862 | 0.464 | 0.603 |
| Synt | 0.895 | 0.375 | 0.529 | 0.816 | 0.276 | 0.412 | 0.920 | 0.472 | 0.624 | 0.890 | 0.134 | 0.232 |
| Synt + Cont | 0.901 | 0.384 | 0.538 | 0.812 | 0.300 | 0.438 | 0.921 | 0.643 | 0.758 | 0.911 | 0.325 | 0.479 |

Table 3.5 – Precision, Recall and F1-score values for classification of different types of actions with different sets of features.

In order to predict the type of a tweet, we trained an individual binary classifier for each of our action types. As a classification method, we used a state-of-the-art approach based on Decision Tree Ensembles¹⁴. Table 3.5 shows its precision and recall results for the four types of actions using 10-fold cross-validation. We observe that the physical action type has the lowest precision and recall among all types. We connect this result to the relative subjectivity in the definition of physical actions and to the high linguistic variety of the tweets of this type. The prominence of physical actions is hard to determine in general, since they can encompass anything from territory cleanups and protests to film-making competitions and tweet-a-thons.

¹³<https://github.com/piskvorky/gensim>

¹⁴<http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html>

Further, the introduction of semantic features extracted from the tweet word vectors leads to a loss in precision and to some improvement in recall. This is due to the semantic representation of the tweets, which allows to identify semantically related tweets and words. For example, in the vector space representation produced by the Word2Vec model, the word “debate” is most similar to the words “politics, issue, discuss, policy, conversation”. Overall, due to the very nature of the tweets (i.e, very limited length, use of slang, pictures, videos, or emoticons), recall is relatively low across all the categories.

As expected, we found that manually constructed syntactic rules result in better precision as compared to the Word2Vec features only. This is caused by the fact that the rules are highly representative of the classes they are built for. Additionally, we observed that domain names play a more important role for meetings, calls for actions and publications, which is explained by the presence of conference websites and specialized websites to gather petition signatures.

3.5.2 Data analysis

In order to detailed content of the campaigns, we ran the tweet type classifiers over all tweets from all campaigns. This analysis does not only consider the general differences between two major types of the campaigns, but also encounter the influence of the campaign agenda to the user involvement pattern. We relied on the classifiers that were trained on all features from the previous section as they achieved the best F1-scores for all message types.

We applied the models on each campaign to identify the amount of contribution of a particular action to the overall contents of the campaign.

A visual summary of the outcome for the two main classes of campaigns is shown on Figure 3.7. We observe major differences in terms of contents; in particular, we see that mobilization campaigns favor calls for actions that motivate the audience to react on the climate change issues, while, having relatively low physical actions. Interestingly, awareness campaigns encourage more physical actions and publication releases, while mobilization campaigns focus more on calls for actions and official meetings. Mobilization campaigns make a high use of official meetings, probably because they tend to raise more attention from the governments or particular stakeholders. To conclude, we see that mobilization and awareness campaigns get organized in very different ways, thus confirming the initial distinction we make between each other.

Following the analysis given in the Section 3.4, we performed a study on user engagement patterns. As shown on Figure 3.8, “one-day” campaigns¹⁵, focus on call for actions tweets which are mainly duplicated rather than retweeted. On the other hand, “regular” campaigns¹⁶ are mostly represented either by regular meetings or physical actions, e.g., annual conferences, campings, etc. Interestingly, “ever-growing” and “multi-burst” campaigns¹⁷, make larger use of publication

¹⁵#helpcovedolphins, #savebucky, #freethearctic30, etc.

¹⁶#climatecamp, #climateweek, #worldenvironmentday, etc.

¹⁷#talkpoverty, #saveanimals, #saveenergy, #actonclimate, #divestment, #fossilfree

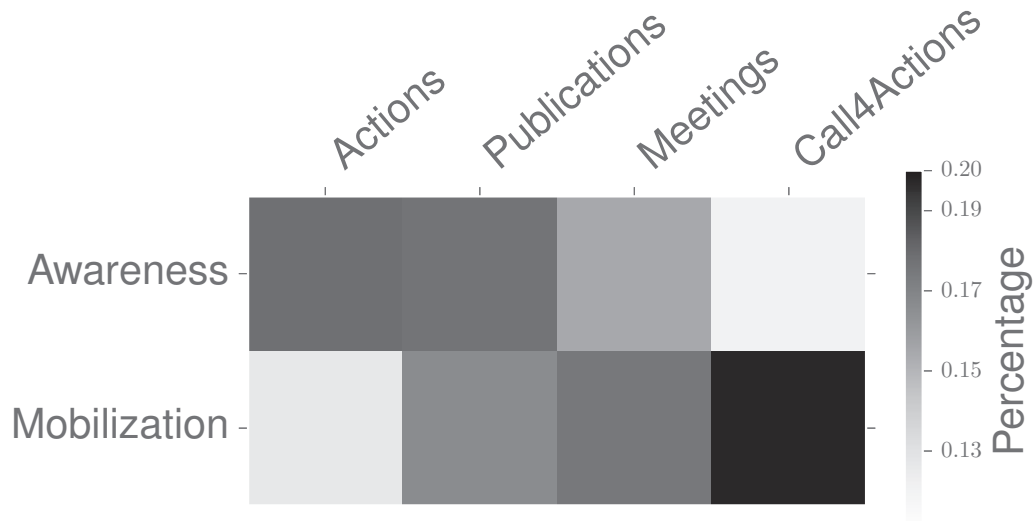


Figure 3.7 – Comparison of the distributions of actions for the types of climate change campaigns defined in Section 3.4.1. Dark indicate greater importance of a particular actions in the campaign type.

and call for actions types, which significantly differs from the awareness campaign strategies in general. This can be explained by the targeted audience and by the issues tackled by those campaigns, such as global poverty, international divestments, dependence on fossil fuels, etc. All of these campaigns share global values and target international audiences around the globe. On the other hand, both awareness and mobilization campaigns show the tendency to have an “ever-growing” pattern which is considered to be successful [281]. Overall, the analysis provides a reasonable background for a further investigation of the successfulness of the campaign in the environmental domain.

Duplicate tweets As described in Section 3.3.2, some tweets from our dataset shared the same contents but were not strictly speaking retweets. This is due to some users trying to promote a tweet into a trending topic on Twitter. We decided to compute the proportions of such duplicated messages to see how they are distributed across different campaign types. We identified the amount of such tweets, since this approach can be considered as a characteristic of a campaign. The main motivation behind segmentation of such content within a campaign, is recent splash of such activity for mobilizing people towards prevention killing animals etc.

To select the threshold at which a message should be treated as a duplicate, we considered the distribution of number of similar messages to the total amount of messages with these number as a half-normal distribution. In such way, the tweet was considered to be a duplicate if the number of such tweets exceeded three standard deviations.

Figure 3.9 illustrates the distribution of duplicate content for the different campaign types. As can be observed, duplicate content is especially significant for the mobilization campaigns, which

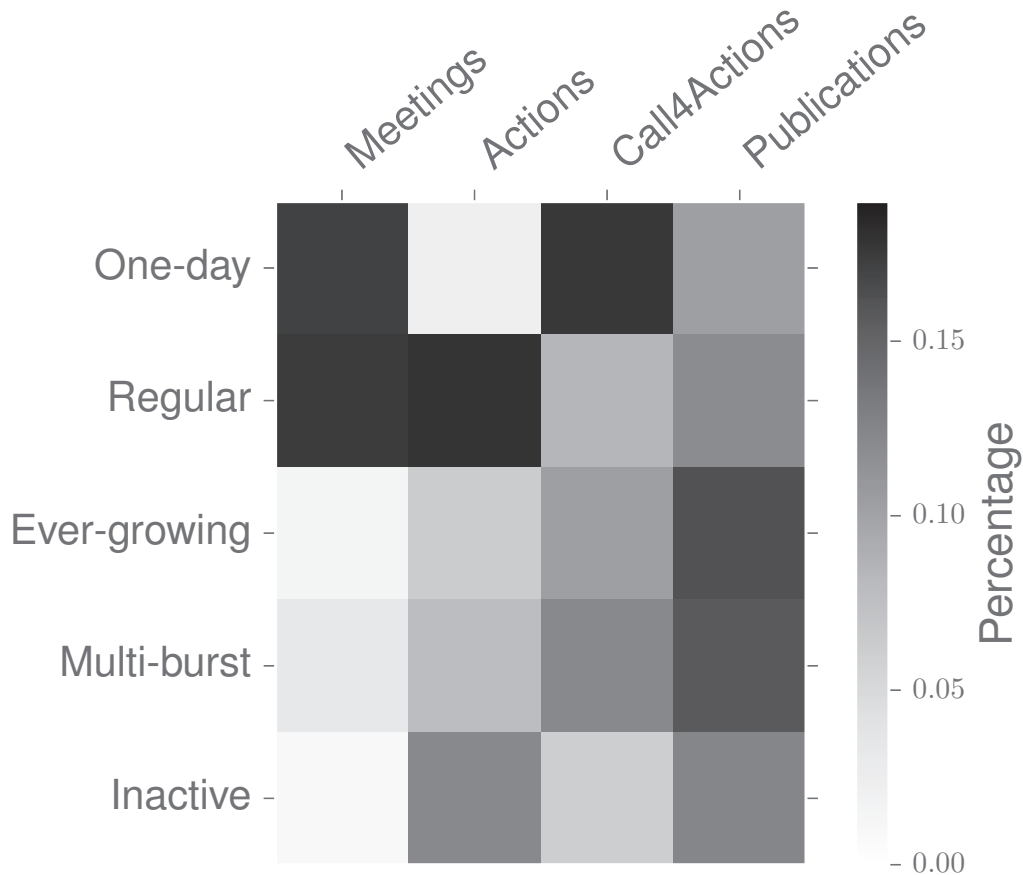


Figure 3.8 – Comparison of the distributions of actions for the two main categories of climate change campaigns: awareness and mobilization. Dark indicates a greater importance of a particular actions in the campaign type.

can be explained by their spontaneous nature and the need to mobilize people in shorter periods of time. Awareness campaign differ in the sense that they typically operate on longer-terms goals. From a user engagement perspective, both regular and inactive campaigns do not contain much duplicated content, which increasing, multi-burst, one-day campaigns make heavy use of it.

Domain usage distribution Users in the climate change community tend to make great use of links to images, facebook pages, youtube videos and petition sites. We explored the general distribution of the top domain names across the campaigns and found that all types of campaigns extensively use visual content (youtube, facebook, photos, etc.). Nevertheless, both ever-growing and regular types of campaigns use such content more parsimoniously comparing to one-day and inactive on average. A similar trend was discovered between awareness and mobilization campaigns respectively. Interestingly, the tendency to overuse visual resources clearly does not affect the popularity of the posted content [311]. Among major domain names whose tweets gained the most retweets, we primarily observe contents related to the campaigns, i.e., the site of

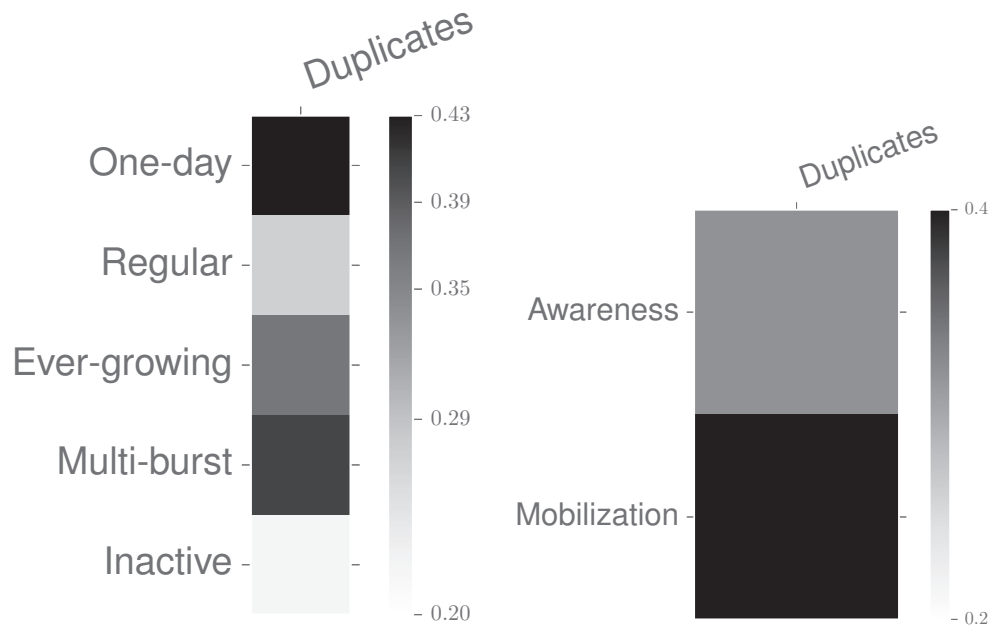


Figure 3.9 – Comparison of duplicate content between campaigns

the campaigns, news and information about related issues.

3.5.3 Retweets

Finally, we analyze the impact of various factors on the degree of retweeting in the campaigns.

First, we checked whether there is any correlation between the number of user followers and the average amount of retweets the user gets for a tweet. Surprisingly, even though having a large group of followers increases the overall probability to be retweeted (0.65 for users with more than 1'000 followers, and only 0.13 for users with less than 100 followers), it only marginally impacts the numbers of retweets for popular users. We observed an average amount of retweets equal to 3.3 for users with more than 1000 followers, and 7.8 for users with more than 10'000 followers. Needless to say, those results should only be considered in the context of climate and environmental campaigns.

Second, we made an analysis on the provenance of the retweets for a subset of 100 popular tweets. We found that on average close to 50% of the retweets are coming from direct followers (direct neighbors in the Twitter social graph). A tiny fraction (close to 5%) of the retweets are originating from second-order followers (i.e., followers of followers of the user). The remaining retweets are coming from users who are not in the user's social graph.

Next, we looked for key textual features that may increase chances of a tweet being retweeted. In that sense, we created a dataset of 100'000 pairs of tweets that were posted on the same day, by

the same user and about the same campaign. By collecting the dataset in this way, we eliminate the impact of the user's personal social graph, as well as time and topic differences. We selected the pairs of tweets such as one had zero retweets while the other had at least one retweet. The first one was considered as a negative example in our analysis, while the second was considered as positive. We then trained a Random Forest classifier on this dataset considering unigrams and bigrams as features. The resulting precision and recall of the classifier reached 0.5 and 0.6 respectively, which is only marginally better than a random guess. Among the features selected by the classifier as important, we could not identify any particular n-grams that predominantly appear in the retweeted tweets.

Finally, since every campaign we analyzed exhibits at least a small degree of duplicate content, we also estimated the likelihood of being retweeted for unique and for duplicated tweets. As can be expected, duplicated tweets get retweeted less often. In fact, their likelihood of being retweeted is reduced by almost a half as compared to unique tweets (0.19 vs 0.35).

3.6 Discussion

In the following, we take a step back, discuss the results we obtained and also make a series of recommendations in the context of public campaigns on social media.

First, we proposed a framework for collecting campaigns and identifying their types. As explained in Section 3.3, we collected over 100 campaigns that were annotated with types, i.e., awareness and mobilization, as well as clustered by their user engagement patterns (Section 3.4). This resulted in a large collection of tweets that were partially annotated with action types using crowdsourcing and further generalized based on an annotated corpus, using a machine learning classifier. Overall, our tweet action type detection technique showed high precision (~90%) and reasonable recall (~60%). This allowed us to automate action identification in tweets and to understand the overall content of specific campaigns.

Subsequently, we focused on the analysis of campaigns classified by their initial goal and their user engagement pattern. The goals of awareness and mobilization campaigns differ significantly, and so do their contents. While awareness campaigns often involve physical actions and promote scientific publications, mobilization campaigns make great use of official meetings and calls for actions; For the mobilization campaigns, the more official meetings organized the more leverage can be obtained from governmental organizations. The analysis of user involvement patterns also showed noticeable differences between campaign types and their agendas. "One-day" campaigns were dominated by calls for actions, while "regular" and "ever-growing" ones contained more physical meetings and publications on climate. This insight represents an important foundation on which specific campaigns studies and their contents can be build.

With the various techniques we leveraged for campaign analysis, we noted major differences in the way users duplicate messages. "One-day" and "ever-growing" campaigns in general

contain 20% more duplicated content as compared to the “inactive” campaigns. In the “one-day” campaigns, this phenomenon can be explained by the spontaneous nature of particular tweets and the need to mobilize people in a short period of time. On the other hand, “awareness” campaigns typically operate on a longer basis, therefore their communities do not actively use duplicated tweets, i.e., on average 15% less duplicates than mobilization. This can be explained by the actions required during shorter periods of time of the mobilization campaigns. In a context of user involvement patterns, both “regular” and “inactive” campaigns do not contain as much duplicated content, while “ever-growing” and “one-day” campaigns make a heavy use of them.

Regarding the effects that drive user engagement, we observe that first-degree neighbors are essential for getting higher numbers of retweets (about half of the retweets from popular tweets originate from direct neighbors), while duplicated content attracts less retweets in general. Finally and most interestingly, the less diverse the main contributors of the campaign, the less likely it is to gain bigger audiences (as shown in Section 3.4.2).

Overall, the work has a potential to empower the governmental and non-profit organizations by facilitating campaign analysis. The analysis of the collected campaigns combined with the analysis of individual tweets provides the foundation for many applications, e.g., detecting public campaigns, identifying means to boost user engagement.

Even though we expanded the campaign coverage by performing several iterations of the data collection, our methodology is focused on English-speaking tweets. The @AlGore and @GreenPeace accounts we used are biased towards the US, and so are the English terms and hashtags that were used for the climate change topic. Therefore, it would be beneficial to further expand the data collection by reiterating over the steps from Section 3.3.1 to sample more campaign hashtags over other languages and countries. At the same time, the same campaign may involve the usage of multiple hashtags and thus affect the results of the analysis.

Finally, we discuss how our framework can be applied to analyse data in other domains. First, our approach to identifying user involvement patterns is not bound to any particular domain and only depends on the actual user involvement during the campaign. Indeed, different domains might yield slightly different classes of user involvement. Nevertheless, each resulting cluster should be examined by experts and labeled manually as we suggested. Second, our type list could be further extended with more categories or transformed to some hierarchical structure to provide finer-grained categorization. New types of actions, however, require the creation of *new regular expression* rules and of *new annotations* for the crowdsourcing tasks. In other words, crowdsourcing can be used to identify less generic types of actions.

3.7 Conclusions

In this work, we analyzed large-scale social media campaigns related to climate change and animal welfare from various perspectives, including analyses on their primary goals, the types

of messages they relay, as well as their user involvement patterns. In the context of climate change and animal welfare, we showed that public campaigns are represented by two main narratives: awareness and mobilization. Our subsequent analysis of user participation revealed that campaigns significantly differ in terms of their user involvement patterns. Finally, we presented a study on the best ways towards increasing user involvement for public campaigns by combining core users, followers, and actions. The high-level patterns that were found in our study lay a solid foundation for future work on specific campaigns and their fine-grained segmentation. As a possible extension, a fine-grained classification of campaign actions could reveal more sophisticated patterns and correlations that appear during the campaign life-span, e.g., in case political or non-profit events exhibit different user involvement patterns.

4 Online Environmental Petitions in Public Campaigns

Social media have become one of the key platforms to support the debate on climate change. In particular, Twitter allows easy information dissemination when running environmental campaigns. Yet, the dynamics of these campaigns on social platforms still remain largely unexplored. In this chapter, we study the success factors enabling online petitions to attain their required number of signatures. We present an analysis of e-petitions and identify how their number of users, tweets and retweets correlate with their success. In addition, we show that environmental petitions are actively promoted by popular public campaigns on Twitter. Finally, we present an annotated corpus of petitions posted by environmental campaigns together with their corresponding tweets to enable further exploration.

4.1 Introduction

The discourse on climate change is often focused on the impact it has on the environment and on wildlife [272]. To bring those issues in the public spotlight, social media campaigns have proved to be an effective instrument to raise awareness and mobilize masses [225]. To further push for concrete actions from governments or public entities, many campaigns resort to e-petitioning [206], whose success is also much easier to assess: reaching or not a required number of signatures. Information about the number of signatures obtained for a given e-petition is often publicly available via e-petitions aggregators websites such as thepetitionsite.com, avaaz.org, change.org etc., and can be used as a proxy for the performance of the public campaigns and petitions themselves.

In this work, we tackle two main research questions.

RQ1: *Which types of the public campaigns use petitions in their agenda?* To answer this question, we study several environmental campaigns that were run in the beginning of 2015 as described in Chapter 3, measuring the incidence of e-petitioning as an instrument for promoting different types of campaigns (awareness, mobilization). We find that petitioning is particularly important

for mobilization campaigns.¹

RQ2: *What makes a petition promoted by a public campaign successful?* We answer this question by making a feature analysis and comparing tweets that belong to public campaigns to individual tweets. We propose a set of social and contextual features and show how the required number of signatures for an environmental petition is correlated to its outcome. Additionally, we release an annotated corpus with the petitions, their corresponding tweets and outcomes². For this study we focus on Twitter, which remains one of the main channels for social media campaigns, also providing relatively easy access to campaign data.

Climate Change Discourse on Social Media. Climate change is a highly discussed topic. Kirilenko *et al.*[154] overview the climate change domain, its polarization, discussion over time etc. Olteanu *et al.*[215] study how various climate-related events are highlighted by various media sources. A variety of public campaigns use social platforms to increase awareness or mobilize people [190]. Tufekci[288] describes how online attention can be driven towards particular politicized persona, while [103] analyzes information transmission during protests. Hestres[123] studies public mobilization and online-to-offline social movement strategies for two major environmental movements. Unlike this prior work, we analyze over 100 environmental campaigns as well as their effects on the success of petitions.

Characterizing E-petitions. Various studies were conducted to analyze e-petitions on various petition aggregators. Hale *et al.*[117] describe a temporal analysis of 8K petitions and discuss early signs of success (e.g., large number of signatures during the first days). Hung *et al.*[132] analyze “power” users that produce petitions. The authors have shown that only 1% of general petitions on change.org reaches their goal. However, to the best of our knowledge, we are the first to analyze which factors predict the success of an environmental petition based on the internal and external attributes of the corresponding public campaign on Twitter. On the other hand, e-petitions can be compared to crowdfunding, as both efforts work towards obtaining a given level of support over a short period of time. Etter *et al.*[78] study various prediction techniques for Kickstarter campaigns. Later, [17] analyze investor activity on Kickstarter and make recommendations based on their activity on Twitter. Unlike those works, we focus on environmental campaigns and petitions on Twitter.

In this work, we found that 25% of the petitions posted with environmental campaigns hashtags on Twitter obtained their required number of signatures. Moreover, we identified a number of features that can act as indicators for the success of the petitions. This information might be of interest to environmental activists and campaign leaders as it can influence the success of the message they are conveying to the public. We also note that the techniques presented below are not restricted to the environmental domain and could be applied to any related setting.

¹Mobilization campaigns refer to the campaigns whose primary goal is to engage and motivate a wide range of partners, allies and individual at the national and local levels towards a particular problem or issue, while awareness campaigns refer to the campaigns whose primary goal is to raise people’s awareness regarding a particular subject, issue, or situation.

²<https://github.com/toluolll/PetitionsDataRelease>

4.2 Data Collection, Cleansing and Insights

Our study is based on the collection of roughly 7,500 tweets and retweets belonging to 240 petitions related and mentioned by campaigns³ on environmental causes, which were posted from Jan 2015 to Apr 2015. Specifically, we consider a tweet to be related to a given petition if it contains the word “petition”. This filter is generic enough to capture mentions from the tweet text and from the URLs while being rather unambiguous.

Campaigns dataset and petition tweets: In order to answer **RQ1**, we created an annotated corpus of environmental campaigns for a given period of time on Twitter⁴. Our campaign corpus consists of 101 public environmental campaigns with over 850K unique tweets. We assume that each campaign has a uniquely identified hashtag, e.g., #saveafricananimals, #tweet4dolphins etc. Moreover, all the campaign hashtags are labeled by (a) their high-level goal, e.g., awareness or mobilization type, and (b) their user engagement pattern over time, e.g., one-day campaigns, ever-growing, annual, inactive⁵. These are the main categories that will be used in our analysis. Among those, “ever-growing” campaigns are the most interesting ones since they are characterized by a constantly growing number of involved people on Twitter.

We extracted all “petition” tweets from the annotated collection of environmental public campaigns tweets. Here we present an example of a tweet with a petition URL: “.@thetimes *Petition: Call for Safer Storage of Nuclear Waste in over 80 USA cities.* <http://tiny.cc/okzicx> #Save-FukuChildren”. Such tweets were identified in 39 (out of 101) campaigns. 15K tweets belonged to unique unresolved links (excluding tweets with broken links). In addition, we resolved, stored and annotated all petition URLs. As a result, we found 294 unique petition links and 158 broken or outdated links. For valid petition links, we stored their resolved URL. We further used this information to eliminate URLs that point to the same petition. This process has resulted in 240 unique petitions.

Tweets with petitions: Regarding **RQ2**, it should be noted that the campaign tweets collection does not account for the overall distribution of the petition tweets across the whole Twitter. Therefore, we collected additional data as we describe below. To minimize the bias in our collection, we further collected tweets that contain one of the 240 petition via backtweets.com. For this task, we used the collection of the extracted URLs with their resolved links (if applicable) and requested backtweets.com to return all historical tweets that mention the given URL. Clearly, this still results in only a subset of the petition tweets since it does not account for the URL redirects and shortening. However, we aim for a best-effort collection, which gives us a clearer picture on the distribution of the petitions tweets. As a result, we enriched the tweet collection with over 1,700 new tweets without campaign hashtag.

³List of campaigns is obtained as it is described in Chapter 3

⁴<https://github.com/toluoll/CampaignsDataRelease>

⁵ Ever-growing campaigns have constantly growing number of users posting with the hashtag. One-day campaigns have most of their user activity happening primarily on the first mention of the hashtag. Annual campaigns are mentioned annually. Inactive campaigns have very low user engagement overall.

Thepetitionsite.com. To compare campaign petitions with other environmental petitions, we additionally collected all the environmental and animal welfare petitions from the major petition aggregator⁶ *thepetitionsite.com* as well as the corresponding tweets from *backtweets.com*. This resulted in over 2,800 petitions with the following properties: (a) 35% of them are successful; (b) 79 of them are in the campaign dataset, (c) 186 of them are mentioned on Twitter with their direct URLs.

Dataset preprocessing To be able to compare petitions with each other, we use both campaign and non-campaign tweets. A petition p is characterized by its signature goal $S(p)$, collected signatures $C(p)$, $SignatureRate = \frac{C(p)}{S(p)}$ and the following set of Twitter related features $T_i(p)$: (1) Number of unique users posted the petition url; (2) Number of tweets with url; (3) Number of followers of the users posting petition tweets with/without a campaign hashtag; (4) Number of tweets with campaign hashtags vs without.

4.3 Petition Analysis

Given the list of petitions corresponding to campaigns on environmental issues on Twitter (described above), we first present an analysis on the petitions usage within different types of public campaigns and then analyze petition success by its visibility on Twitter.

4.3.1 Petitions and tweets stats

Table 4.1 includes the basic figures extracted from our list of petitions⁷. Surprisingly, we notice that failed petitions aimed to gather only about half as much signatures as successful campaigns. Furthermore, in our data, about a quarter of the petitions were successful, as opposed to only 1% as found by [132] across a broader range of petitions. Overall, the tweets corresponding to the successful petitions are more likely to be passed on, i.e., they are retweeted about 4 times more frequently.

After a deeper inspection of the petition collection, we identified that over 6% of the petitions in our dataset have a low signature goal $S(p)$, i.e., under 1,000 required signatures, out of which 13% are identified as successful (as they reach their goal). On the other hand, around 50% of the petitions have a high initial goal (over 30,000) among which 35% are successful. Additionally, we observed that 39 petitions reached over 100K signatures while 130 petitions collected over 10K signatures. The distribution of collected signatures is shown in Figure 4.1; it follows a Zipf distribution.

⁶Accessed on the 16th Feb 2016

⁷Latest petition signatures reassessment was on 28 Jan 2016.

| | <i>Successful</i> | <i>Failed</i> |
|--|-------------------|---------------|
| Petitions | 61 | 179 |
| Original tweets | 601 | 716 |
| Original tweets users | 245 | 313 |
| Retweets | 4828 | 1451 |
| Retweets users | 3965 | 1207 |
| Median $S(p)$ | 50000 | 15000 |
| Median $C(p)$ | 62997 | 6226 |
| <i>Petition tweets without campaign hashtags</i> | | |
| Tweets | 1054 | 707 |
| Users | 626 | 472 |

Table 4.1 – Global statistics of the petition dataset of environmental campaigns. We show the data for the successful and failed petitions, as well as total numbers. Users are unique individuals who tweeted the petition URLs at least once. $S(p)$ and $C(p)$ for successful and failed petitions are highlighted in the table. Additionally, we show statistics of the petition tweets that do not have a campaign hashtag.

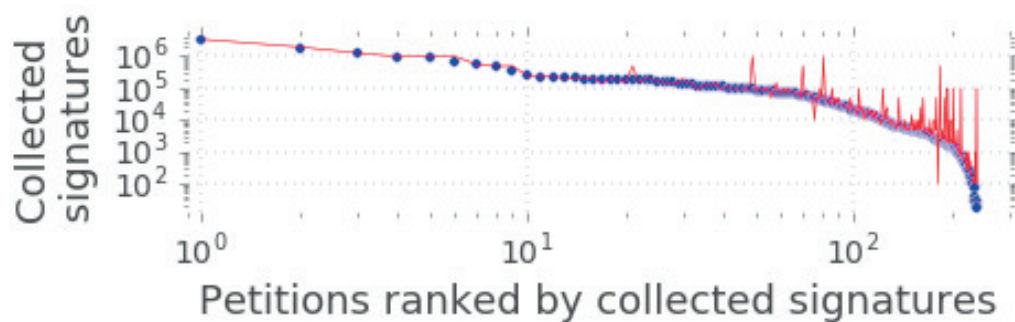


Figure 4.1 – The final number of signatures received by each petition. The red line indicates the required number of signatures. A change in the slope of the zipf distribution occurs at 1K signatures, which represents a threshold for a petition to make a potential impact.

4.3.2 Petitions in public campaigns on Twitter

The following subsection provides answers for **RQ1** based on our analysis. With only two exceptions, all the petitions were promoted through mobilization campaigns. The two exceptions are “#talkfracking” and “#worldlovefordolphins”, which are both awareness campaigns. Interestingly, these petitions for aforementioned two public campaigns hashtags were directed towards long-term plans, e.g., preventing “covering up” hydraulic fracturing by some organizations, or legalizing hemp farming.

As described in Section 4.2, the campaign corpus is also annotated according to user engagement patterns for each campaign, and consists of four main types: one-day, ever-growing, annual, inactive. We found that “ever-growing” campaigns (“#saveafricananimals”, “#tweet4dolphins” etc.) are the most active at tweeting about the petitions. The rest ~15% of the campaigns are mainly “inactive” (“#savethereef”, “#votegreen2015”). Not surprisingly, “one-day” campaigns do not use petitions as their instruments given the very short timespans of such campaigns. Among

campaigns with petitions, we also identified one “annual” campaign (“#worldlovefordolphins-day”) that is advertising multiple “Protect Dolphins” petitions that tend to have a high failure rate. Overall, there is no clear distinction between campaigns in terms of successful petitions. However, mobilization and “ever-growing” campaigns were the most active with petitions on Twitter.

4.3.3 Campaigns’ petitions on Twitter

After data collection, cleaning and preprocessing, we extracted a number of features from the tweets containing a petition URL. This process is explained in Section 4.2 in detail. To answer **RQ2**, we built a binary decision tree classifier⁸ over our petition tweets collection using our set of features.

On average, the resulting tree has a relatively high branching factor, however a few paths are better at predicting the petition success. We observe that the higher the signature goal, $S(p)$, of a particular petition, the more likely it is to succeed. In particular, for the signature goal between 100K and 300K 88% of the petitions were successful. However, setting a high petition goal may not guarantee its success. Success might also be correlated with various external factors, i.e., problem that a petition tries to address, external promotion (Facebook etc.), location of the petition owner etc. Hence, the success factors for those campaigns are very different from the success factors of Kickstarter campaigns, for which failed campaigns have goals (amount of money) about three times higher than successful campaigns [78].

In our case, over 92% of the petitions with $S(p)$ higher than 100K obtained their required number of signatures. Regarding $T_3(p)$, the lower the average number of followers a campaign activist has, the less likely the petition is to attain the required number of signatures. Similarly, the higher the average number of followers a user posting the petition URL without campaign hashtags has, the more likely the petition is to attain the required number of signatures. We observe that the average number of followers is 10x higher for users outside of the campaign compared to campaign activists.

Further Insights Towards RQ2 Since it is not trivial to provide step-by-step instructions on how to drive your petition towards success in general, we would like to highlight some additional key points from our analysis.

Does petition success correlate with the number of tweets? - Yes. We observed uniform distribution for the petitions with 0 tweets found on *backtweets.com* in terms of *SignatureRate*. On the contrary, for the petitions with several tweets carrying its direct URL, $T_2(p)$, we observed a very high fraction of successful petitions (88%). Pearson correlation for petitions with multiple tweets is 0.64 with $p < 0.05$. This effect is particularly strong when we consider only retweets or

⁸ <http://scikit-learn.org>

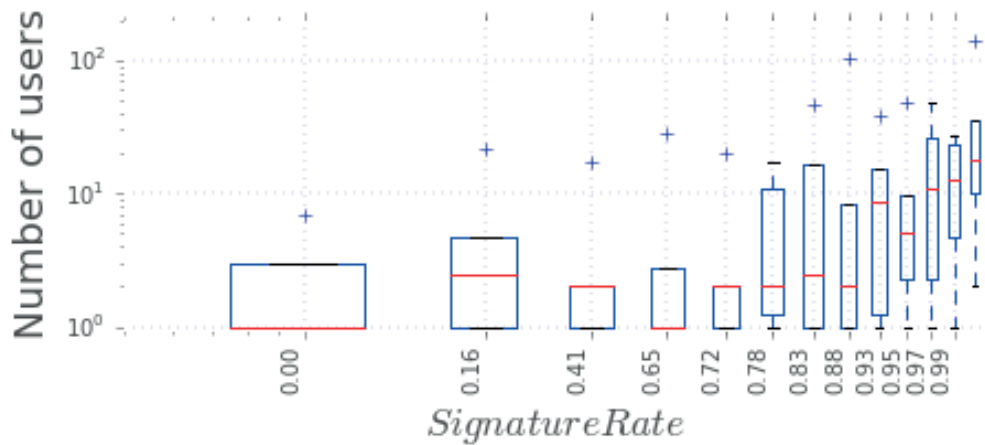


Figure 4.2 – *SignatureRate* against number of unique users posting about a petition on Twitter.

tweets without campaign hashtags, $T_4(p)$. We observed similar behaviour for *thepetitionsite.com*.

Does the number of users posting about the petition affect its success? - Yes. We binned the petitions from the campaign corpus based on the *SignatureRate*, and extracted the average number of unique users posting about the petition in each bin. Figure 4.2 shows a boxer plot with the 25th, 50th and 75th percentiles for each bin. As a result, Pearson correlation is over 0.7 with $p < 0.003$.

Is it common to post (a) identical tweets without acknowledging original tweets or (b) retweet? - Retweet. In our petition dataset we did not identify any duplicated tweets, i.e., tweets that are identical. Moreover, as shown in Table 4.1, the number of retweets for the successful petitions is several times higher than the corresponding number for the unsuccessful ones.

Which word features are more representative for tweets with successful petitions? - Uppercased. We discovered that tweets with successful petitions have more words and uppercase words on average, by 9% and 12% respectively. We compared the distribution of the uppercase words between the collections of successful and failed petitions by computing the relative change for each word. We define it as follows: $RelativeChange = \frac{W_{succ} - W_{fail}}{W_{fail}}$, where W_{succ} and W_{fail} are the term frequencies of uppercase word W for tweets with successful and failed petition. The top words from the successful collection are: “ACTION”, “URGENT”, “WAZA”, “PETITION”, “SIGN”, while the unsuccessful petitions did not uppercase those words at all.

4.4 Conclusions

In this chapter, we introduced a dataset of environmental petitions that were promoted by major environmental campaigns on Twitter. We studied the use of petitions as one of the instruments of

a public campaign. We proposed a model to identify successful petitions and highlighted key aspects to obtain the required number of signatures. Although our dataset is limited in size, we could observe the petitions spread within the environmental campaigns and identify the major factors that lead to the success of the petitions. Our findings provide helpful directions for all public campaigns, its participants, petition initiators, and signers.

In this piece of work, we quantified the positive effects of the intense petition promotion on Twitter, e.g., the number of retweets, unique users, user followers and attention uppercased words correlating to successful petitions. In the next Chapter, we are enhancing current petitions' dataset by collecting the hourly information on petitions' engagement as well as explore the time series of the signatures and which quantify factors affect the engagement. Moreover, we expand on this work with respect to the analysis of the petition signers and users who promote petitions on Twitter.

Modelling Part

5 Predicting the Success of Online Petitions

Applying classical time-series analysis techniques to online content is challenging, as web data tends to have data quality issues and is often incomplete, noisy, or poorly aligned. In this chapter, we tackle the problem of predicting the evolution of a time-series of user activity on the web in a manner that is both accurate and interpretable, using related time series to produce a more accurate prediction. We test our methods in the context of predicting signatures for online petitions using data from thousands of petitions posted on *The Petition Site*—one of the largest platforms of its kind. We observe that the success of these petitions is driven by a number of factors, including promotion through social media channels and on the front page of the petitions platform. The interplay between these elements remains largely unexplored. The model we propose incorporates seasonality, aging effects, self-excitation, external shocks, and continuous effects. We are also careful to ensure that all model parameters have simple interpretations. We show through an extensive empirical evaluation that our model is significantly better at predicting the outcome of a petition than state-of-the-art techniques.

5.1 Introduction

The ability to predict user activity or engagement on the web has many applications in a wide range of domains. This includes, for instance, predicting the number of people who will install an application in an app marketplace, buy a product from an online retailer, or participate in an e-government action. Ideally, a forecast of user involvement should be generated as *early* as possible, in a manner that is both *accurate* and *interpretable*. The quest for interpretability is due to the importance of knowing what are the elements that are driving predictions up or down as a process unfolds, in order to take corrective actions whenever possible.

The problem of generating early, accurate, and interpretable predictions on the web probes our understanding of complex interactions over time, and is further complicated by data availability and data quality issues. The data that we obtain from the web is almost invariably noisy and incomplete, and often comes from several heterogeneous sources. Additionally, and despite recent

advances in empirical methods for predicting information dissemination [281, 324], we lack a general parametric modelling framework to predict user involvement in a *reinforced* process, i.e., one that is driven actively by the efforts of a person or organization through some sort of online campaign. In this Chapter we consider a particular instance of a campaign - *online petition*. For instance, the mobilization of people through a particular online campaign might involve several sources of reinforcement: various social media, traditional news media, and word-of-mouth or viral advertising.

In this chapter, we present new forecasting models for online content dissemination that are able to take into account several elements: self-excitation, seasonality, web platform artifacts, and the presence of external factors (e.g., in the form of social media postings). Our main contribution, beyond presenting a combined parametric model that has better predictive power than the state of the art, is being able to incorporate a time series of related observations to produce a more accurate and earlier prediction, and to further enhance the interpretability of the results.

We evaluate our models by using them to predict the number of signatures an online petition will gather over time. Online petitions are, in our opinion, representative of a broad class of online phenomena involving active public mobilization, and thus represent a relevant scenario for testing our methods. The setting we consider might generalize to the spread of online ideas or memes, in the sense that it exhibits *active* promotion, instead of simply *passive* diffusion. People promoting online petitions, and people who sign petitions tend to encourage others to sign, instead of expecting that people simply learn about these petitions through a contagion process, which also takes place but does not fully explain what we observe.

Our contributions. In this work, we present models for user behaviour with respect to online petitions. We make the following contributions:

- we analyze thousands of online petitions from one of the largest petitions sites on the web (Section 5.3);
- we present a model to predict user involvement in a reinforced manner combining self-excitation, seasonality, aging, and external evidence as a continuous signal; this model has easily interpretable parameters (Section 5.4);
- we show that our proposed model is more accurate in both short-term and long-term predictions of user involvement, when compared with state of the art methods (Section 5.5).

The rest of the chapter is organized as follows. We start with an overview of related work in Section 5.2. We describe our process for collecting petition data, as well as the insights we gained through that process in Section 5.3. We present our new predictive model and compare it to existing models in Section 5.4. We experimentally evaluate the models and discuss them in Section 5.5. Finally, we summarize our results and outline future work in Section 5.6.

5.2 Related Work

This section outlines previous work on popularity prediction on the web in general, and for online petitions. We also position this chapter with respect to these previous contributions.

5.2.1 Popularity prediction on the web

Predicting the popularity of user generated content on the web is a problem that has been studied extensively [283]. Many different settings have been considered; typical content types include online videos [178], online news [45], social bookmarking sites [171], social networking services [324], crowdfunding campaigns [78], among others. Most works on this topic tackle one of three main tasks: (i) *classify as successful/unsuccessful*, meaning trying to predict whether a particular piece of content will exceed a certain popularity threshold or not; (ii) *predict the overall popularity*, i.e., predict the final number of views or votes a piece of content will receive; and (iii) *time series forecasting*, i.e., modelling the popularity dynamics over time. Regardless of the specific task, two main types of approaches are observed: feature-based and model-based. *Feature-based* techniques rely on a set of (hand-)crafted features extracted from a single or multiple sources, for the purpose of classification or regression. *Model-based* techniques assume a specific parametric model for the process that defines the phenomenon; they are usually harder to formulate, but often produce better insight into the studied phenomenon. We summarize these approaches and include references for each one in Table 5.1.

This chapter goes beyond analyzing “meme”-like content that spreads virally, and study a phenomenon that involves active promotion; hence, we need to consider external signals. External information is used by previous work adopting feature-based approaches that extend Szabo and Huberman [281] (such as [45]), but not in model-based methods, as we do in this work. Our approach is based on modelling the conditional mean of a Hawkes process, as Kobayashi and Lambiotte [156] suggested. However, we extend their model with a more flexible aging, i.e., raise and decay, and include both internal dynamics (self excitation) and external factors (social network, front page effect). Moreover, each external factor is modeled as a continuous effect on the signature dynamics, rather than a series of single external shocks. This allows us to efficiently fit the model and easily interpret the results. To the best of our knowledge, we are the first to present a model that captures interaction between multiple platforms in a model-based framework and with easily interpretable parameters.

5.2.2 Analyzing the dynamics of online petitions

Signature acquisition in online petitions is a complex and multi-dimensional problem. From the perspective of online activism, it is not only important to predict whether a petition will gain the required number of signatures or not, and what the final number of signatures will be, but also to start from valid assumptions about how the number of signatures evolves over time, and how external factors shape this evolution. Understanding these factors can help the organizers of the

| Approach | Data source(s) | Examples |
|--|----------------------------|---|
| Classification | | |
| .. Feature-based | Twitter | Hong <i>et al.</i> [126], Ma <i>et al.</i> [187], Cui <i>et al.</i> [69], Jenders <i>et al.</i> [138], Cheng <i>et al.</i> [55] |
| .. Social transfer | Multiple sources | Roy <i>et al.</i> [251] |
| .. Model-based | YouTube | Crane and Sornette [67] |
| Popularity prediction | | |
| .. Feature-based | Digg, YouTube | Szabo and Huberman [281] |
| .. Feature-based | Online news | Castillo <i>et al.</i> [45] |
| .. Feature-based & logistic regression | Twitter | Kupavskii <i>et al.</i> [162], Bao <i>et al.</i> [26], Hong <i>et al.</i> [126], He <i>et al.</i> [120] |
| .. Model-based | Twitter | Zhao <i>et al.</i> [324] |
| .. Model-based | Earthquake, neurons, crime | Ogata <i>et al.</i> [213], Pillow <i>et al.</i> [229], Mohler <i>et al.</i> [205] |
| .. Model-based | Multiple sources | Choi <i>et al.</i> [56] |
| .. Social dynamics | Digg | Lerman <i>et al.</i> [171] |
| Series forecasting | | |
| .. Model-based | Twitter | Kobayashi and Lambiotte [156], Gao <i>et al.</i> [92], Shen <i>et al.</i> [266] |
| .. Model-based | Multiple sources | Linderman <i>et al.</i> [182], Xu <i>et al.</i> [310], Xu <i>et al.</i> [311] |
| .. Time series clustering | YouTube, Digg, Vimeo | Ahmed <i>et al.</i> [6] |

Table 5.1 – Selected works on popularity predictions in social media. Typical tasks in this context are to classify as successful/unsuccessful (top), to predict the overall popularity (middle), and to forecast the popularity time series (bottom).

petitions to further enhance the engagement of the public with their campaigns.

Hale *et al.* [117] describe a temporal analysis of 8,000 petitions and discuss early signs of success (e.g., a large number of signatures during the first days). However, it remains unclear why some petitions become popular and others do not, or what are the factors that can lead to an increase in popularity. Huang *et al.* [132] analyze “power” users on petitions platforms and how user involvement changes over time on a petitions platform. Proskurnia *et al.* [233] study the effect of petition success on user involvement. In contrast, we link social media and petitions together to gain insights on the underlying process, focusing on modelling its evolution considering multiple factors, including external influence.

Online petitions can be compared to crowdfunding campaigns, as both efforts work towards obtaining a given level of support over a bounded period of time. Etter *et al.* [78] study various prediction techniques for crowdfunding campaigns on Kickstarter. An *et al.* [17] analyze investor

activity on Kickstarter and make recommendations based on their activity on Twitter. Unlike these works, we focus on signature rate dynamics using co-evolving time series information, and we do not limit ourselves to signals from social media, but also utilize further available information, including the effect of being featured on the front page.

5.3 Data Collection and Insights

Our study is based on petitions obtained from *The Petition Site*¹, one of the top-3 online petitions site according to Alexa.² *The Petition Site* allows anyone to create an online petition and to gather signatures. There are 14 categories in which the petitions can be started, including Environment and Climate, Education, Health, and Human Rights. Petitions have a headline (e.g., “Help stop the Taiji dolphin slaughter”), the name of the person or entity to whom the petition is addressed (e.g., “International Marine Trainers Association”), the name of the person who creates the petition, dates of opening and closing of the signature gathering, and a description and/or letter describing the contents of the petition. Petitions also include a target number of signatures, decided by its author; we consider that petitions that reach this target are *successful*, otherwise they have *failed*.

5.3.1 Data collection

We collect two kinds of information on those petitions: list of signatures, and tweets pointing to the petitions. The entire data collection pipeline is illustrated in Figure 5.1. Each signer is represented with an object that contains her full name, country of location, time of the signature. Each signer is uniquely identified on the petitions’ platform, thus we assume that a user can sign a petition only once³. When signing a petition, a user must authenticate herself on the platform. The user has an option to sign petition *anonymously*. In this case private information (such as name and country) are not displayed on the platform, thus, we obtain only information about the signature time.

Petitions data. Petitions data were obtained using a custom-made web crawler and scraper to collect petitions created after August 1st, 2016 across all the topics. The resulting petitions garnered around 85 million signatures from about 5 million unique users. While there are old petitions in the data we collected—some dating back to 2003—we decided to focus solely on petitions that started after August 1st and were active for at least 10 days. These petitions comprise 85% of the total number of signatures in the entire collection. Given our focus on petitions reaching a target goal of signatures, we additionally removed five outlier petitions having unattainable goals (requiring more than 1 billion signatures).

Each petition has a web page including public information about the people who signed the

¹<http://thepetitionsite.com/>

²<http://www.alexa.com/topsites/category/Society/Activism/Petitions>

³Theoretically, some users might register several accounts, however, different email address should be used.

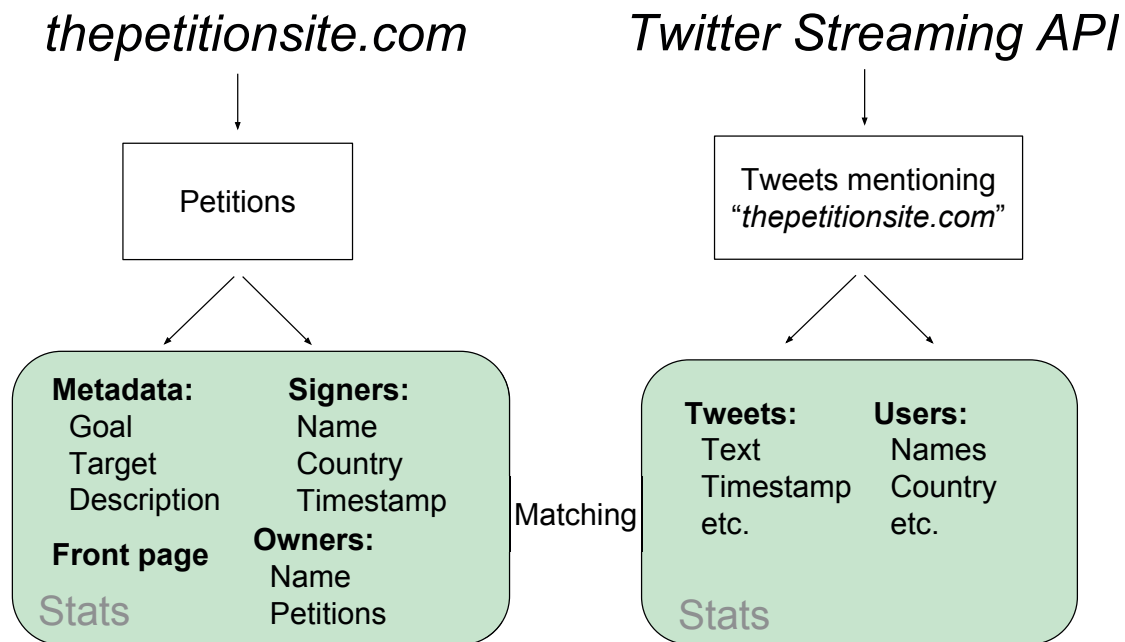


Figure 5.1 – Collection pipeline for the petitions dataset.

petition. Each signer is authenticated on the platform by providing an e-mail address, whose ownership must be verified before the signature is recorded. Once the e-mail address is verified, signers may chose to remain anonymous (listing only the signature timestamp on the website), or to disclose more information (such as their first name and country of residence). Additionally, we collected hourly data for top 10 petitions promoted on the front page starting August 1st.

Twitter data. In addition, we used Twitter’s streaming API to collect all tweets containing a link to any URL containing “thepetitionsite.com.” Tweet timestamps were processed and normalized to be in the same format (POSIX) as signature timestamps from *The Petition Site*. Tweet collection was conducted from August 1st, 2016 through October 1st, 2016, collecting over 250K tweets.

Time normalization. *The Petition Site* and Twitter has different time representation. The former stores all the information about the signatures relatively to the PST timezone, while Twitter stores tweet’s unix timestamps relatively to UTC with occasionally available user timezone⁴. We perform the following time transformation in order to reason about (1) relative time between signature and tweet, (2) local time of the day when tweet or signature were added. Former is obtained by converting both signature and tweet times to UTC Unix timestamps. Later is achieved by (1) applying twitter user utc offset to the tweet timestamp, (2) converting a signature time to the UTC time zone and further localizing the timestamp to the country specified by the signer.

Owner identification. Fine-grained information about the petition owners is not identifiable

⁴UTC is Coordinated Universal Time, PST is Pacific Standard Time and -8 hours behind the UTC

Petition target categories

| | President, Government | Parliament, Minister Congress, Senator | Governor, State | Mayor | Agency, Commission Institution, Department | Council | CEO, Director, Corporation | Single Person | Other |
|-----------|-----------------------|---|-----------------|-------|---|---------|-------------------------------|---------------|-------|
| N | 305 | 290 | 171 | 63 | 190 | 34 | 140 | 44 | 226 |
| $S(p), K$ | 43 | 33 | 10 | 12 | 9 | 21 | 12 | 15 | 20 |
| $C(p), K$ | 27 | 39 | 29 | 21 | 56 | 37 | 45 | 26 | 19 |

Table 5.2 – Number of petitions, main signature goal and mean number of collected signatures (in thousands) that are directed to a particular category of the petition target.

from the petition’s web page. The only information that is available about the owner is her name which is spelled and specified by the owner manually. Thus we designed a heuristic that identifies owners among the first signers of each petition. The heuristic matches lower cased combinations of the first and last names of the first 10 signers to the owner’s name. This way we have identified over 800 owners as well as other petitions that are signed or created by them.

Petition target. Fine-grained information about the petition target is not available similarly to the owner’s one, i.e., it is free formatted by the owner, e.g., “Dean on EPFL” etc. Moreover, 3% of the petitions did not have a target specified at all. Thus, we have categorized petition targets into the categories ranked by their importance for a particular country. Table 5.2 provides basic statistics for each target category for the 1463 petitions that we were annotated. We used annotations from the CrowdFlower⁵. Category “other” has targets that are either not specific or addressed to all people or communities, e.g., “contractors”, “all people”, “animal lovers” etc.

Collected petitions are labelled by the fine-grained subcategories which are assigned manually by the petitions owners from the offered taxonomy. Table 5.3 presents the taxonomy provided by the *thepetitionsite.com* platform, while Table 5.4 shows the fraction of the petitions in each particular category that reached the signature goal.

5.3.2 Data insights

The overall characteristics of the collection are shown in Table 5.5. As expected, the distributions of the number of signatures collected by successful and failed petitions are significantly different

⁵Crowdsourcing configuration: 3 answers per each petition; 10% of the targets are annotated by the author of the thesis as a golden standard; label accepted if all labels are the same; labels are verified manually if they are not equal (this case account for 23%). Annotator agreement is 87%. Among ambiguous categories are the following: institution vs authority, congress vs council, corporation vs single person.

| | |
|--------------------------|---|
| Animal Welfare | animal abuse; animal research; farm animals; pets; |
| Environment and Wildlife | arctic; endangered species; wildlife; environmental health; oceans; oil drilling; global warming and climate change; rainforest; national parks and forests; pollution; whales; |
| Human Rights | women rights; LGBTQA rights, death-penalty; refugees; etc; |
| Politics | conservative; international; progressive; |
| | etc. |

Table 5.3 – Petition categories labeled by *thepetitionsite.com*.

| Category | Success rate | Mean number of collected signatures (std. deviation) |
|---------------------------|--------------|--|
| Animal Welfare | 0.193 | 8,451 (31,963) |
| Environment and Wildlife | 0.213 | 12,091 (44,784) |
| Politics | 0.783 | 34,128 (25,161) |
| Corporate accountability | 0.870 | 49,196 (92,972) |
| Human rights | 0.462 | 22,545 (23,584) |
| Health | 0.750 | 39,165 (42,101) |
| Media, art, culture | 0.333 | 27,088 (28,022) |
| Spirituality and religion | 0 | 2,177 (178) |

Table 5.4 – Petition categories’ success rates.

($p \ll 0.001$). On the other hand, both successful and failed petitions have similar timespans, 50 and 42 days on average respectively.

Table 5.5 – Dataset characteristics. Each characteristic in this table shows a significant difference at $p \ll 0.001$.

| | Successful | Failed |
|-------------------------------|------------|--------|
| Petitions | 1,219 | 3,505 |
| Median signatures goal | 4,319 | 43,838 |
| Median signatures collected | 51,986 | 5,687 |
| Anonymous fraction | 0.023 | 0.044 |
| Fraction of signers’ comments | 0.031 | 0.045 |
| Petitions with tweets | 90% | 27% |
| Mean number of tweets | 83.3 | 37.1 |
| Mean number of retweets | 31.2 | 24.7 |
| Mean number of unique users | 62.7 | 26.8 |

We observe that successful petitions have more modest goals but also collect more signatures than failed ones and while the majority of people include their first name and country, signatories of failed petitions are almost twice as likely to remain anonymous—they might be less willing to be publicly associated to these petitions. We also observe that petitions that are successful have on average more activity on Twitter: they are three times more likely to have tweets, and have an

average number of tweets that is more than twice the number of tweets failed petitions receive. The cumulative distribution of signatures for over 4,000 petitions is shown in Figure 5.2 (left). From the figure, we observe that over 70% of the failed petitions did not reach 1,000 signatures, while all successful petitions obtain at least 1,000 signatures and over 20% of the successful petitions reached over 100,000 signatures.

As previous works [117, 269], we observe that the higher the number of signatures a petition receives early on, the more likely it is to gain the required number of signatures. Figure 5.2 (right) shows the distribution of the number of signatures for the first 3 hours of a petition. Almost all failed petitions acquire less than 10 signatures during the first 3 hours, but almost 60% of the successful ones also acquire less than 10 signatures. As a result, a significant part of the successful petitions are indistinguishable from failed petitions during the first hours and, thus, it is not trivial to make an accurate prediction on whether they will succeed or not using only this data. Observations done using the first 24 hours of each petition, omitted for brevity, show a similar lack of separation between successful and failed petitions.

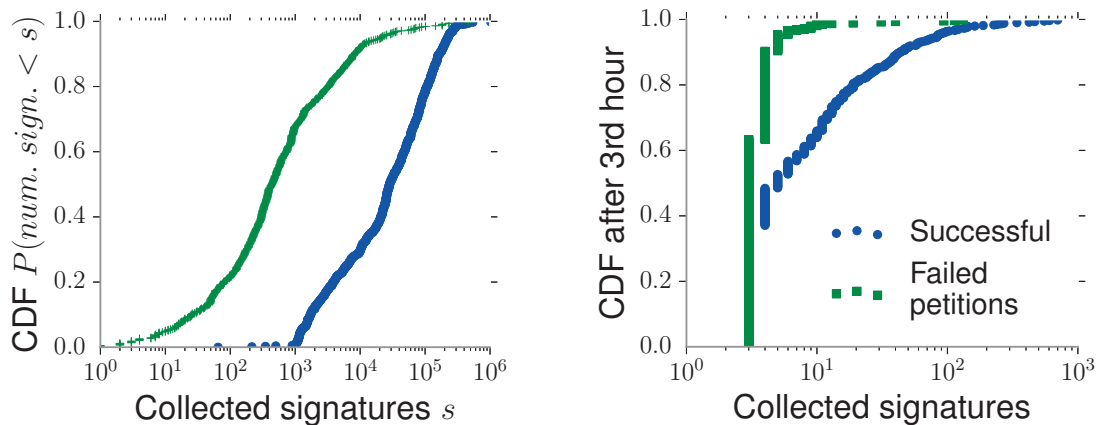


Figure 5.2 – Cumulative Distribution Function (CDF) of the signatures collected by successful and failed petitions during their entire history (left) and during their first three hours (right).

We clustered the petitions’ time series using Dynamic Time Warping [97]. We varied the number of clusters from 2 to 30 and found that cluster quality, in terms of inter-cluster distance, stabilizes at about 4 clusters. The corresponding centroids are shown in Figure 5.3. Each cumulative distribution function for the petition signatures has been rescaled to the unit interval and to have the same number of time bins. Again, we observe that successful petitions tend to gather a large share of their signatures early on.

Interestingly, this findings could be generalised to particular countries, e.g., the higher the participation of the USA signers the more likely the petitions to gain required number of signatures.

The following provides a set of key questions and findings that describe our dataset from the perspective of the user’s participation and their geographical activity.

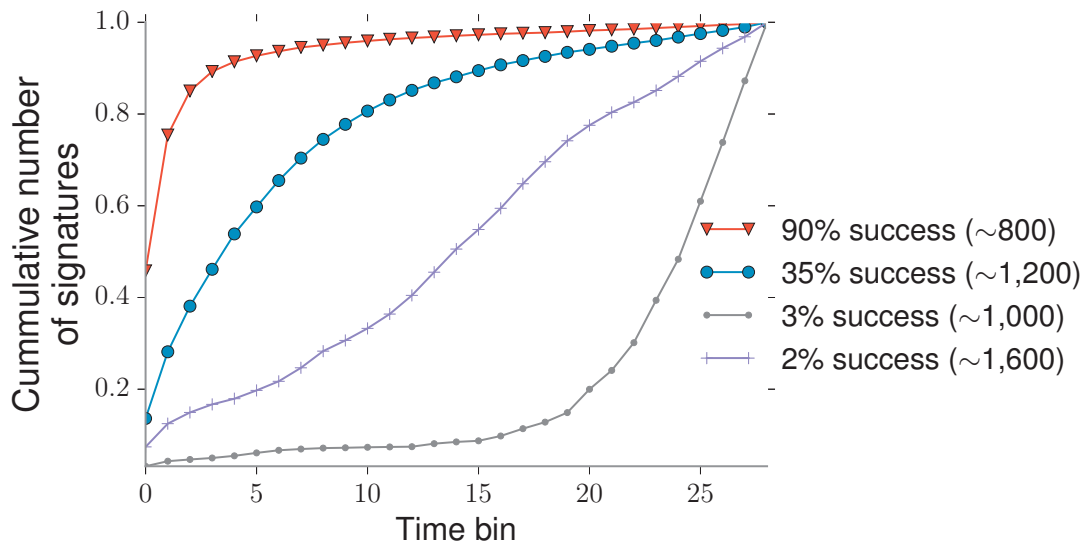


Figure 5.3 – Average of normalized Cumulative Distribution Function (CDF) in four clusters of petitions. Clustering was performed using Dynamic Time Warping (DTW). Numbers in parenthesis represent the size of each cluster.

How different countries participate in environmental petitions?

Figure 5.5 shows the distribution of the top countries⁶ that are the most active for environmental petitions. In particular, “us” signers account for almost 41% of the contributors to the petitions on average, while the following “gb” signers has only 12% on average.⁷

Do users sign multiple petitions (category) and how often?

To identify whether signers of the particular petition are active for the other ones in the domain. Figure 5.6 shows, that among 5M total users over a half signs only a single petition (right most point). On the other hand, ~50 users contribute to ~1000 petitions.

We have considered categories provided by the petition platform that are shown on Table 5.3. As a result, we found that over 130 users signed petitions from all the categories while ~ 3.5M users has signed contributed to only one subcategory (out of while ~ 2.8M users signed only one petition overall).

Do owners sign their petitions or petitions (category) of others?

We have examined the owners of the 2,120 petitions. We have identified that among those only 890 petitions were signed by the owners publicly with the average number of created petition 1.15⁸. This does not imply that other petitions are not signed by the owners but rather that

⁶Country codes can be found in <https://countrycode.org/>

⁷However, it should be noted that the origin of the petition owner, thus petition, is known for only 890 owners. As a result the data presented in the chart is not normalized by the petition origin and might be biased towards petitions of US origin.

⁸Only 54 owners had 2-3 created petitions

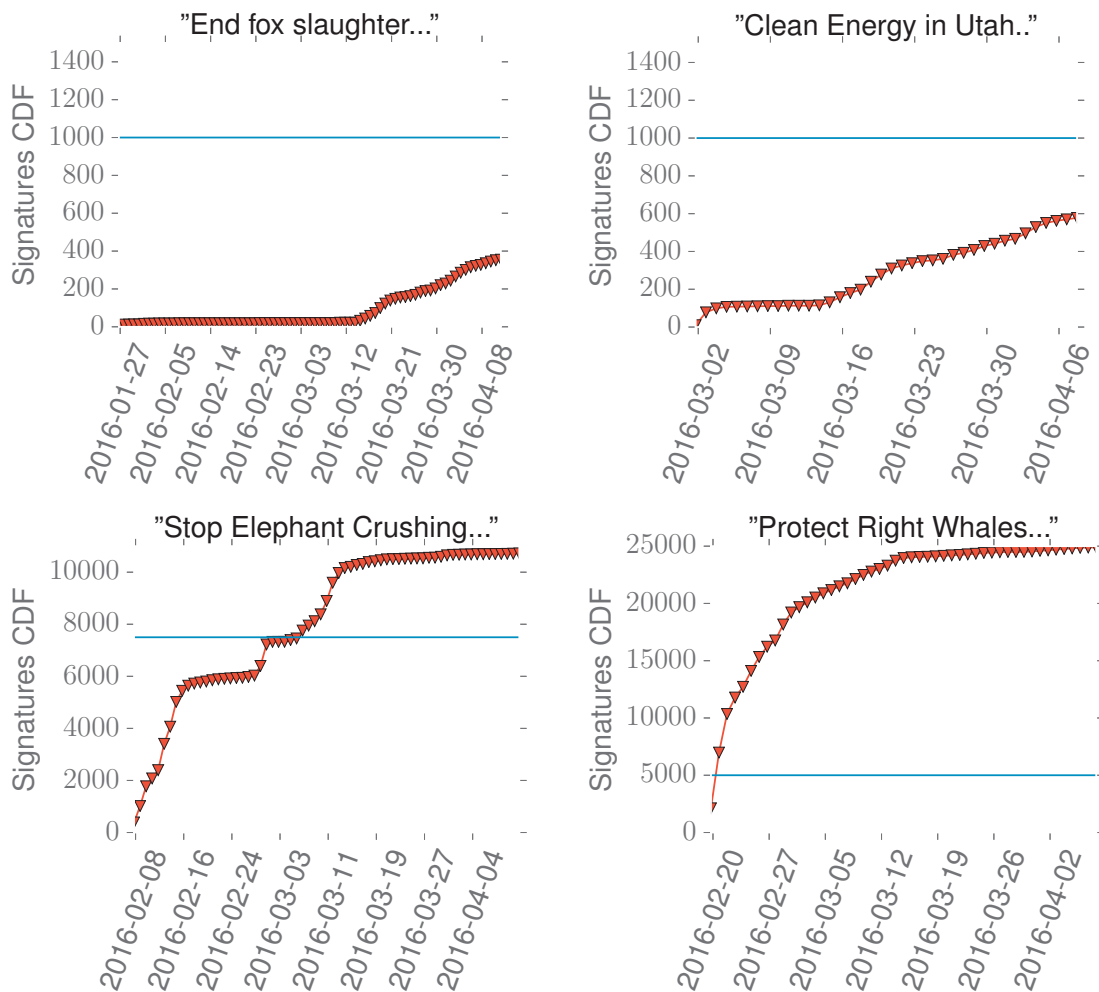


Figure 5.4 – Petition signature daily cumulative distribution function. Straight line corresponds to the signature goal set by the owner. First two distributions belong to the failed petitions while the last two belong to the successful ones.

their owners sign them anonymously⁹. Not surprisingly, we have found that owners sign their petitions as one of the first ones (being ranked on 1.41 position on average). For the owners we have identified that on average they participate in 388 other petitions covering 13 out of 15 subcategories. On the contrast, the average among all signers is 9 and 2 for number of petitions and number petition categories respectively. This observation proves that the petition owners are more likely to be active in other related petitions apart from their own.

Does owners location affect who signs their petitions?

Since we could extract more detailed information about the owners of the petitions, such as their location, we have identified whether owners origin attracts signers from the same country. As can be seen on Figure 5.7 owner origin does not imply that the petition will be signed by the users

⁹“Anonymously” means that the user is authenticated in a system but her name is not displayed



Figure 5.5 – Average number of signatures per petition that are contributed by the top active countries.

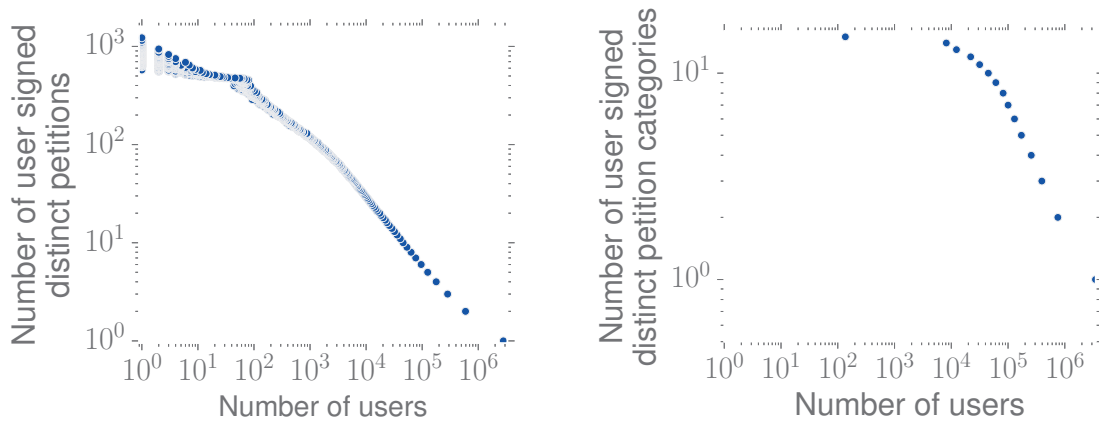


Figure 5.6 – The number of the distinct petitions (left) and distinct petition subcategories (right) signed by the users, e.g., (left) 101 users (x axes) signed 324 petitions (y axes); (right) 136 users signed petitions from all environmental categories. A change in the slope of the distribution occurs for the users with less than 300 petitions signed which correspond to the average number of petitions signed by the owners.

from the owner’s country. This is an important observation since response from the governmental authorities usually require signers from the specific country, e.g., if the president of Ukraine is the target then only Ukrainian signers count.

Interestingly, petitions created by the following countries are never dominated by the signers from the same country: Sweden, Switzerland, Greece, Ireland, Poland, Portugal etc. This can be explained by an extreme activity of the users originated from USA, Great Britain and Canada.

Does petition description correlate the petition performance?



Figure 5.7 – Top petition creators are USA, Great Britain and Canada. On the x axes there are country codes of the petitions owners. Left y axes (bars) shows the average rank of the owners country among top 10 countries that contribute to the petitions. Right y axes (scatter) shows absolute number of petitions owners from a given country (country code).

Surprisingly, we have not identified any significant correlation¹⁰ between the length of the description and number of signatures accumulated by the petition. Similar finding holds within various petition subcategories.

5.3.3 Circadian cycles and external influence

In this section we observe two key characteristics of the time series of signatures that we subsequently use for building our prediction model.

Circadian cycles. We binned the petition signatures and corresponding tweets into 10 minute time intervals. In addition, we aligned the petition signatures and tweets with the corresponding time of the day in the users' country. Both activities clearly follow a circadian rhythm, with the signature activity showing a stronger circadian pattern than the tweets. In particular, we can observe a peak (at around 10am) in signature activity as shown in Figure 5.8.

External effects. In order to estimate whether social media and front page affect the signatures, we performed a Granger causality [109] study between signature time series, social media and front page appearances. We examined a random sample of 30 petitions from each cluster in Figure 5.3 with their corresponding tweets and their presence in the front page of *The Petition Site* (as detailed in Section 5.4.2). Specifically, we ran the algorithm to discover the latent network structure for point processes from Linderman and Adams [182], which determines the influence of a time series on the prediction of another time series, e.g., whether signatures affect tweets or vice versa. As a result, we discovered that for the cluster containing more successful petitions,

¹⁰R=0.03 and p < 0.0005

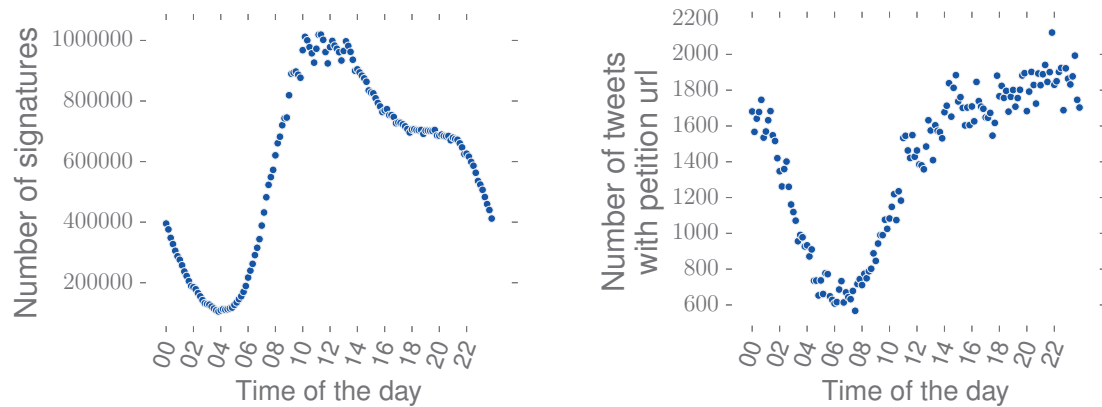


Figure 5.8 – Daily pattern of the signature (left) and tweet activity (right) with 10 minutes time intervals. Both activities can be fitted by using a sinusoidal function: $a + b \sin(2\pi(t + t_0)/24)$.

Granger causality from Twitter to the number of signatures can be observed in 90% of the cases. This fraction is lower for the remaining clusters that have less probability of success: 72%, 35%, 20% respectively. This suggests that Twitter can accelerate the signatures early in the lifetime of a petition. We confirm this later, in Section 5.5.3, by showing that it mostly influences our predictive capability early in the petition lifetime. Interestingly, in the case of petitions that were promoted to the front page of *The Petition Site*, we identified cases where signatures influenced the front page time series and vice versa equally. We further study the front page effect in Section 5.3.5. This might be explained as follows: (a) the most popular petitions get promoted on the front page (signatures influence front page ranking), (b) popular petitions that are not among the top by popularity, get promoted on the front page which result in their better performance.

5.3.4 Matching twitter users and signers

The main goal of this subsection is to establish a deeper connection between signatures and social media postings (*tweets*) beyond Granger causality. We performed a one-to-one matching between Twitter accounts and the names of petition signers/owners. Information about signers is represented in a structured format on the petitions platform. We adapted the method by Goga *et al.* [102] with matching parameters set according to our data. The procedure is explained in more details by MATCHTWITTERUSERTOIGNERS. In particular, we used the following attributes to match the profiles: (1) signer full name and Twitter name/user name, (2) signer location and Twitter user location, (3) signer petitions and tweeted petitions. We tried various combinations of these three matching dimensions, and found that using all of them resulted in the maximum number of unambiguously matched users. Contrary to the common social networks [102] petition platform does not provide neither signer photo information nor social circles.

```

MATCHTWITTERUSERTOSIGNERS(TwitterUser, Signers)
1  matches = {}
2  user_name = {TwitterUser.name,
               reversed(TwitterUser.name),
               TwitterUser.screen_name}
3  for signer ∈ (Signers.name ∈ user_name)
4     if signer.country = TwitterUser.country
5         if TwitterUser.petition_ids ∩ signer.petition_ids
6             Add(matches, signer)
7  return matches

```

The main idea behind the matching is to investigate user patterns while signing the petition, specifically whether people post a tweet after signing, or signs after posting a tweet. This fine-grained matching further allows us to trace the number of followers that signed the petition and retweeted it. As a result, we were able to match 3,157 accounts (out of 37K unique users). On average, each signer was matched to 1.47 Twitter accounts (with the maximum number of matches being 45); 2,641 accounts were matched one-to-one to Twitter accounts in a non-ambiguous manner; these are the ones included on Table 5.6. The first observation from this table is that most

Table 5.6 – Characteristics of the user profiles that were unambiguously matched between *The Petition Site* and Twitter.

| | |
|---|------------|
| Fraction of petition overlap (signed and tweeted) | 12% |
| Mean number of distinct petitions tweeted | 15.34 |
| Mean time between first signature and tweet | 26 hours |
| Mean number of tweets per petition | 16 |
| Mean number of petitions signed | 113 |
| Mean delay between signature and tweet | 19 hours |
| Median delay between signature and tweet | 15 minutes |
| Fraction of users that post a tweet after signing | 74% |
| Fraction of users that sign after posting a tweet | 26% |


people who sign a petition and post a tweet first sign the petition, and then tweet. The distribution of user sign/tweet behaviour can be depicted with the following sparkline: , where on the left of the red line we have users that first tweet and then sign. About 80% of the users perform signing and tweeting almost at the same time. In particular, 74% of users that sign and tweet almost simultaneously, tweet less than 10 minutes after signing a petition. We note that no matching scheme across websites is perfect, and this particular one might have false positives (some of the signer profiles had several identical matches on Twitter), however, it provides relevant insights on the interaction between these platforms.

Table 5.7 – Comparison of petitions that were promoted to the front page (FP) against similar petitions that were not promoted (\neg FP). A significant difference at $p < 0.01$ is denoted by **.

| | FP | \neg FP | |
|--|--------|-----------|----|
| Petitions | 75 | 75 | |
| Median signatures before t_S^* | 2,146 | 2,038 | |
| Mean signatures before t_S^* | 9,285 | 9,314 | |
| Success rate | 100.0% | 83.5% | |
| Median signatures after 2 days | 14,835 | 8,049 | ** |
| Average signatures after 2 days | 24,485 | 16,035 | ** |
| Petitions that perform better after 2 days | 60 | 15 | |
| Median among better performed | 8,198 | 2,864 | ** |
| Mean among better performed | 11,401 | 2,522 | ** |

5.3.5 Front page effect

We identified 75 petitions that were promoted to the front page, and measured whether petitions that are promoted to the front page are already on track to be successful, and if promoting those petitions causes their success. The short answer corroborates the results of the Granger causality analysis of Section 5.3.3: yes to both. To arrive to this answer, we used a standard tool from observational experiments, a *matching* study, where we matched these 75 petitions featured on the front page with 75 similar petitions that were not featured on the front page. First, we computed the number of signatures that each of the 75 petitions promoted to the front page obtained before it got promoted at time t_S^* . Second, we matched each petition promoted to the front page with one that is within a 10% range of the number of signatures but was not promoted (\neg FP) at time t_S^* . On average petitions appear on the front page after 27 hours (79 hours median) and remain for 14 days (6 days median). Statistics of these two samples are compared in Table 5.7.

Table 5.7 strongly suggests that the petitions that are promoted are not randomly chosen. Failed petitions constitute about 75% of our sample, and hence a petition chosen uniformly at random should have about 25% success rate. In comparison, the matched \neg FP set has a success rate above 80%. However, the same observations also confirm that being promoted on the front page has a drastic effect on these petitions. Beyond ensuring success (as the success rate of promoted petitions is 100%), it significantly increases the number of signatures received. For example, after only 2 days of being promoted on the front page, petitions gained almost twice as much signatures as \neg FP.

5.4 Petitions Modelling

In this section, we introduce new methods to model the evolution of the number of signatures. Our models take into account circadian rhythms, information aging, self-excitation, and external signals that influence the signature rate over time. Experimentally, these signals correspond to postings related to each petition on a social media platform, and the position in which a particular

petition was present on the front page of the petitions site.

First, we introduce a deterministic model that mimics the circadian nature of the underlying phenomenon we are studying and that includes information aging and self-excitation. Next, we extend this model by incorporating the external influence of social media and front page display, describing an end-to-end prediction pipeline.

5.4.1 Circadian rhythm and aging

The engagement of users with petitions, this is, the signature rate over time, exhibits two important temporal characteristics: circadian cycles and temporal decay. Circadian cycles are visible as daily oscillations in the signature rate, as we showed in Figure 5.8; they affect all petitions and remain stable within a particular time zone. Decay is expected due to the aging of the petition; sometimes the signature rate starts to decrease immediately, while in other cases it increases and then decreases. Based on these observations, we propose a model called Circadian rhythm with Rise and Decay (CRD). We discretize the time using a time step $\delta t = 1(h)$, while the signature rate (number of signatures between t and $t + 1$) is described as

$$\hat{s}_{p,t} = \left\{ a_p + b_p \sin\left(\frac{2\pi}{T}(t + \phi_p)\right) \right\} t^{k_p} e^{-t/\tau_p}, \quad (1)$$

where a_p is the intensity, b_p is the amplitude of the oscillation, ϕ_p its phase (with respect to an oscillation cycle of $T = 24h$), τ_p is the decay parameter, and k_p describes the initial rise in the petition activity. Parameters are fitted by minimizing the square error $E^p = \sum_{t=1}^{T_{train}} \{\hat{s}_t^p - s_t^p\}^2$, using Levenberg-Marquardt's algorithm [194]. The parameter range of τ_s is restricted to $0.5 < \tau_p < 75$ hours similarly to Kobayashi and Lambiotte [156]. We also explored an alternative Circadian with Decay (CD) model with $k_p = 0$, which in all of our experiments performed worse than CRD; we thus decided not to report on CD in this chapter.

5.4.2 Self-excitation and external influence

The CRD model is extended to incorporate self-excitation and external influence. The external influence we model comes from two sources. The first one is social media, and is expressed as $n_{sm}(t)$, the number of social media postings at time t . The second one is being featured on the front page of *The Petitions Site*, something we express as the rank in the front page $n_{srank}(t)$ that contains 10 petitions at a time, with an arbitrary value of $n_{srank} = 1,000$ for petitions not featured in the home page, which are the majority.

$$\hat{s}_{p,t} = \left\{ a_p + b_p \sin\left(\frac{2\pi}{T}(t + \phi_p)\right) \right\} t^{k_p} e^{-t/\tau_p} + \sum_{i=0}^{T_{mem}} \left(c_{self}(i) s_p(t-i) + c_{sm}(i) n_{sm}(t-i) + \frac{c_{front}(i)}{n_{srank}(t-i)} \right), \quad (2)$$

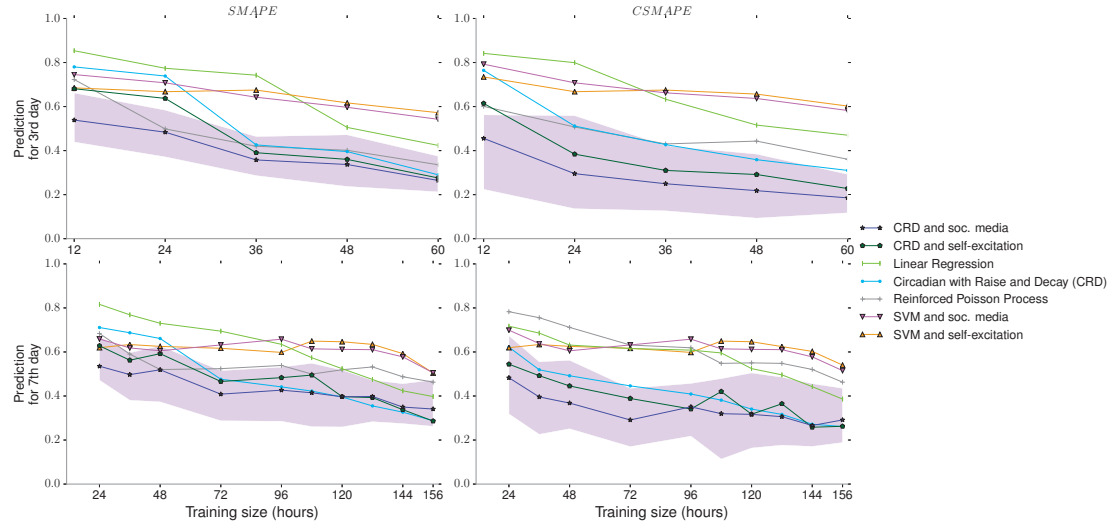


Figure 5.9 – Error in the prediction of signatures after 3 days (top) and 7 days (bottom), in terms of SMAPE (left) and cumulative SMAPE (right). For each timestamp t_s (x axis) a predictor was trained using $\{s^p(i)\}, i < t_s$ for all petitions p in the data set. The shaded area depicts the 20th and 80th percentile of the performance of the best model (CRD and social media).

where $T_{\text{mem}} = 10h$ is the size of a memory window indicating the number of time steps to be used in the estimation, and memory kernels $c_{\text{self}}, c_{\text{sm}}, c_{\text{front}}$ are, respectively, the relative importance of self-excitation, the external influence from social media, and the impact of being featured on the front page of the petitions site over time. The memory kernels are determined by minimizing the squared error after fitting CRD parameters τ_p, ϕ_p and k_p .

5.5 Experimental results

In our experiments, we consider two main prediction tasks: short-term $T_{\text{tot}} = 72$ (3 days) and long-term $T_{\text{tot}} = 168$ (1 week) prediction. We vary the size of the input that is available to each model T_{train} (from 12 hours to 71 or 167 hours respectively).

5.5.1 Metrics

Two metrics were used for calculating prediction performance of different prediction models.

Cumulative Symmetric Median Absolute Percentage Error (CSMAPE) measures the median deviation between the predicted and actual cumulative signature counts for a predicted period over N petitions.

$$\text{CSMAPE} = \text{median}_p \left| \frac{\hat{S}_p(T_{\text{train}}, T_{\text{tot}}) - S_p(T_{\text{train}}, T_{\text{tot}})}{\hat{S}_p(T_{\text{train}}, T_{\text{tot}}) + S_p(T_{\text{train}}, T_{\text{tot}})} \right|,$$

where $\hat{S}_p(T_{\text{train}}, T_{\text{tot}})$ and $S_p(T_{\text{train}}, T_{\text{tot}})$ are the predicted and actual number of signatures of the p -th petition in the prediction period $(T_{\text{train}}, T_{\text{tot}}]$, respectively. We use median to reduce the effect of outliers, similarly to previous works on web predictions [156, 324].

Symmetric Median Absolute Percentage Error (SMAPE) measures the median hourly deviation between the predicted and actual time series signature counts for a predicted period over N petitions:

$$\text{SMAPE} = \text{median}_p \frac{1}{T_{\text{tot}} - T_{\text{train}}} \sum_{t=T_{\text{train}}}^{T_{\text{tot}}} \left| \frac{\hat{s}_{p,t} - s_{p,t}}{\hat{s}_{p,t} + s_{p,t}} \right|$$

where, $\hat{s}_{p,t}$ and $s_{p,t}$ are the predicted and actual number of signatures of the p -th petition between t and $t+1$.

5.5.2 Baselines

We compared our methods against three state-of-the-art baselines.

Linear Regression. We trained the linear regression model proposed by Szabo *et al.* [281], which is a standard method for popularity prediction. The logarithm of the cumulative number of signatures $S(T)$ at time T is fitted by a linear function $\log S(T) = \alpha_T + \log S(T_{\text{train}}) + \epsilon_T$. Parameter α_T is obtained by minimizing the squared error of the prediction on a training set, and ϵ_T is a Gaussian random variable with zero mean and unit variance.

SVM with self-excitation and SVM with soc. media. A strong and simple baseline to predict complex time series is SVM regression with the Gaussian radial basis function (RBF) [64]. Similarly to our model, SVM with self-excitation and SVM with soc. media are given $s_p(t-i)$ and $n_{sm}(t-i)$ for a time window $T_{\text{mem}} = 10$ respectively. The best performing parameters for the model determined experimentally for our case are $C = 1000$ and $\gamma = 0.1$, where C is the soft margin penalty parameter and γ is the kernel coefficient.

Reinforced Poisson Process (RPP) The RPP model has been used for modelling the cumulative number of citations published by the American Physical Society [267]. The signature rate λ_t is expressed as $\lambda_t = c f_\gamma(t) r_\alpha(R_t)$, where c represents the attractiveness, $f_\gamma(t) \propto t^{-\gamma} (\gamma > 0)$ describes the aging, and the reinforcement function $r_\alpha(R_t) (\alpha > 0)$ models the ‘‘rich gets richer’’ phenomenon. The parameters c, γ, α are determined by maximizing the likelihood function.

5.5.3 Prediction

We train linear regression and SVM models for each input size T_{train} and prediction length $T_{\text{tot}} - T_{\text{train}}$. As training data, we use 70% of the petitions selected uniformly at random, and trained the model to predict the number of signatures occurring at an arbitrary hour in the future,

as well as the cumulative number of signatures up to that point. Hourly signature $s_p(t)$ and tweet $n_{sm}(t)$ counts from the training dataset were used. We then tested the prediction on the rest of the petitions. These experiments were performed 10 times, and we report their average performance. Estimation of the parameters of our model is performed in two steps. First, we estimate the parameters of seasonality and aging using the plain CRD model for each petition. Second, we train a linear regression model with self-excitation $c_{self}(i)$ and soc. media $c_{sm}(i)$ component separately using the results of the previous step. For the former, we make a one step prediction $\hat{s}_p(t)$ that we use further as a feature for the self-excitation and social media components. The process continues until we obtain the prediction for the 3rd or 7th day respectively. For the latter, we estimate the future postings (on the social media) on the training set using Eq. 1, since this information is not known. Figure 5.10 shows the hourly average of social media exposures as well as its estimation by CRD model. Upon prediction we reestimate parameters a and b of Eq. 1 based on the actual social media exposures. Further, we utilize the predicted values as $n_{sm}(t)$ in Eq. 2.

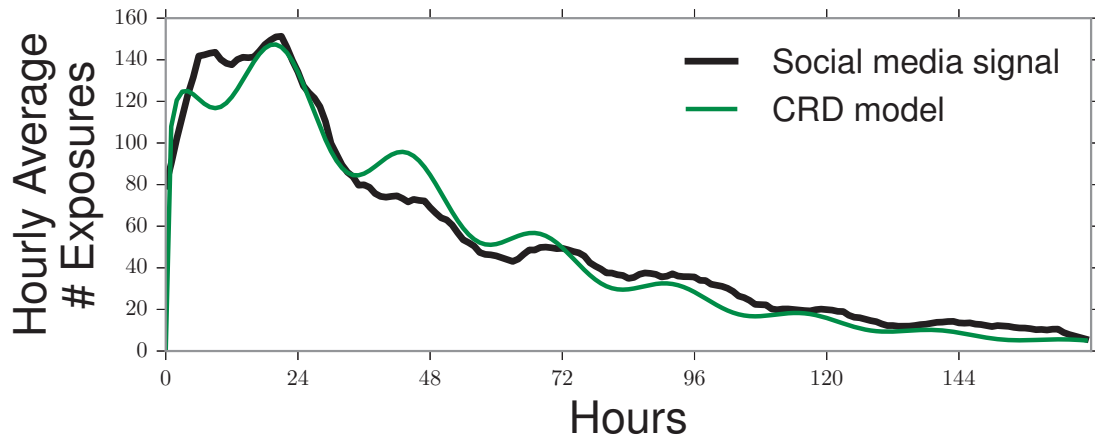


Figure 5.10 – Hourly average social media exposure for a petition during its first week. CRD model has the following parameters: $a = 102.14$, $b = 16.67$, $\phi = 8.90$, $k = 0.25$, $\tau = 37.86$

Prediction accuracy. Figure 5.9 shows the results of predicting the total number of signatures a petition gathers after 3 and 7 days. The X axis corresponds to the amount (in hours) of training data each method receives. We observe that the performance of the SVM-based methods is the lowest, linear regression and reinforced Poisson process have intermediate performance, and the performance of CRD, CRD with social media and CRD with self excitation are the highest. The latter two behave similarly, except when little training data is available, at the very beginning of a petition. In that case, CRD with social media is better than CRD with self excitation.

Given the size of the entire collection, the average improvement of considering front page information for 75 petitions is relatively small. However, among them, the front page effect brings an improvement of about 5% in terms of prediction accuracy metrics, with respect to models in which c_{front} is forced to be 0.

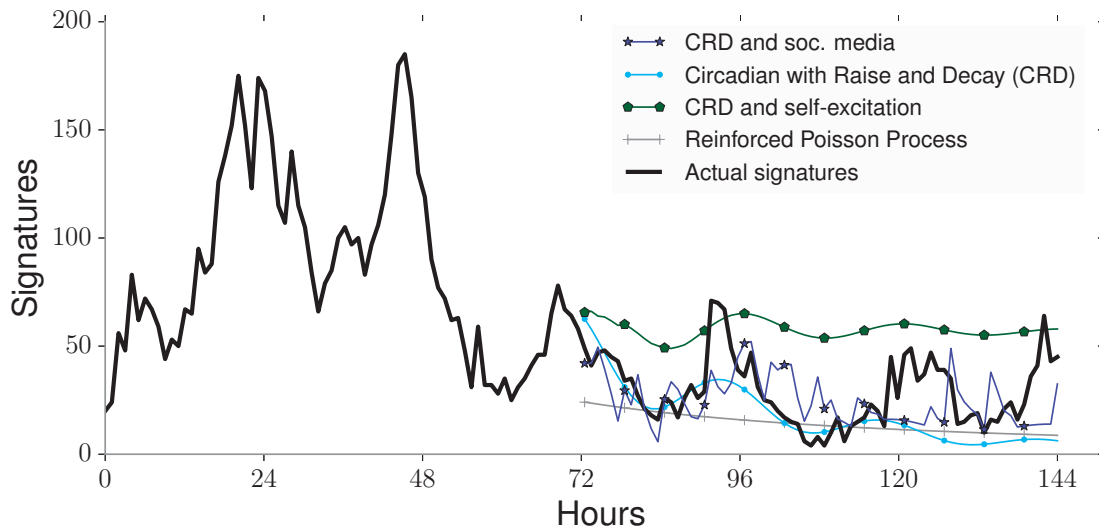


Figure 5.11 – Example showing the prediction of the number of signatures of one petition, after 3 days of observation, by a sample of methods including the best performing ones.

Figure 5.11 shows an example of a typical time series for signatures. Again, we show the advantage of incorporating information from social media in terms of generating a prediction that follows more closely the actual evolution of the number of signatures.

5.5.4 Analysis of estimated parameters

This subsection describes the analysis of the estimated parameters of the CRD models as well as its external influence functions.

Circadian Rhythm and Aging. As a by-product of modelling each petition using the Circadian with Rise and Decay (CRD) model given in Eq. 1, we obtain a distribution for each parameter across all petitions. These distributions are shown in Figure 5.12, where we are separating failed petitions from successful ones, as well as a special case of successful petitions, which are the ones promoted on the front page.

As expected, we observe that the intensity parameter a , which corresponds to the vertical shift of the series of signatures per unit of time, is higher for successful petitions than for unsuccessful ones. Interestingly, the amplitude parameter b shows that the oscillations of the series are larger for failed petitions. The growth parameter k , which influences the day at which a petition reaches its peak, shows that successful petitions tend to be more popular early on in comparison with failed petitions, and that the peak of the petitions that are promoted on the front page happens later in time—usually at the moment the petition ranks the highest on the front page. The decay parameter τ can be much larger for successful petitions, meaning that they sustain interest for a longer period of time (in the model this appears as $e^{-t/\tau}$). Finally, most of the petitions have a similar shift of the circadian rhythm, given by phase parameter ϕ , since most of them are created

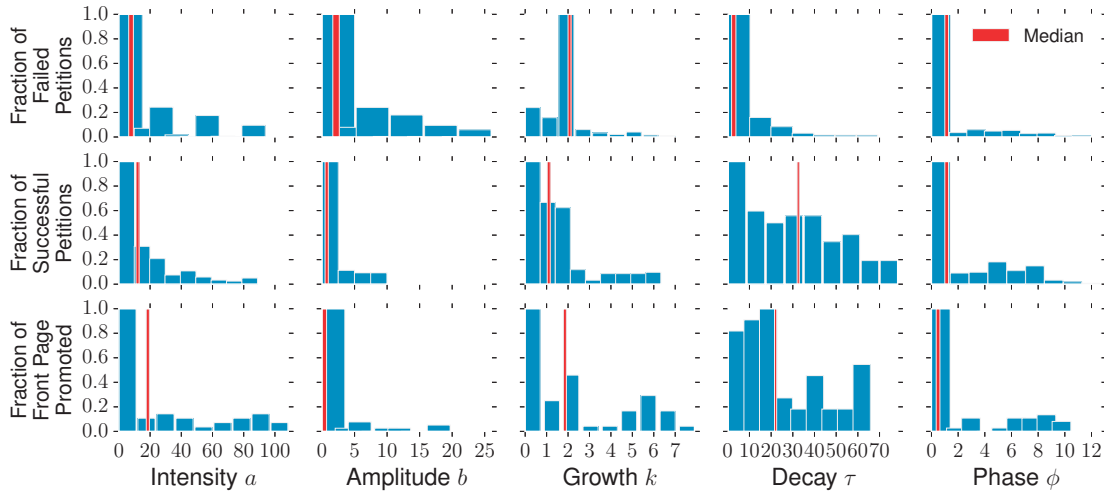


Figure 5.12 – Distributions of parameter estimations for failed petitions (top) and successful petitions (middle). We also consider separately petitions promoted on the front page, all of them successful (bottom). Details about each parameter are provided in Section 5.4.1.

in the USA and signed by people in the same country, in time zones that are close to each other (the distributions are almost equal so they are omitted from the figure).

Self-Excitation vs External Influence. Our model uses a time window of size T_{mem} hours, which allows to incorporate information from the recent past in its estimation of the future. Each of the coefficients for the influence of self-excitation $c_{\text{self}}(i)$, social media $c_{\text{sm}}(i)$, and front-page effect $c_{\text{front}}(i)$ can be seen as a time-indexed vector reflecting the importance of different moments of the recent past for each specific influence across successful petitions. If we are predicting the popularity on $t + 1$ hour, the influence function corresponds to the vector of size T_{mem} that contains the effect of each $t - i$ hour of observation from self-influence, social media or front-page influence, where $i = 0, 1, \dots, T_{\text{mem}}$. The centroids of these vectors are shown in Figure 5.13.

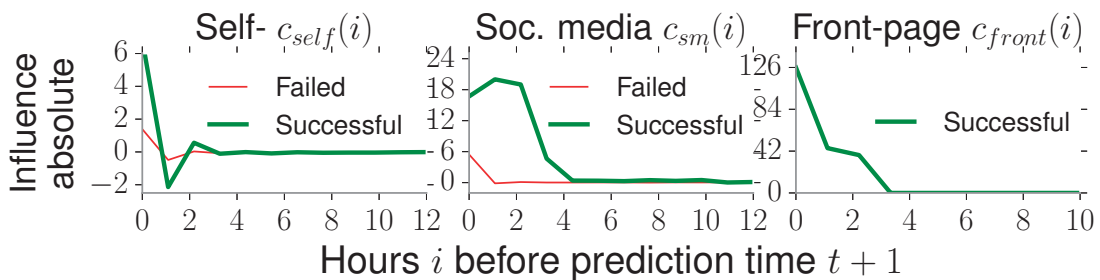


Figure 5.13 – Influence function estimation for self-excitation, the influence of social media, and the front-page effect. A value of i on the X axis refers to the median influence of this aspect i hours in the past. On each plot, the Y axis presents an absolute scale for successful and failed petitions and shows the multiplicative effect on the number of signatures. Petitions promoted on the home page are all successful.

Several interesting observations can be made from Figure 5.13. First, self-excitation seems to be

largely memory-less, with the immediately preceding step being the most influential element. Second, social media (Twitter in this case) has an influence that can last up to four hours for the successful petitions, and peaks at about 3 hours; this means a posting at time t mostly affects the signature rate between times $t + 2h$ and $t + 3h$. Failed petitions seems to be less affected by social media with a short memory of 1 hour. Third, the front-page effect has an effect that lasts about two hours. In absolute terms, social media has a stronger effect than self excitation, and being featured on the front page has a stronger effect than social media activity: it results in a big boost of signatures, consistently with our observations from Section 5.3.5.

5.6 Conclusions

Online user engagement is a complex phenomenon, which can be better captured when considering potential influences that might be affecting it. In general, interdependent phenomena across websites are less studied than phenomena happening on a particular website. In this chapter, we studied an important form of engagement—signing an online petition—we modeled two external influences: activity on social media, and promotion to front page. In both cases, we demonstrated significant improvement in modelling and predicting engagement when those influences are taken into account. In addition, we showed that the circadian rhythm of human activity, and the fact that interest decays over time, also need to be considered.

We analyzed the effect of social media and found it to be impactful in two ways. First, at a micro level, as demonstrated by the matching of people signing a petition and then posting about it shortly afterwards. Second, at a macro level, where we analyzed the effect of Twitter on the signature rate using a Granger causality test, and showed significant improvement in prediction accuracy when using social media—improvements that are particularly important to reduce the amount of time/data needed to perform an accurate prediction. We were also able to determine that the effect of an increase in postings on Twitter lasts for about 5-6 hours and peaks at about 3-4 hours. These findings are probably relevant beyond online petitions, as many campaigners in social media (e.g., advocating for brands, causes, or candidates) also perform similar activities in order to boost user engagement.

Specifically for online petitions, we showed that online petitions that are successful tend to peak early and to continue receiving attention for longer. In other words, it is not just about having a “strong start,” but about being able to sustain this engagement day after day. Petitions can be boosted by activity on social media, and/or by featuring them prominently to a large audience of potential signatories, as demonstrated by the front page effect that we have modeled and measured. These findings are relevant for people running other types of campaigns, and may be particularly important for crowdfunding campaigns.

In general, running a successful campaign on the web requires sustained attention and punctual interventions. In that context, interpretable models that can provide actionable insight about how a campaign is evolving are vastly more useful than opaque models, even if the latter were to

provide small advantages in terms of prediction accuracy.

Recommendation for the future campaigners. *First*, we have seen that most of the successful petitions experience increased user participation during the first initial days, thus, it is important to prepare the material and the meticulous plan on how to engage more people and explore the possibilities to using multiple channels to convey the idea of the petitions as early as possible. *Second*, we recommend the activists to establish the connection with the petition platforms' owners and request them to feature the petition on the front page. This showed to have the strongest effect on the user gain compared to social media. *Third*, in this work we have not made a great distinction between various topics of the petitions, however, we have seen some evidence that users are less likely to post sensitive topics (LGBTQA, women rights, abuse) on social media. Therefore, other means to promote and spread the information shall be found for such topics.

Future Work. We believe that this chapter is an important step towards better modelling and predicting how reinforced information spreads online. It can be extended in a number of ways. In terms of new methods, it would be interesting to explore how the effects of several petitions on each other could be modeled, and how social media communities, defined both topically and through network structures, could be incorporated into our models. Moreover, impact functions could be represented through parametric distribution functions. In terms of enhancing the prediction accuracy, further sources of social media, and new features, could easily be incorporated into our model. Since we are modelling the petitions at an individual level, it might also be interesting to build and compare our model to a batch model and apply it over specific clusters of petitions. Finally, a prediction using a stochastic Hawkes process might be compared to the deterministic one presented in this chapter.

Facilitating Part

6 Efficient Document Filtering Using Vector Space Topic Expansion and Pattern-Mining

The Case of Event Detection in Microposts

Automatically extracting information from social media is challenging given that social content is often noisy, ambiguous, and inconsistent. However, as many stories break on social channels first before being picked up by mainstream media, developing methods to better handle social content is of utmost importance. In this chapter, we propose a robust and effective approach to automatically identify microposts related to a specific topic defined by a small sample of reference documents. Our framework extracts clusters of semantically similar microposts that overlap with the reference documents, by extracting through frequent pattern mining combinations of key features that define those clusters. This allows us to construct compact and interpretable representations of the topic, dramatically decreasing the computational burden compared to classical clustering and k-NN-based machine learning techniques and producing highly-competitive results even with small training sets (less than 1'000 training objects). Our method is efficient and scales gracefully with large sets of incoming microposts. We experimentally validate our approach on a large corpus of over 60M microposts, showing that it significantly outperforms state-of-the-art techniques.

6.1 Introduction

Social media—and in particular Twitter—have reshaped the news industry. Billions of users contribute live updates on events that are happening in their vicinity using social media platforms, thus allowing not only journalists but also citizens or stakeholders to follow breaking news in near real-time. A number of activities such as journalism, activism, or disaster recovery can be facilitated by means of social media [291]. However, *extracting relevant information from social media in a reliable and efficient manner* still remains a challenge.

In this chapter, we tackle the problem of efficiently identifying documents that are relevant to a given query. A query in our context is represented by a small set of textual documents that are relevant to a specific topic of interest. As a particular instance of this problem, we focus on *extracting microposts that are relevant to a given event* in the following.

Several methods have been proposed for this problem, we summarize them in Section 6.2. One approach is to apply some semantic matching (e.g., edit distance or lexical overlap) between the description of the event and a series of microposts. In general such methods identify messages that are similar to the query of interest, but are computationally expensive and yield a poor recall, i.e., fail to produce comprehensive results. Another approach is to leverage knowledge bases, thus taking into account semi-structured or unstructured descriptions of well-known entities and events in the matching process. In such approaches, however, domain-specific knowledge is generally underrepresented. Finally, a number of classification and clustering approaches have been proposed recently for this problem. These approaches are typically computationally expensive, require a well-defined and accurate metric of similarity between two texts, and usually require a large corpus of labeled data, thus limiting their potential domain of application.

We propose a novel methodology that is both efficient and requires a very small labelled training set while performing on par with methods that utilize much larger training datasets or that are computationally very expensive. Specifically, we propose a technique based on frequent itemsets (patterns) extracted from the query. We measure the distance between various query items to extract the patterns. Specifically, we leverage text similarity metrics that rely on word embeddings that are pre-trained on very large collections of microposts. Our solution is task-independent and is evaluated on the complex task of event extraction from social media streams. We show that our method outperforms state-of-the-art baselines (lexicon, embedding similarity, k-nearest neighbours and classification based on word embeddings) and that it is computationally efficient compared to etalon-based (k-NN) approaches. Broadly speaking, we show how syntactic or semantic clustering can be efficiently replaced by semantic pattern extraction for event detection.

Our contribution. We present a new method for filtering microposts that match a specific query. The query is a textual description of the topic of interest; in our running example and in our experiments, this topic of interest is an event. Based on this description, our method automatically generates a small *seed set* of microposts, based on text similarity. Then, we apply frequent pattern (itemset) mining on the seed set. Among the extracted patterns, we select those that are associated with semantically homogeneous groups of microposts. These patterns (called topical patterns) are then compared to an incoming stream of messages in order to select all microposts matching the query. Our technique is presented in detail in Section 6.4.

In summary:

- we describe an efficient solution requiring minimal annotations to filter and identify microposts that are relevant to a given query;
- we present an extensive evaluation with multiple baselines and show that our approach

- outperforms them over a large dataset of 3TB of Twitter messages spanning two years;
- finally, we release the source code of our technique as well as a collection of annotated event messages for different classes of events.

The rest of the chapter is organized as follows. We start with an overview of related work in §6.2. We describe our process for collecting the data from social media as well as identifying seed microposts related to the events in §6.3. We present our topical document extraction model and compare it to existing models in §6.4. We experimentally evaluate the models and discuss them in §6.5. Finally, we summarize our results and outline future work in §6.6.

6.2 Related Work

On demand extraction of online content based on a seed document or query is challenging and typically requires large amounts of annotated data used to build supervised models [9, 29, 70, 144, 247, 298]. In some cases, the query is not known a priori and is only implicitly represented through a set of documents that are relevant to a topic of interest [152, 165, 177]. Similarity-based approaches tend to be inefficient [66] and difficult to scale.

Another approach to tackle the topical document detection problem is to rely on content clustering and topic modelling (see Table 2.6). However, these approaches work best for document extraction relating to past events (thus, specific details are known and can be used for the extraction) and are hard to adapt to a stream processing context (where neither particular details nor dates are known ahead of time). A number of techniques leverage a lexicon that can effectively and accurately represent a given topic, yielding a high precision but a rather low recall. For instance, [217] uses pseudo-relevance feedback to improve recall for the lexicon-based methods, which however hampers their capacity to detect new events [244]. Finally, a range of new deep learning architectures have been recently proposed to both represent the document in a semantic space as well classify the documents by topics based on their vector space representation [152, 165, 177]. Such methods are supervised and require a large corpus of annotated data.

Contrary to the variety of methods described in Table 2.6, such as classification methods requiring a substantial amount of annotated data, or methods based on query similarity, that require pairwise similarity comparison between each query text and input data, we propose a method that is more efficient and accurate than aforementioned techniques, as well as require a very small training dataset. Furthermore, we show that our approach can achieve high performance with minimal initial input.

6.3 Data Collection and Seed Extraction

In this section, we introduce the data sources we use and our data collection process (Section 6.3.1), explain how seed messages describing the events are extracted from the Twitter

Stream (Section 6.3.2), and describe the data annotation process that is used to evaluate the quality of our results (Section 6.3.2).

6.3.1 Data collection

The topics we use in our examples correspond to large-scale events that are covered widely by international media. Specifically, we focus on terrorist attacks: uses of violence to create fear, for ideological purposes, and aimed at civilians or noncombatant targets [196]. We create a database of attacks by integrating information from Wikipedia and from the Global Terrorism Database (GTD).

Wikipedia data. ¹ We crawled all attacks in 2014 and 2015, which are available on 24 separate pages indexed by month, and contain information on 650 events. This list applies the definition of violence from a non-state actor, without considering the restriction of being against civilians. Hence, author of this work manually annotated the events to discard those perpetrated against combatants or armies. The attack was added to the database when the agreement between the annotators was 100%. A total of 592 events were selected and are listed on Table 6.1, along with information on the country and type of the attack as described on Wikipedia².

Global Terrorism Database (GTD). ³ GTD contains over 15K records from the same period, including minor and major incidents involving civilians. The GTD dataset was created to enhance the initial descriptions we obtained from Wikipedia. We use GTD and Wikipedia attack descriptions as the input queries.

Twitter data. We performed a rate-limited data collection from Twitter, collecting up to 5%⁴ of all tweets posted during 2014 and 2015 using Twitter’s Streaming API. The dataset resulted in over 3TB of data, out of which 60M tweets were posted in English. We relied on the NLTK python library⁵ to detect the language. This is the dataset over which we all extraction techniques are evaluated in the following.

6.3.2 Seed extraction

Our method leverages a small set of seed microposts (training dataset) that are later used to extract patterns to determine which microposts should be selected. Since we have at our disposal a database of attacks along with their description (see above), we use this information to the seed microposts. The quality of this seed set is of utmost importance, in the sense that it ideally

¹https://en.wikipedia.org/wiki/List_of_terrorist_incidents

² Information from Wikipedia contains a variety of metadata, including the location and date, a summary of the event, the number of casualties, and the suspected perpetrator.

³<https://www.start.umd.edu/gtd/>

⁴5 Twitter accounts were requesting Twitter Streaming API during the 2 years.

⁵<http://www.nltk.org/>

| Country | Bombing | Attack | Shooting | Raid | Events | Tweets |
|---------------|---------|--------|----------|------|--------|--------|
| Iraq | 84 | 2 | 0 | 0 | 90 | 811 |
| Israel | 3 | 55 | 9 | 13 | 84 | 1001 |
| Nigeria | 48 | 5 | 6 | 0 | 72 | 2408 |
| Afghanistan | 42 | 3 | 3 | 0 | 53 | 657 |
| Pakistan | 39 | 7 | 4 | 0 | 51 | 3189 |
| Egypt | 25 | 2 | 3 | 0 | 31 | 344 |
| Yemen | 19 | 2 | 0 | 0 | 22 | 175 |
| Syria | 21 | 0 | 0 | 0 | 21 | 483 |
| Somalia | 16 | 1 | 1 | 0 | 18 | 251 |
| Cameroon | 10 | 0 | 4 | 0 | 15 | 76 |
| India | 6 | 1 | 2 | 0 | 14 | 353 |
| Lebanon | 14 | 0 | 0 | 0 | 14 | 74 |
| Philippines | 6 | 2 | 2 | 0 | 12 | 54 |
| Libya | 10 | 0 | 1 | 0 | 11 | 29 |
| Mali | 3 | 1 | 5 | 0 | 11 | 279 |
| Kenya | 0 | 6 | 3 | 0 | 10 | 1135 |
| United States | 0 | 1 | 8 | 0 | 9 | 2614 |
| Saudi Arabia | 6 | 1 | 1 | 0 | 9 | 49 |
| Turkey | 6 | 1 | 1 | 0 | 8 | 101 |
| Chad | 6 | 0 | 0 | 0 | 7 | 246 |
| Niger | 1 | 0 | 1 | 0 | 7 | 82 |
| France | 1 | 2 | 1 | 0 | 6 | 506 |
| China | 3 | 1 | 0 | 1 | 6 | 220 |
| Tunisia | 1 | 1 | 3 | 0 | 5 | 347 |
| Ukraine | 1 | 0 | 0 | 0 | 5 | 419 |
| Australia | 0 | 1 | 2 | 0 | 3 | 684 |

Table 6.1 – Wikipedia dataset characteristics. We list the top 25 countries and the top 4 attack types as described on Wikipedia. The column “Event” corresponds to the total number of events for a country, while “Tweets” contains the number of matching microposts for each attack description.

should contain many relevant microposts and few irrelevant ones. This training dataset is directly provided to the method described in Section 6.4. To build the ground truth and identify relevant tweets, we start with a set of terrorist attack descriptions that contain the relevant information about the type of the event, people involved, number of causalities, locations, particular details about the timelapse of the event etc (as described in the beginning of the section), which we then use as a proxy to measure micropost similarity (TF.IDF score) to event description. Moreover, we made sure all tweets we match the location where the attack happened. We extract an initial seed of relevant microposts, we apply the TF.IDF-based algorithm described in Snippet 1. The algorithm identifies whether a micropost describe any of the attacks in the database. First, we initialize the set of matched attacks for a given tweets to be empty in Line 1. Further in Line 3 we eliminate the attacks with no clear evidence of attack location. The lower bound in this condition is meant to solve potential inconsistencies in the way the source datasets we use manage timezones.

Next, for the events (Line 5) in our dataset we compute an event matching score as the sum of

Algorithm 1 Identification of seed microposts. *Events* correspond to the set of event descriptions, D_{tweet} is the database of microposts, and θ a threshold for text similarity.

```

MATCHEDEVENTS(Events,  $D_{\text{tweet}}$ ,  $\theta$ )
1  matched_events = []
2  for event  $\in$  Events
3      if MATCHLOCATIONKEYWORD(event,  $D_{\text{tweet}}$ )
4          matched_events = matched_events  $\cup$  event
5  for event  $\in$  matched_events
6      event_score = GETIDF(event[keywords]  $\cap$   $D_{\text{tweet}}$ [keywords])
7      if event_score  $<$   $\theta$ 
8          REMOVEFROMLIST(event, matched_events)
9  return matched_events

```

| Terrorist attack description | TF.IDF similarity | Extracted tweet |
|---|-------------------|---|
| A suicide bomber attacked a police academy in 5th police district, Kabul city, Kabul province, Afghanistan. In addition to the suicide bomber, 25 people were killed and 25 others were wounded in the blast. The Taliban claimed responsibility for the incident. | 0.30 | #KCA #VoteJKT48ID guardian: Taliban attack parliament building in Kabul with suicide car bomber and RPGs |
| Assailants opened fire on Dr. Waheedur Rehman in Dastagir area, Karachi city, Sindh province, Pakistan. Rehman, a Karachi University professor, was killed in the attack. No group claimed responsibility for the incident. | 0.33 | F.B Area Block-16 Me Firing Se Karachi University Shoba Ablagh-e-Aama Ke Assistant Professor Syed Waheed Ur Rehman S/O Syed Imam Janbahaq. |
| Assailants abducted seven Coptic Christian Egyptians from their residence near Benghazi city in Benghazi district, Libya. The seven Egyptians were killed the same day. No group claimed responsibility for the incident. | 0.43 | #IS in #Libya claims responsibility for abducting 21 Egyptian #Christians, http://t.co/32l8YCLL35 #Egypt #ISIS |
| Two suicide bombers opened fire and then detonated inside a classroom at the Federal College of Education in Kano city, Kano state, Nigeria. In addition to the two bombers, at least 15 people were killed and 34 others were injured in the blasts. Boko Haram claimed responsibility for the attack. | 0.44 | Boko Haram claims responsibility for Kano bomb blast, share photo of the male suicide bomber: B... |
| A rocket landed inside a community and detonated in Sdot Negev regional council, Southern district, Israel. There were no reported casualties in the blast. No group claimed responsibility for the attack. | 0.56 | #BREAKING: A rocket from #Gaza hit Sdot Negev Regional Council in southern #Israel. No damage, no injuries |

Table 6.2 – Examples of seed microposts related to the attacks.

the TF.IDF values for the keywords in the intersection between the attack description and the micropost (Line 6). Finally, if the obtained score is lower than a threshold (Line 7) the event is discarded from the resulting list. To set the similarity threshold θ between the event descriptions and the microposts, we manually annotated (as described in the Annotation paragraph below) a random sample of 300 tweets related to some attack for various thresholds. As a result, we

picked the threshold to $\theta = 0.27$ as this value yields the best precision (95%) on our sample.

In total, we obtained 17'093 seed microposts related to terrorist attacks. Table 6.2 shows some examples of attack descriptions and the related microposts. The tweet ids that correspond to the attacks as well as the attack database are available on <https://github.com/toluolll/ShortTextFiltering>.

Data annotation We adopt a consistent process to annotate the microposts and to determine the quality of our results (Section 6.5). Specifically, the author of this thesis and a colleague of hers manually annotate the relevance of the microposts selected by the algorithm (or by any of the baselines).

6.4 Method description

This section describes the method we propose to filter relevant microposts for a given query. Our method first represents each input query by a *seed set*; it mines “topically homogeneous” patterns from the seed set. The patterns are then placed into an index which is used for efficient filtering, i.e., to select the microposts that contain a pattern and are hence relevant to the input query. We give an overview of the whole method in Section 6.4.1. Next, we describe the text similarity metric (Section 6.4.2) and pattern extraction approach (Section 6.4.3). Finally, we compare this approach to alternative clustering methods in Section 6.4.4.

6.4.1 Overview

Figure 6.1 outlines the major steps of our approach, which combines two key insights: (1) an appropriate distance metric can be leveraged to estimate their topical similarity between microposts, and (2) we can take the best of two worlds by using pattern extraction techniques to combine both supervised and unsupervised learning.

Similarity metric adjustment. The only part of the process that requires human supervision is the selection and the adjustment of the distance metric between documents, which has to be performed once per corpus – in the present case just once as there is a single input corpus containing all microposts.

The metric adjustment assumes the following:

- all possible pairs of documents (microposts) existing in the input corpus belong to one of the following classes: identical (x, y_{ident}), similar (x, y_{similar}), topically related (x, y_{related}), or unrelated ($x, y_{\text{unrelated}}$); and
- there exists a distance metric d that defines the following order on the pairs of documents:

$$d(x, y_{\text{ident}}) < d(x, y_{\text{similar}}), \quad d(x, y_{\text{similar}}) < d(x, y_{\text{related}}), \quad d(x, y_{\text{related}}) < d(x, y_{\text{unrelated}})$$

If those two assumptions hold, we can determine a threshold d_{related} that separates pairs of

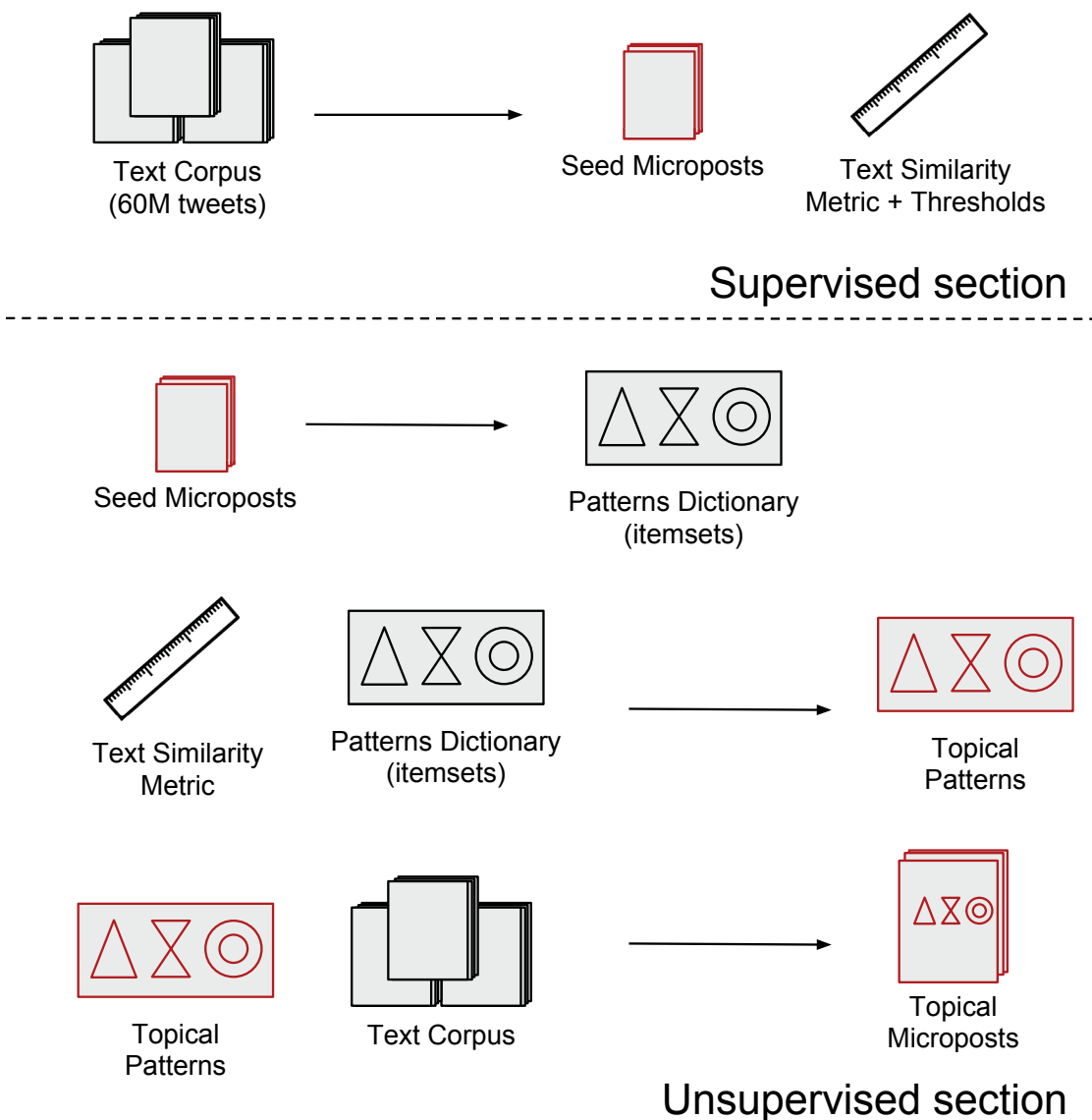


Figure 6.1 – Pipeline overview.

topically related documents, from unrelated pairs of documents. Fortunately, there is a large body of literature on this topic and we do not need to invent a new text similarity metric. The threshold value d_{related} can then be estimated empirically on a validation set, for a target type of documents (in this chapter, microposts). Details on this are provided next in Section 6.4.2.

Pattern mining. We extract frequent patterns (itemsets) from the seed microposts and use them to filter the input to produce a larger set of relevant microposts. Given that there might be many such patterns potentially (with many patterns not representing any relevant subset of microposts), we need to filter those patterns. Towards that goal, we note that *a relevant pattern induces a topically homogeneous set of microposts*.

We call a pattern *topically homogeneous* if all the microposts that it matches are topically related to each other. To measure topical homogeneity, we estimate the expected pairwise distance between a pair of microposts selected by a pattern by randomly sampling pairs of microposts containing the pattern. If the expected distance is lower than a threshold value d_{related} estimated during the similarity metric adjustment step, then the pattern is considered to be topically homogeneous.

Pattern extraction has several benefits compared to other approaches:

- Unlike most of supervised learning approaches, the performance of our method (especially the precision) depends less on the size of the seed. The resulting accuracy of text selection is boosted by the effectiveness of the distance metric and the selected thresholds.
- Compared to the instance-based machine learning methods (like k-NN), pattern extraction is more flexible and efficient from a computational perspective. In general, for every new document, k-NN would require computing the distance between this document and all the seed documents, which in the most simple case yields a complexity of $O(|\text{Docs}| \cdot |\text{SeedDocs}| \cdot \text{AvgWordsPerDocument})$ (if we are using distance metrics that only weigh word overlap of the documents). With large training sets and an elaborate distance metric (like the one we are using in this chapter), k-NN rapidly becomes impractical. Another important drawback of k-NN is the necessity of a proper set of negative samples. In the context of topic extraction, one needs to create a set of neighboring topics, which is often a very complex task. Our method on the other hand does not require negative samples explicitly. The computational complexity of pattern extraction in general is NP-hard, though limiting the length of the patterns and the textual features dramatically limits the number of possible patterns that can be extracted from the seed documents. With topically homogeneous patterns, we reduce the number of elements to take into account to a few hundreds or thousands even for large seeds. Every pattern is a conjunction of a limited number (maximum 5 in our case) of textual features. In that case, checking whether a document contains at least one topical pattern can be done in sublinear time (in terms of the size of the text) with proper indexing techniques.
- Extracted patterns are easy to interpret.
- The support values of the patterns can be used to rank the documents with respect to their relevance to a topic.

6.4.2 Text similarity metric

Our method requires a metric for measuring text similarity (see above). In this chapter, we picked Word Mover's Distance (WMD) proposed by [165]. WMD attempts to find an optimal transformation between documents d and d' . The method is solving the following linear optimization task with constraints: WMD attempts to find an optimal transformation between documents d and d' . The method is solving the following linear optimization task with constraints:

$$WMD(d, d') = \min_{T \geq 0} \sum_i^n \sum_j^m T_{ij} c(i, j)$$

subject to:

$$\sum_j^m T_{ij} = 1/n, \quad \forall i \in \{1, \dots, n\}$$

$$\sum_i^n T_{ij} = 1/m, \quad \forall j \in \{1, \dots, m\}$$

where:

- n, m are the number of words in documents d and d' ,
- T_{ij} is the weight of word i (WDM works with nBOW representations of documents, so a word weight is equal to $1/|d|$) from document d that is going to be transferred to word j of document d' , and
- $c(i, j)$ is the “traveling” cost between words d_i and d'_j .

In [165], the traveling cost was selected to be equal to the Euclidean distance between vector representations (in the `word2vec` embedding space) of words. According to our experiments, however, the Euclidean distance suffers from the so called “curse” of dimensionality for a high-dimensional vector space (over 100 dimensions) as most of the distances end up having similar values. On the other hand, the cosine similarity empirically yields less skewed distance distributions. Hence, we rely on cosine similarity in the following, and the distance metric between words for our method becomes:

$$c(i, j) = 1 - X(d_i)X(d'_j) \quad (6.1)$$

where $X(m)$ is the vector representation of word m . For our method we use the FastText [177] vector model with a dimensionality of 300.

We now assess how WMD values can be leveraged to identify related documents. To identify reliable threshold values for topical similarity in the WMD space, i.e., to determine d_{ident} , d_{similar} , and d_{related} , we sample document pairs for every WMD value interval from 0.1 to 1.2 with a step of 0.1. The sample sizes were equal to 100, giving us 1'200 document pairs in total. For every WMD interval sample, the author of this thesis and a colleague of hers checked the pairs and labelled them as (1) copies, (2) semantically identical texts, (3) topically related texts, or (4) different texts. The distribution is shown in Figure 6.2. Based on those results, we selected a WMD value of $d_{\text{related}} = 0.5$ as threshold for topical similarity. A pair of documents with WMD smaller than 0.5 has a probability of more than 90% to be topically related (as it is close to 80%

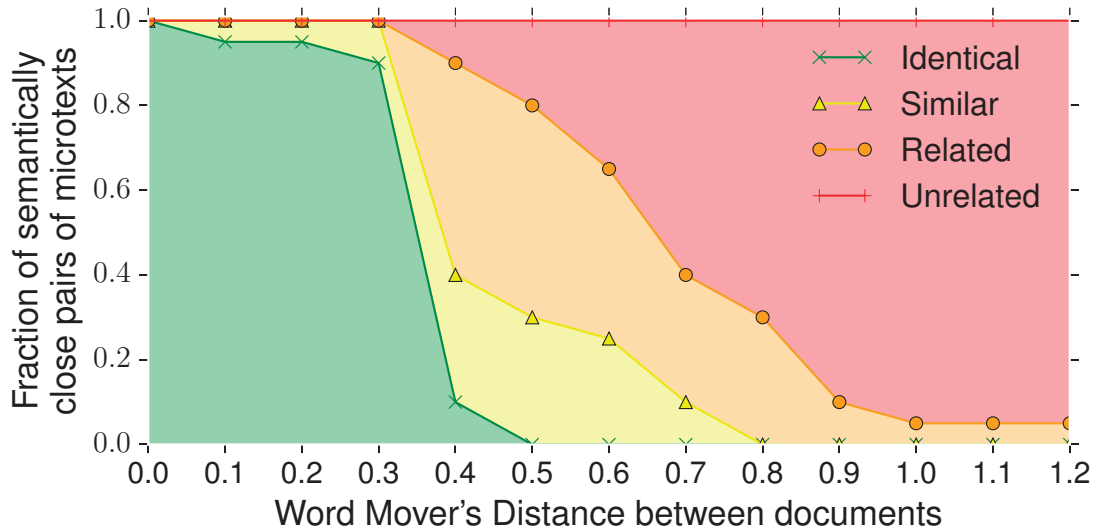


Figure 6.2 – Relatedness of documents as function of their pairwise distance.

for $WMD=0.5^6$ and increases for lower values of WMD).

6.4.3 Pattern extraction

The problem of mining patterns (or associations) from item sets was introduced in [5]. Pattern extraction from text can be formally introduced as follows: Let $D = i_1, i_2, \dots, i_m$ be a set of m distinct attributes (we call these attributes *markers* in the context of this chapter). Each document in a corpus T has a unique identifier TID and is associated with a set of markers (itemset). As such, it can be represented as a tuple $\langle TID, i_1, i_2, \dots, i_k \rangle$. A set of markers with k items is called a k -itemset. A subset of length k is called a k -subset. An itemset is said to have a *support* s if at least s documents in T contain the itemset.

Originally, the pattern extraction task consists of two steps: (1) mining frequent itemsets, and (2) forming implication rules among the frequent itemsets. In our method, for the point (2), we concentrate on the extraction of topically homogeneous itemsets, i.e., patterns that are present in documents that are topically related to each other.

To answer whether a set of documents containing the itemset is topically related, we estimate the mean WMD value between the documents by calculating the average WMD value for a sample of documents pairs. If the resulting value is less than the threshold value for WMD topical relatedness ($d_{related}$), then we consider the set of documents as being topically related.

⁶ Robustness tests of various WDM thresholds against short text filtering as presented in Table ???. Threshold of 0.4-0.45 results in low recall (about 1.5-2 times less than for 0.5). Threshold of 0.55-0.6 results in 1.5-2 times higher recall (this improvement reduces for larger training sizes) and lower precision. In terms of F1, 0.5 precedes 0.6 for larger number of training examples and vice versa for smaller number of training examples.

| Seed | Synset |
|---------|--|
| bomb | bombing, bomber, explosives, detonated |
| shot | shooting, shoot, shots |
| kill | kidnap |
| nigeria | kenya, ghana, uganda, benin |
| huge | massive, enormous, tremendous |
| gas | hydrocarbon, combustion, sulfur, methane |

Table 6.3 – Examples of extracted synsets.

Pattern mining algorithm. Starting with the full dictionary of markers present in the seed microposts, we use the ECLAT algorithm [317] for pattern mining. ECLAT is a scalable (due to initial parallelization) depth-first search family of pattern mining algorithms. The minimum support of an itemset is defined by a minimum sample size of document pairs that is required to reliably estimate the mean of the pairwise distances between documents that contain the pattern. In our case, we chose this value to be more than 40, so that assuming normal distribution of pairwise distances we will have enough pairs of documents to estimate the mean distance.

To speed-up the process of pattern extraction, we add two pruning criteria. First, we stop growing topically homogeneous itemsets, since all their supersets will be producing subclusters of the current cluster of topically related documents. Second, we also define a maximum pattern length; in our experiments, we only use patterns composed of at most 5 markers.

Types of attributes. For this chapter we only use two type of attributes - stemmed and lowercased words presented in the text, and synsets (clusters of semantically similar words) that we describe below. For stemming we use the Porter stemmer. We also remove stop words, since their absence helps to significantly reduce the amount of irrelevant patterns.

Sets of related words (synsets). We leverage word embeddings constructed as explained in Section 6.5.1 to construct sets of related words from our Twitter dataset. We call them “synsets” in the following, but note they are not necessarily synonyms of each other, but closely related words. As shown by [262], skip-gram models in combination with cosine similarity yield similarity estimations on par with more complex state-of-the-art techniques. Author of this thesis and her colleague manually evaluated several cosine similarity thresholds ranging from 0.5 to 1.0. A threshold of 0.65 resulted in the most coherent pairwise semantic word proximity. For the 30K most frequent words in the whole Twitter dataset, we construct the synsets greedily in a “snowball” fashion, i.e., for each word we identify a set of most semantically similar words; each of those words is then used in turn to find semantically similar words, and so on. Each word is added to the synset if it is similar to at least 30% of the words that are already there, reducing topic drift. Some examples of synsets are shown in Table 6.3. On average, synsets have 3.6 terms, with a median of 3 terms.

6.4.4 Patterns vs clustering: a case of coverage

One may ask whether frequent itemsets cover a significant part of topically-related documents, particularly when compared to potentially higher-recall methods, such as clustering.

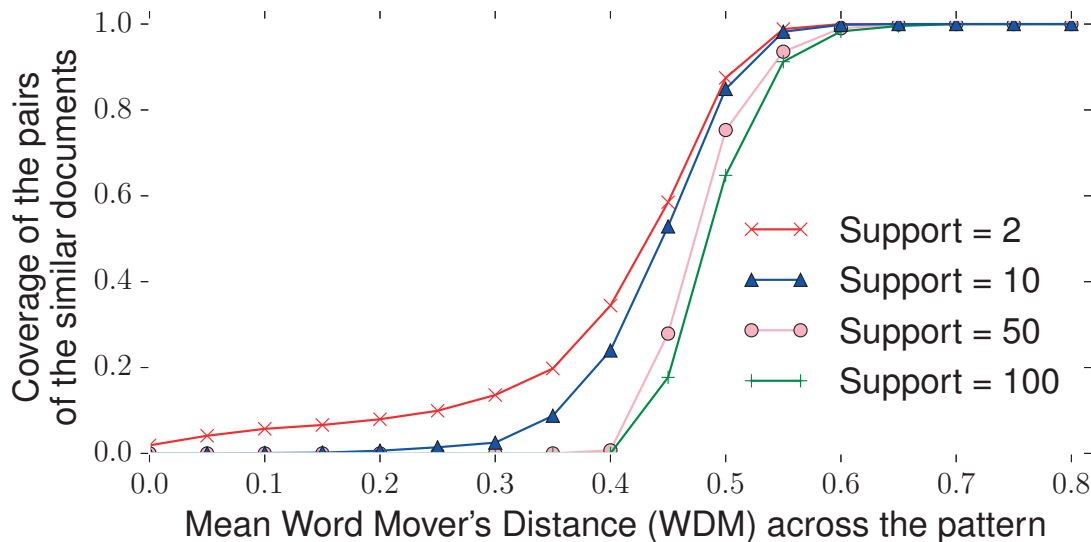


Figure 6.3 – Estimated part of topically-related document pairs that can be covered by patterns. The graph is generated for the training size of 5,000 examples.

To examine this question, we sampled 200K document pairs that were considered at least topically related (WMD value smaller than 0.5), i.e., those that were produced/filtered based on the patterns extracted from the 5,000 training set. Then, for every pair we looked for itemsets that were included in both documents and had a given support value. The distributions of coverage (percentage of pairs covered by patterns with defined parameters) as a function of minimum pattern support and mean of WMD are shown in Figure 6.3. We observe that at least 90% of the sampled pairs can be covered with topical patterns with support greater than 10, and approximately 75% of the pairs can be covered with topical patterns with a support greater than 50. This means that in the most pessimistic case, the selected support value of 40 guarantees that we cover at least 75 – 80% of the documents related to the topic.

6.5 Experimental Results

To evaluate our approach, we compare it to several baselines (Section 6.5.1) in terms of precision and relative recall (Section 6.5.2); results are summarized in Section 6.5.3 and discussed in Section 6.5.4.

6.5.1 Baselines

We compare our methods against a number of state-of-the-art baselines that cover the main approaches for topical document extraction (Section 6.2). Specifically, we implemented a Pointwise Mutual Information (PMI) based lexicon expansion, and three methods based on word embeddings: a semantic centroid classifier, a FastText similarity ranking technique, and a proximity-based (kNN) method.

These baselines require training data; we use synthetically generated training examples which are nevertheless of high quality. The training examples are obtained using the seed selection method described in Section 6.3.2, which as discussed is more than 95% precise. We vary the size of the training set available to each baseline N_{train} from 1 to 10K randomly sampled positive examples. For methods requiring negative examples, we select an equally-sized set of negative examples, which are sampled from all microposts that are not in a seed dataset. The assumption here is that the presence of tweets related to terrorist attacks in the general dataset of tweets is negligibly small, so false negatives will be minimal. However, this heuristic is not appropriate for k-NN, so we had to slightly modify it, as explained below.

For the methods based on word embeddings, we trained a FastText skip-gram model [39] over the 60M English tweets described in Section 6.3 with default parameters: vector size – 300, window size – 5, negative sampling, minimum words count – 10.

1. Corpus-based PMI. In this chapter, we used the PMI-based term scoring method described in [217] that measures the difference between the relatedness of a term t to (1) an event class a and (2) a non-event class $\neg a$. This is defined as follows:

$$PMI(t) = \log_2 \frac{p(t|a)}{p(t|\neg a)} \quad p(t|a) = \frac{\text{count}(t, a)}{\text{count}(MARKERS, a)}$$

where $p(t|a)$ and $p(t|\neg a)$ are the probabilities of t appearing in event-related and not event-related microposts, respectively. *MARKERS* can be any syntactic representation of a text; we use unigrams and bigrams in this evaluation. The top ranked unigrams and bigrams for the union of all the events (terrorist attacks) are shown in Table 6.4.

2. Semantic centroid classification. As a second baseline, we use a linear classifier trained on the semantic representations of the microposts, as described by [152]. Every word in the training data is represented by a set of 300-dimensional features that correspond to the embedding representation of each word. We derive this feature vector by averaging each dimension of the words in the sentence.

3. FastText-based similarity ranking. This approach enhances the previous baseline by learning how to combine word embeddings into a text representation as described by [144]. Thus, the resulting text representations are better to distinguish topical tweets. To find similar tweets, each short text is sent through the classification model so that task-specific embeddings

| | | | |
|------------------|----------------|--------------------|------------|
| Unigrams | bernardino | gunmen | bombers |
| | haram | kano | militants |
| | jerusalem | garissa | mandera |
| | parenthood | baghdad | siege |
| | copenhagen | synagogue | bombings |
| | leytonstone | blasts | tunisia |
| Uni- and bigrams | bernardino | planned_parenthood | kano |
| | san_bernardino | copenhagen | garissa |
| | haram | in_peshawar | baghdad |
| | boko_haram | leytonstone | synagogue |
| | jerusalem | shooting_in | injured_in |
| | parenthood | gunmen | in_nigeria |

Table 6.4 – Top features extracted by PMI using unigrams only (top) or unigrams and bigrams (bottom). Results are obtained taking the entire training data for all events (terrorist attacks) as the positive class.

are obtained. Then, the representations of the target tweets are compared (using cosine similarity) to the unlabelled ones. Several similarity threshold are tested to select the final results; a distance threshold of 1.1 radians results in the best accuracy.

4. k-NN-based on WMD metric. We use the k -nearest neighbors (k-NN) method described in [66]. The distance between each new, unseen element to all training examples is computed using the WMD metric as described in the previous sections.

k-NN requires negative samples in addition to the positive samples. Using as negative examples a sample of documents from the main document corpus will not be helpful in this case. Taking into account the abundance of possible topics in the microblogging space, a small training sample of tweets not related to the target topic cannot guarantee that a topic of a randomly picked document will be present in the training dataset (as this probability will be very small). So, for the majority of documents, all documents from the training dataset are equally far and majority vote provides a nearly random answer.

To avoid this problem and be able to use the k-NN approach (since it is one of the very few methods that can be effective even with small training samples) we modify it so that it can work without negative training samples. The idea is to assign the positive class to the documents that have at least K positive documents from the training seed in their near proximity. We selected K to be equal to 3 (as lower numbers significantly decrease precision) and the radius to 0.5 WMD (according the result that we discussed in the previous section).

6.5.2 Metrics and their estimation

We report standard information retrieval metrics: precision, recall and F-measure. Precision was evaluated using 3 random samples of 200 tweets each, which were labeled by the author

and her colleague of this chapter with annotators agreement of 95%⁷, following the procedure described in Section 6.3.2. Computation of recall is challenging since human annotation of the full corpus of 60M tweets is beyond our resources; thus, we rely on relative recall. Relative recall is computed by taking the union of all microposts that are positively labeled by all methods. We report recall as: $RR_{\text{method}} = \frac{TP_{\text{method}}}{\sum_{m \in \text{all_methods}} TP_m}$, where RR stands for relative recall, and TP_{method} is a true positive rate for a given method. Finally, we report F-measure as follows: $F_{\text{method}} = \frac{2 * P * RR}{P + RR}$.

6.5.3 Results

Results are summarized in Table 6.5. Our method performs better than the baselines in terms of both precision and recall when we allow it to use 5'000 or more automatically selected seeds. Synset-based variation of the attributes also performs better than the baselines in terms of F-measure when we use 100 or more automatically selected seeds. In addition, we compare the results of the micropost extraction task using a less sophisticated approach for the synset generation, e.g. when synsets are generated by using the top-10 most similar words for each of the 30K most frequent words in the dataset. This experiment yields a reduction of 3% and 1% for recall and precision respectively, compared to the results obtained using synsets generated by our method (see textit“Ours - unigrams” and “Ours - synsets” in Table 6.5). The baselines perform worse in terms of both precision and recall when the number of positively labeled examples are over 5K; in principle this cannot be attributed to the training set quality, as according to our tests it was 95% precise as discussed in Section 6.3.2.

Our method, in contrast, loses recall on smaller input sizes but wins precision depending on the number of automatically selected seeds to be used, with the best values of F-measure obtained when using around $k = 5,000 - 7,000$ automatically selected seeds. Overall, we observe that our method with any number of $k \geq 100$ automatically selected seeds outperforms all baselines in terms of F-measure, even in cases where they use 10,000 manually labeled items⁸. Table 6.6 presents samples of patterns and associated documents that are generated by our method.

6.5.4 Discussion

Our approach is most similar to the nearest neighbors approach (kNN); indeed, the results of both approaches on small trainings sets are comparable. However, our approach does not have the limitations that kNN has:

- Unlike kNN, we do not require objects with negative class labels. Collecting samples of tweets that are not related to the topic is often impractical, as the number of potential topics

⁷Microposts describing an event from the past lead to the most annotation disagreement, since those were not specifically reflecting an event that has recently happened. We included such examples to the training set.

⁸We have also performed a robustness test against noise in the training set (1%, 2%, 5%, 10% of noise). As a result, P and R were equivalent to the results presented in Table 6.5 for any size of the training set. However, the number of the extracted patterns on average were 20% lower.

| | Synthetic training examples (baselines) or seeds (ours) | | | | | |
|-------------------------|---|--------------|--------------|--------------|--------------|--------------|
| | 100 | 500 | 1,000 | 5,000 | 7,000 | 10,000 |
| Extracted volume | | | | | | |
| PMI | 1.7M | 953K | 473K | 290K | 140K | 60K |
| Centroid | 1.0M | 429K | 427K | 196K | 135K | 115K |
| FastText | 2.0M | 664K | 259K | 97K | 116K | 101K |
| KNN | 3.6K | 13.4K | 31.2K | - | - | - |
| Ours - unigrams | 6.2K | 15.1K | 33.5K | 112.2K | 149.8K | 171.2K |
| Ours - synsets | 5.0K | 16.8K | 26.1K | 114.3K | 143.4K | 169.1K |
| Precision | | | | | | |
| PMI | 0.005 | 0.005 | 0.01 | 0.020 | 0.050 | 0.100 |
| Centroid | 0.004 | 0.030 | 0.080 | 0.120 | 0.210 | 0.220 |
| FastText | 0.005 | 0.040 | 0.100 | 0.190 | 0.240 | 0.270 |
| KNN | 0.810 | 0.740 | 0.670 | - | - | - |
| Ours - unigrams | 0.880 | 0.760 | 0.690 | 0.570 | 0.540 | 0.460 |
| Ours - synsets | 0.880 | 0.770 | 0.690 | 0.560 | 0.550 | 0.460 |
| Recall | | | | | | |
| PMI | 0.409 | 0.601 | 0.621 | 0.623 | 0.627 | 0.629 |
| Centroid | 0.544 | 0.600 | 0.634 | 0.641 | 0.671 | 0.703 |
| FastText | 0.557 | 0.630 | 0.643 | 0.646 | 0.703 | 0.722 |
| KNN | 0.102 | 0.265 | 0.32 | - | - | - |
| Ours - unigrams | 0.090 | 0.269 | 0.348 | 0.682 | 0.745 | 0.787 |
| Ours - synsets | 0.130 | 0.283 | 0.384 | 0.701 | 0.775 | 0.797 |
| F1 score | | | | | | |
| PMI | 0.010 | 0.010 | 0.020 | 0.039 | 0.093 | 0.173 |
| Centroid | 0.008 | 0.057 | 0.142 | 0.202 | 0.320 | 0.335 |
| FastText | 0.010 | 0.075 | 0.173 | 0.294 | 0.358 | 0.393 |
| KNN | 0.181 | 0.390 | 0.433 | - | - | - |
| Ours - unigrams | 0.163 | 0.397 | 0.463 | 0.621 | 0.626 | 0.581 |
| Ours - synsets | 0.227 | 0.414 | 0.493 | 0.623 | 0.643 | 0.583 |

Table 6.5 – Evaluation results for the micropost extraction task of the four baseline methods against our method. The average size of a synset pattern was 204, 373, 465, 439, 451, 462 attributes for 100 - 10,000 training examples respectively.

to cover can be very large.

- Our method is more robust to large training samples (which are potentially more noisy) and complex distance metrics. Word Mover’s Distance has a computational complexity $O(w^3 \log(w))$, where w is the average length of a document. Multiplied by the size of a training set $|T|$ and the size of the text corpus $|D|$ makes it impractical for large collections. In our case, we were not able to get results for training sets larger than 1’000 documents for kNN, as the extraction process on a cluster of 50 machines was still running after several days.

Empirically, topics are mixtures of sub-topics. Compared to the baselines, our method shows stable performance across all seed sizes. It is noticeably more selective, especially on smaller samples, where the non-kNN methods perform quite poorly. With more seeds, our methods still maintains a high precision and outperforms the baselines in terms of recall. The level of precision

| Mean WDM | Pattern | Support | Micropost examples |
|----------|--------------------------------------|---------|--|
| 0.436 | attack, claim, {egypt} ¹ | 99 | “isis claims responsibility for tunisia attack that killed 13 people” “islamic state claims responsibility for tunisia attack statement reuters” ... |
| 0.476 | boko, {attack} ² | 686 | “boko haram gunmen attack nigeria villages kill 43publish date feb 13 2014 new vision #bokoharam” “flash buhari explains legal basis for accepting suvs after boko haram attacked him in kaduna in 2014” “boko haram gunmen attack nigeria villages kill 43” ... |
| 0.375 | boko, bomber, femal, haram | 41 | “alleged boko haram suicide bombers dressed as females die in an accident in borno see photos” “see photos of the 13 year old female boko haram suicide bomber” “ttw today s news suspected boko haram female suicide bombers blow up market in nigeria” “breaking boko haram attacks maiduguri again as female suicide bombers did this via” |
| 0.474 | bomber, polic, suicid | 128 | “turkey suicide bomber wounds 5 turkish police during r #trending #news #startups #howto #diy #android #howto #apps” “muslim b tch blows up police dog heroic k9 diesel blown up by female suicide bomber in paris #mcgnews” “french suicide bomber killed during raid was blonde woman yelling help me to police before she detonated bomb bb4sp” “french honor diesel hero police dog blown up by suicide bomber during terrorists last stand #jesuischien” “is says dutch suicide bomber struck iraq police #middleeast #politics” “police suicide bombers one of them 11 female target nigeria market” |
| 0.345 | attack, government, militant, somali | 53 | “al shabaab militants attack somali government building at least 5 dead mogadishu reuters a #breakingnews” “somali militants raid government base at least eight people are killed in an attack by suspected al shabab mi” “world somali police say 7 dead in attack on baidoa government hq mogadishu somalia suspected islamic militants” |
| 0.399 | attack, blast, kabul | 87 | “blast and gunfire in kabul s diplomatic district second attack in a day fighting season ends in the battlefield begins in kabul” “rt updated story deadly blast at kabul airport as taliban attacks surge” “after blast in kabul taliban say they made suicide attacks against guesthouse for foreigners” “rt after blast in kabul taliban say they made suicide attacks against guesthouse for foreigners” “updated story deadly blast at kabul airport as taliban attacks surge” |
| 0.440 | claim, sinai | 52 | “rt isis branching out islamic state affiliate claims attacks in sinai” “isis in sinai claims attack on hamas in gaza” “breaking just in islamic state s wilayat sinai account claims responsibility for the coordinated attack in northern sinai” “#israel under attack radical sinai salafi group claims responsibility for rockets fired at eilat via” “#breakingnews egyptian islamic state group affiliate claims deadly sinai attacks” “sinai based militants claim rocket attack on israel #egypt #israel #sinai” |

Table 6.6 – Examples of patterns and associated documents generated by our approach. Mean WDM refers to mean pair-wise WDM document distance. Pattern is presented as a combination of stemmed words.

¹ egypt, syria, libya, tunisia, cairo

² attack, attacking

is guaranteed by the topical compactness of the extracted patterns, which is a key element of our method. The increase in recall is also expected for higher numbers of seed documents as it allows us to cover more subtopics and consequently more relevant microposts.

One possible reason why the Centroid and FastText approaches do not significantly benefit from growing seed sizes is that they conceptually try to find a clear center in the embedding space that is supposedly the pivot of the topic. This is in contrast to an empirical observation, which

shows that each topic is typically a mixture of numerous smaller subtopics that have little overlap between each other. For example, here are several tweets that were considered to be related to terrorist attacks, but that do not have much in common:

- “Amnesty International Says Boko Haram Kills Thousands in Nigeria’s Baga Town ...”
- “Palestinian Kidnapped Near #Jenin #westbank”
- “ISIS releases internet video purportedly showing American journalist Steven Sotloff’s beheading”
- “Twin suicide bomb blast rocks northern #Cameroon village”
- “As usual terrorist attacks take place in Sinai, while military will strike back against university students and women in rest of #Egypt”

PMI adjusts to general words like “attack”, “terrorist”, “massacre”, “killed”, etc., which explains the reason why it shows a relatively high recall on training sets of different sizes. This also explains the low precision values: that generality does not allow PMI to discern terrorism from other topics related to casualties, deaths, or violence.

Precision, in our approach, slightly degrades with larger training sets. We attribute this to a growing number of outliers that are included into training samples.

6.6 Conclusions

In this section we introduced a generic and flexible framework for semantic filtering of microposts. Our framework processes microposts by combining two key features: semantic pattern mining and document similarity estimation based on the extracted patterns.

Compared to the baselines, our method shows stable performance across all document seed set sizes. It is noticeably more selective, especially on smaller samples, where the non-kNN methods perform quite poorly. In particular, our approach leverages word embeddings that are trained on event-specific microposts, thus enhancing the event representation on particular Social Media platforms. Our approach makes no use of external knowledge bases (e.g., WordNet) nor of linguistic tools (parsers) that are computationally expensive. Our empirical results show that our algorithm is efficient and can process high-velocity streams, such as the Twitter stream, in real-time. We demonstrate its efficiency on a large corpus and showed that our topical extraction outperforms state-of-the-art baselines.

Future Work. Our current method of topical pattern extraction uses a very simple set of features that represent the documents: stemmed unigrams and synsets. Our plans is to expand it with other potentially more expressive features like n-grams, synonyms, entity types, etc. With a richer set of features our method could adapt to finer topical nuances.

To make our approach more efficient, we plan to optimize the pattern extraction process even further. The idea is to apply restrictive pattern growing techniques that prevent the emergence of

multiple patterns based on similar sets of documents.

Another potential area of enhancement is to assign weights to extracted patterns, as well as estimating the confidence with which our approach attributes documents to the target topic. The number of matched patterns, their support values, average pairwise distances, and other observable values could provide a rich input for the prediction of a confidence value.

Finally, as embeddings are usually highly dependant on the input, mixed embeddings (e.g., trained on both Social Media content and Wikipedia) could be leveraged to make our method more robust.

Data and code availability. Code and anonymized data are available at <http://github.com/toluolll/attacksProfiling>.

7 Template Induction over Unstructured Email Corpora

We have seen that a great volumes of content on the Internet are either duplicates or near-duplicates of each other. For example, in Chapter 3 we have discovered that about 40% of the content of the mobilization campaigns are near-duplicates of the several original messages. Another instance of this phenomena is the case of news media articles that usually are multiple variations of the original material published by only few media sources. Thus, there is a need to organise such content into a structured template that both users and machines would benefit from. Therefore, we aim to develop a solution that would assist in template induction over unstructured text. About 1% of the whole email user base produces a lot of repetitive content. As a result, we chose email documents as our use case in this Chapter to study the extend to which repetitive content can be templated.

Unsupervised template induction over email data is a central component in applications such as information extraction, document classification, and auto-reply features. The benefits of automatically generating such templates have been shown for structured data, e.g. machine generated HTML emails. However much less work has been done in performing the same task over unstructured email data.

We propose a technique for inducing high quality templates from plain text emails at scale based on the suffix array data structure. We evaluate this method against an industry-standard approach for finding similar content based on shingling, running both algorithms over two corpora: a synthetically created email corpus for a high level of experimental control, as well as user-generated emails from the well-known Enron email corpus. Our experimental results show that the proposed method is more robust to variations in cluster quality than the baseline and templates contain more text from the emails, which would benefit extraction tasks by identifying transient parts of the emails.

Our study indicates templates induced using suffix arrays contain approximately half as much noise (measured as entropy) as templates induced using shingling. Furthermore, the suffix array approach is substantially more scalable, proving to be an order of magnitude faster than shingling even for modestly-sized training clusters.

Public corpus analysis shows that email clusters contain on average 4 segments of common phrases, where each of the segments contains on average 9 words, thus showing that templating could help users reduce the email writing effort by an average of 35 words per email in an assistance or auto-reply related task.

7.1 Introduction

Template induction, the technique of generating a skeleton of repeated content based on previously seen examples, has seen substantial success for structured content such as web pages, where metadata such as the underlying DOM¹ provides multiple presentational and structural signals that can be exploited algorithmically. These structures can be useful for tasks such as automatic labeling, plagiarism detection, duplicate detection, and structured information extraction. Despite the success of template induction for structured data, we have found little prior research for the same task, but for data that is not explicitly structured (i.e. plain text). This duality has difficult implications for a domain such as email or other unstructured content. Despite often having some amount of structure, emails almost always contain some significant portion of freeflowing plain text that have, to date, not yet been sufficiently modeled in induced structured templates.

In this chapter, we develop a template induction algorithm that focuses on the plain text content. We use email as the target domain, as we envision two potentially high-impact applications of templates generated for plain text content: 1) structured information extraction, where particular important pieces of information in the email are extracted; 2) document autocompletion, where the system suggests content to add during the composition of a document, based on the already present content the user has added; and 3) facilitation of the spam detection and filtering (similarly to the set up of Chapter 6). All of this use case are not only suitable for the general email use case, but also within a broad online activism domain, where message autocompletion could save time for the activists to communicate their ideas to various people. The intuition driving our approach is that the more often a person composes an email, a document or a message, the more likely they are to repeat words and phrases in their generated content. One could imagine that in a near-ideal situation, a near-complete document would be generated by the system based on only a few keystrokes from the user. From now on we will focus only on the email use case in the Chapter, however, most of the findings can be generalized to other types of documents.

Information overload, the idea that we are receiving more data than we can effectively process, has steadily worsened since the advent of email as a form of communication. The majority of efforts to manage the growing influx of messages has centered on organizing or filtering messages [35, 208]. These efforts focus on improving the situation by affecting the incoming stream of emails in some way. Only recently have there been efforts to reduce the cost required to actually respond to an email [147].

We look to further reduce the response cost to users by investigating the usefulness of a template

¹http://wikipedia.org/wiki/Document_Object_Model

suggestion mechanism that is initiated when composing emails from a set of automatically constructed templates.

Although the potential benefits are considerable, targeting email documents provides nontrivial challenges in addition to the lack of explicit structure. Unlike the public domain for web pages, email documents, private messages, cloud word documents are virtually always considered private, making it difficult to obtain training data. This is a sharp contrast to the freely observable and highly-structured web domain. Consequently, there are vastly fewer suitable datasets available for this kind of investigation, and we were unable to find any direct prior research on this topic. Given this setting, the primary task becomes one of establishing that template induction can be effective in the unstructured content domain.

A template creation method consists of two parts: First, clustering similar messages together. Second, for each cluster, determining the parts which are considered “fixed” and storing the information in a standard representation, which is the produced template. In this chapter, we focus on the latter.

To determine the fixed regions, we use an implementation of the suffix array data structure [301], which is efficient in space and time complexity, and can be easily parallelized. We show that the quality of produced templates created with our approach is consistently better than the baseline regardless of the quality of the clusters and with better latency performance. The results of the public corpora analysis determine that text suggestion would affect a significant number of users and would save them a significant volume of writing in an autocompletion task. In addition, we show that the portion of emails detected as fixed-text is larger than for the baseline, which would allow an automated information extraction system to focus on fewer parts of the emails when extracting transient information.

To determine the effectiveness of using a suffix array to generate templates, we compare it against a standard baseline approach for template creation. We test both methods on a synthetically generated corpus as well as a publicly available corpus of emails from the Enron Corporation². The results of these experiments indicate that the suffix array serves as a superior approach to correctly identifying an optimal number and span of fixed regions.

The main contributions of this work are:

- An extensive analysis of the feasibility of unstructured documents, such as emails, templating for emails sent by a given user or bulk sender;
- A scalable technique that results in high quality templates regardless of the clustering quality;
- A novel application of the generalized suffix array algorithm to detect common phrases over similar documents; and
- An evaluation of the efficiency and quality of both the suffix array and baseline approaches.

²<https://www.cs.cmu.edu/~enron/>

The rest of the chapter is organized as follows. Section 7.2 discusses prior work in the area of email template creation. Section 7.3 describes the template creation task in detail as well as our suffix-array based approach to generating templates from pre-formed clusters. Section 7.4 describes the experimental setup for comparative analysis, and concludes with the results of those experiments. Finally, Section 7.5 presents further applications and potential future work based on the results of this study.

7.2 Related Work

In this section we discuss state of the art techniques related to this work. Specifically, we cover methods for email content mining (e.g. spam classification, labeling, and threading), template induction (for web and for emails), and autocompletion systems.

7.2.1 Email content mining

Most email content mining techniques are originally derived from more established web information extraction methods, such as template or wrapper induction [20, 115, 164, 166, 259]. In addition to information extraction, email content mining techniques also span document classification tasks, such as spam filtering, automatic labeling, as well as document clustering for email threading.

Spam classification. The popularity and necessity of email as a communication medium has also made email a popular target for spam attacks. Simple manual spam classification filtering rules have given way in the last decade to more complex and effective machine learning systems that now serve as the defacto defenders tasked with detecting and removing spam messages from many inboxes [18, 223, 319]. More recently, these binary classification techniques have expanded beyond textual classification to include much richer feature sets, collaborative filtering techniques, and peer-to-peer and social networking ontology-based semantic spam detection techniques [35, 43].

Label classification and ranking. The expansion of email as one of the primary communication media in both personal and commercial use has led to the notion of *email overload*, in which users become overwhelmed when the rate of incoming messages in their inbox outpaces the rate at which they can process those messages [71]. Automatic email foldering has been proposed to alleviate this problem [32, 160, 179]. Bekkerman *et al.*[32] discuss the challenges of applying traditional document classification techniques, such as naive Bayes classifiers and support vector machines to the email foldering task. However, these works present methods for email organization based on the underlying context of each message rather than occurrences of specific strings.

In many of these scenarios, sparsity of labeled data remains a hindrance to accurate and robust model development. Kiritchenko and Matwin [155] deal with the sparsity issue by using a

co-training algorithm to build weak classifiers, then label unlabeled examples, and add the most confident predictions to the labeled set. Somewhat similarly, Wendt *et al.*[302] utilize a graph-based label propagation algorithm to label unlabeled emails from a small set of labeled emails, but do so at the template level to improve scalability in very large mail provider systems.

Possibly one of the most popular and widely known automatic email organization systems is Google's Gmail priority inbox, which distinguishes between important and non-important emails by predicting the probability that the user will interact with the email (e.g. open, respond) within some time window from delivery [3].

Threading. Email threading is another solution to the *email overload* problem that assists in inbox organization and can furthermore reduce the user's perceived inbox load by clustering emails from the same conversation together [173]. Current threading techniques cluster messages together by header information, such as sender, subject, and subject prefixes (e.g. 'Re:', 'Fwd:'). Personal correspondence emails are generally threaded very accurately using this technique, however, many commercial emails, such as purchase receipts, tracking numbers, and shipment confirmations are often split into multiple threads due to their different subject lines despite arriving from the same sender domain and belonging to the same semantic thread. Ailonp *et al.*resented techniques to thread such commercial emails through leveraging email templates and learning temporal causal relationships between emails from similar senders [8]. Similarly, Wanga *et al.*tempt to recover implicit threading structures by sorting messages by time and construct a graph of conversations of the same topic, however their analysis is limited to newsgroup style conversations [299].

7.2.2 Template induction

Web data is generally formatted in a human-readable format which a machine renders but might not necessarily understand. Take for example an event web page whose body contains images and text organized either using HTML tables or division tags and CSS. When rendered, the information will often be presented in an appealing way to the user, however the rendering machine will not know what rendered information is most pertinent to the page's purpose (e.g. event title, location, date, start and ending time). Web extraction techniques have been proposed to solve this issue of extracting information structured for human consumption from data.

The web is comprised of over a trillion documents in the public domain [13], many of which are dynamically created and generated using templates. Hence, web information extraction is a well explored topic and is often closely coupled with template induction techniques. Since many emails utilize HTML markup and it is estimated that nearly 60% of emails are created from templates (e.g. B2C emails) [8], it is natural to also apply similar web-based techniques for template induction and information extraction on emails. However, emails are among the most sensitive data on the Internet. Hence, very little research has been presented on email template induction due to privacy constraints, although recent work has proposed methods for enforcing anonymity in web mail auditing [73].

In web template induction, multiple training examples for a single template can often be created by tweaking the parameters comprising a dynamic URL (e.g. the product ID in the URL of an e-commerce page). In email template induction, multiple training examples for a single template must first be derived by clustering emails together which have a high probability of being derived from the same template.

While there is very little published work on using structural templates for processing commercial email data, templates have been used for annotating semantic types within the DOM trees of emails [321] and used in hierarchical classification of emails [302]. To our knowledge, no techniques have yet been published that propose template induction for plain text email content.

[172] proposes a methodology to summarize the short messages about same events into their shorter representation. However, it has two major drawbacks that are not suited for the email use case. First, clustering of the messages relies on the edit distance which does not scale well over average size emails and might produce clusters where all words are different. Second, the method requires strong preprocessing.

7.2.3 Autocompletion

The general task of autocompletion can benefit the user by minimizing errors or reducing the time to issue a query, thereby minimizing repetitive typing and resurfacing familiar results. Autocompletion can also benefit the provider by reducing load and potentially improving cache hit rates.

Autocompletion of text in the context of search queries has primarily focused on predicting short strings and incorporating large indexing data structures and processing times [30]. Hyvönen and Mäkelä [135] generalized the idea of syntactic autocompletion on a semantic level by autocompleting typed text into categories instead of words.

7.2.4 Assisted email composition

SmartReply is the first email-specific machine-learned tool designed to assist the user in composing email. The tool is built on recurrent neural networks (one to encode the incoming email and the other to predict possible responses) that automatically suggests replies to email messages [147]. The work presented here aims to provide assistance that is learned specifically from an individual email sender, whereas SmartReply provides canned responses meant to satisfy as many users as possible. In an assisted email composition context SmartReply would infer intent, while our work would detect content that was written before by the same sender.

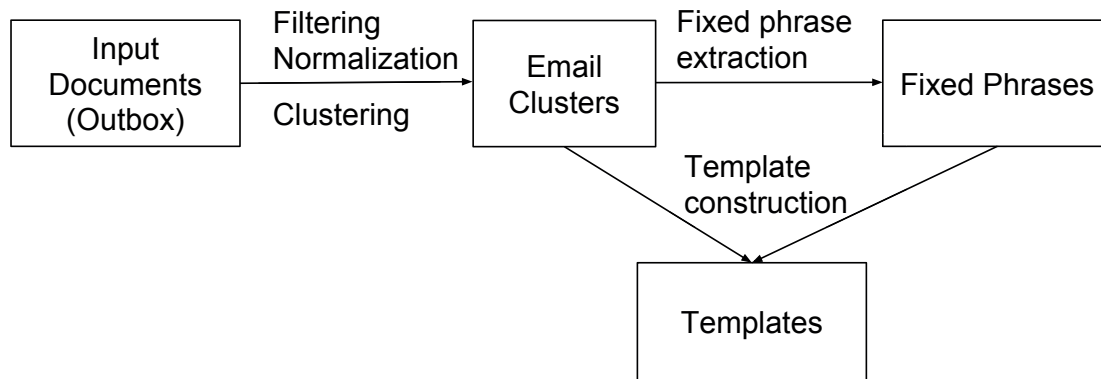


Figure 7.1 – An overview of the clustering and template induction.

7.3 Methodology

The design of the template induction system is shown in Figure 7.1. We first describe the preprocessing and clustering steps which generate the input for the template induction algorithms. We then present the template induction algorithms.

7.3.1 Preprocessing

During the preprocessing phase, we perform necessary enrichment and filtering of the raw input emails to make them suitable for use as experimental input for comparison.

For each email in the raw corpus, we determine if the message is a reply (its subject begins with the pattern `re:`) or a forward (its subject begins with the pattern `fwd:`) of a prior email. All other emails are treated as original messages. Any non-original message is analyzed, and if all of the content of the message is simply a quote of the original message, the email is filtered out. Of the remaining emails, all messages determined to be non-English (or consisting primarily of non-English content) are additionally filtered out. Each of the remaining emails is then tokenized and each produced token is a structure that carries the original form of the token as well as its lemma, using the WordNet database for lemmatization [202].

7.3.2 Clustering

After normalization and filtering, we then cluster similar emails as follows. Given the task at hand, a natural basis for similarity is to first group all emails sent from the same email address together. Each of these initial groups is then further processed to produce the final set of email clusters, which are then passed as input to the template generation algorithms. Since we aim at the efficiency and scalability, we present a greedy clustering methods as described below. However, other clustering techniques could be used depending on the volumes of the input data.

Formally, let D_i and D_j be two emails in our corpus, represented as term count vectors. We define the distance between these two emails to be:

$$\delta(D_i, D_j) = \left\| \frac{D_i}{\|D_i\|_1} - \frac{D_j}{\|D_j\|_1} \right\|_2 \quad (7.1)$$

which we use as a measure when constructing the final set of clusters. Note that based on this definition, $0 \leq \delta \leq \sqrt{2}$ ($\sqrt{2}$ indicates orthogonal document vectors). A maximal distance indicates that two documents have disjoint vocabularies, while minimal distance indicates that, at the very least, the token frequency histograms of the two documents are identical. This measure does not account for token ordering differences; we assume that statistical similarity is sufficient when considering documents from a single sender.

We use the distance defined in Equation 7.1 and a given distance threshold $\theta \in \text{range}(\delta)$ to partition each input sender-based group into smaller clusters based on email distances. Using these definitions, we present Algorithm 2, which receives an input cluster C_{in} (e.g. every email sent by a given user or bulk sender), and a threshold θ to produce a set of output clusters C_{out} ³.

Algorithm 2 Creating distance-based clusters.

PROCESSCLUSTERS(C_{in}, θ)

```

1   $C_{\text{out}} = \{\}$ 
2  for  $D \in C_{\text{in}}$ 
3      if  $C_{\text{out}} == \emptyset$ 
4          INITIALIZECLUSTER( $C_{\text{out}}, D$ )
5      else  $C_{\text{best}} = \text{argmin}_{C \in C_{\text{out}}} \delta(D, \text{REP}(C))$ 
6          if  $\text{SIM}(D, \text{REP}(C_{\text{best}})) > \theta$ 
7               $C_{\text{best}} = C_{\text{best}} \cup \{D\}$ 
8          else INITIALIZECLUSTER( $C_{\text{out}}, D$ )
9  return  $C_{\text{out}}$ 

```

The INITIALIZECLUSTER function adds a new cluster to the output set C_{out} , and places D in the new cluster as the “representative” of that cluster. The REP function returns the assigned representative email of a given cluster, which in this instance is the first email added to the cluster. While we could drop this assumption and attempt to find the best representative email for the entire cluster, performing this operation reduces to an instance of the set cover problem, which is known to be NP-complete. Although we may be able to find the optimal representative for some smaller-scale sets, such an approach would not scale to real-world collections. Therefore we use the pre-selected representative to keep the problem tractable.

³The large volume of emails in a real-world scenario makes the usage of complex clustering techniques not scalable. Although cluster creation is not the focus of this work, we observed that Algorithm 2 and k-means clustering did not show any significant differences in the quality of the resulting clusters, however, the method in Algorithm 2 showed a significant gain in speed.

Line 1 initializes an empty result, where the main iteration over the emails in the input cluster occurs on lines 2–8. We initialize the first output cluster with the first email in line 4. For each subsequent email, we calculate the distance of the email from all the representatives. If the best score is greater than θ , the email is added to the cluster that provided the best score; otherwise it creates a new cluster, with the email set as the representative (lines 5–8). After all emails are processed, the resulting set of clusters is returned in line 9. Each of the final clusters represents a “template” of an email - each email in the cluster is a minor variant of a mostly static template.

7.3.3 Baseline phrase extraction

We now describe the two methods that we will compare when conducting experiments. In both cases, we assume the input documents have already been clustered based on our selected distance metric. The task is to determine the parts of content that are frequent enough in the cluster to consider them “fixed”, which we can subsequently use to construct a template representing the clustered emails.

We adopt a greedy version of a longest common subsequence algorithm as a baseline approach. To identify fixed phrases for the templatization task, our baseline algorithm operates sequentially on emails of a given cluster. Pseudocode of the algorithm is presented in Algorithm 3. We define the input $0 \leq \gamma \leq 1.0$ as a threshold for determining if a token is fixed. Additionally, let t_i be the i^{th} token in an email vector i.e., $D = t_0 \dots t_{|D|-1}$. A token will be classified as fixed if it is present in $\gamma \cdot |C|$ documents. Given a cluster $C \in \mathbf{C}_{\text{out}}$ from the clustering step, and a value for γ , the algorithm will return a set of terms considered as fixed for the given cluster.

Algorithm 3 Baseline method for finding fixed text.

```

BASELINEEXTRACTPHRASES( $C, \gamma$ )
1   $df = \text{DOCUMENTFREQUENCIES}(C)$ 
2   $acc = \{\}$ 
3  for  $D \in C$ 
4       $phrase = []$ 
5      for  $i = 0$  to  $|D| - 1$ 
6          if  $df[t_i] \geq \gamma \cdot |C|$ 
7               $\text{APPEND}(phrase, t_i)$ 
8          else
9              if  $|phrase| > 0$ 
10                  $\text{INSERTINC}(acc, phrase, 1)$ 
11                  $phrase = []$ 
12             if  $|phrase| > 0$ 
13                  $\text{INSERTINC}(acc, phrase, 1)$ 
14  $\text{REMOVEINFREQUENT}(acc, \gamma \cdot |C|)$ 
15 return  $\text{KEYS}(acc)$ 

```

On line 1, we iterate over C to calculate the document frequencies of the unique tokens in C , and store the tabulated data in df . Line 2 initializes the acc variable, which acts as an accumulator for the frequency counts of observed phrases. Lines 3–11 loop over every email in C , filling acc with candidate phrases. The function `INSERTINC` inserts the entry if it does not exist, or increments the count of the existing entry. Lines 5–11 construct the candidate phrases by iterating over the token sequences from the current D . The $phrase$ (line 4) variable tracks the “current” candidate phrase. If a token has a high enough document frequency, it is appended to the current phrase (lines 6–7). Otherwise if the current phrase is non-empty, its count is incremented in acc , and the phrase list is reset to empty (lines 9–11). The final step on line 14 involves iterating over the entries in the acc variable, and removing any candidate phrases that are below the $\gamma \cdot |C|$ threshold.

One of the major advantages of this approach is that the space and time complexity constraints of the algorithm are both linear with respect to the input size. Both aspects of the algorithm are $\mathcal{O}(|C||D|_{\max})$, where $|D|_{\max}$ is the length of the longest email in C . Constructing df requires a scan over each email in C , while the execution of Algorithm 3 is a linear scan over only the emails in C . In terms of space complexity, the greedy construction of the candidate phrase table makes the space requirements be linear as well.

Limitations

While Algorithm 3 is straightforward in its execution, it suffers from the assumption that frequent tokens tend to co-occur with each other. Only the longest identified phrases will be computed for the given emails; any subphrases that may also pass the frequency threshold, but are not in the same order or consecutive, will not be included in the output set of fixed phrases unless they occur as unique phrases elsewhere in the emails. For example, if the current tracked phrase is “Hi your order is here”, the potentially higher-frequency phrase “your order is here” will not be added as a resulting fixed phrase simply because it occurs as a subphrase. Let us consider the following two clusters with 5 emails each, with 5 words in each document:

$C1$: (A B C D E),(A B C D F),(A B C D G),(A B Q C D)

$C2$: (A B C D E),(E A B C D),(A X C Y E),(A K C L E)

Let $\gamma \cdot |C| = 2$ for this example. For the first cluster $C1$, based on the first three emails, the resulting fixed phrase would be A B C D, since it is greedily constructed and passes the frequency checks. However, because A B C D was greedily constructed with the first three emails, the more useful fixed phrase pair A B and C D are not considered, resulting in a template with suboptimal coverage over its constituent emails. For the second cluster $C2$, the resulting fixed phrases would be A, C and E, while more optimal A B C D and E are not emitted.

7.3.4 Suffix array based approach

In order to overcome the limitations discussed in the previous section, we introduce an algorithm to perform template induction based on a suffix array. Since the performance and precision of the extracted phrases play an important role in template extraction and user profiling, we examine both the quality and scalability characteristics of this approach.

Our approach uses two main data structures to compute fixed phrases. The first is a suffix array (SA), which is the lexicographically sorted array of all suffixes of the input documents (only pointers to the original positions are stored). Suffix arrays typically operate over the character space of the input, however in order to ensure that we produce valid fixed phrases, we need the SA to operate on the token space. Therefore, when constructing the SA, we only permit suffixes to be added at standard token separators such as whitespace and punctuation. The second data structure is an array of the longest common prefix (LCP), which is produced while computing the SA. Entries in the LCP correspond to the number of common characters in the prefixes between two consecutive suffixes in the SA.

Algorithm 4 provides a sketch of the fixed phrase selection process using the SA and LCP. We define $\mu \in \mathbb{N}$ to be a threshold for the minimum number of shared characters allowed between two suffixes, and provide it as an input argument. In line 1, we initialize an empty accumulator *acc*, and we construct the SA and LCP structures. In this instance, *acc* has the added functionality of maintaining the insertion order of the entries, which will be needed in Algorithm 4. Practically speaking, this can be achieved by internally maintaining both a table and list to accommodate both access patterns.

The main loop of the algorithm (lines 3–15) iterates over the contents of the SA, using each iteration to examine a particular suffix and then decide whether to track it as a candidate fixed phrase. On line 4 we only admit suffixes that have a high enough overlap with the previous value; all admitted suffixes then have their counts incremented on line 5⁴. The PHRASE function extracts the actual phrase from the input by its position (SA) and length (LCP).

The next action taken depends on the delta of the “current” LCP; when there is a decrease in the LCP, we know that the currently tracked phrase has ended, and we need to check its frequency (lines 6–11). The completed phrase’s count is updated (lines 7–11). The PREVLARGER function emits all previously inserted *acc* entries as long as the corresponding LCP values are higher than the current value. Similarly, the PREVSMALLER function emits contiguous prior entries with lower corresponding LCP values. If the frequency of the recovered phrases is below the $\gamma \cdot |C|$, threshold, then the phrase is dropped from *acc*. Alternatively, if the LCP increases, this indicates that a new phrase has started, and we need to increment (or insert) it and all contained subphrases (lines 12–15). In the case where the LCP delta is zero, no additional action is taken. After the loop completes, the remaining keys in *acc* are returned as the fixed phrases (line 16).

⁴ Since all the documents are treated as a single string during the SA construction, we maintain an additional data structure *D* that contains indexes of beginnings of the documents. Thus, ids of the documents are obtained with SA.

Algorithm 4 Fixed phrase extraction based on SA, LCP.

```

EXTRACTPHRASES( $C, \gamma, \mu$ )
1   $acc = \{\}$ ;  $SA, LCP = \text{BUILDSUFFIXARRAY}(C)$ 
2   $\text{INSERTINC}(acc, \text{PHRASE}(SA[0], LCP[0]), 1)$ 
3  for  $i = 1$  to  $|SA| - 1$ 
4      if  $LCP[i] > \mu$ 
5           $\text{INSERTINC}(acc, \text{PHRASE}(SA[i], LCP[i - 1]), 1)$ 
6          if  $LCP[i] < LCP[i - 1]$ 
7               $c = acc[\text{PHRASE}(SA[i], LCP[i - 1])]$ 
8               $\text{INSERTINC}(acc, \text{PHRASE}(SA[i], LCP[i]), c)$ 
9              for  $phrase \in \text{PREVLARGER}(acc, LCP[i])$ 
10                 if  $acc[phrase] < \gamma \cdot |C|$ 
11                      $\text{REMOVEKEY}(acc, phrase)$ 
12             elseif  $LCP[i] > LCP[i - 1]$ 
13                  $\text{INSERTINC}(acc, \text{PHRASE}(SA[i], LCP[i]))$ 
14                 for  $phrase \in \text{PREVSMALLER}(acc, LCP[i])$ 
15                      $\text{INSERTINC}(acc, phrase, 1)$ 
16  return  $\text{KEYS}(acc)$ 

```

Let us again consider the simple example from the end of Section 7.3.3. Table 7.1 provides a view of the contents of the SA and LCP after constructing them over the example emails. Recall that the SA is an array with pointers to the suffixes' positions in the original input, (e.g., 1, 11, 21, and so on), and the LCP stores the number of shared characters between two consecutive suffixes. For the course of the example, we make the following assumptions: the cluster contains only the suffixes present in Table 7.1, each suffix comes from a different document⁵, and $\gamma \cdot |C| = 3$.

We start with the first suffix "A B C D", which belongs to the first document. The suffix has an overlap of 9 characters with the next suffix (i.e., there are at least two documents that have a repeated phrase of length 9 including spaces) and thus we insert the phrase into *acc* with count 1 (line 5). As we progress in the example, we see the decrease of the LCP between IDs 3 and 4 which is valid, i.e., number of remaining overlapping characters passes μ . A negative LCP delta indicates that the suffix has shortened, and the current phrase has ended. In this case, its accumulated frequency is checked, the phrase passes, and it is emitted. We observe that the LCP value drops to 5 characters, thus the count of the corresponding suffix should include the frequency of the previous larger phrases, and *acc* is updated accordingly (line 7-8). As a result, "A B" is added to *acc* with count 3. When the LCP increases, the frequency of the current tracked phrase should continue to accumulate and longer phrases should be added to *acc* with count 1 (lines 12–15). Following this example to its termination, we would produce A B C D, A B, C D for the first cluster and A B C D, E for the second one, which prove to be more optimal selections than those produced by Algorithm 3.

⁵In the algorithm 4 we proceed ensure that phrase increment only happen for phrases in different document.

| ID | LCP | SA | “A B” : 4 |
|-----|-----|-----|----------------|
| 1 | 9 | 1 | A B : C D E... |
| 2 | 9 | 11 | A B : C D F... |
| 3 | 5 | 21 | A B : C D G... |
| 4 | 1 | 31 | A B : Q C D... |
| 5 | 7 | 3 | B C D E... |
| 6 | 7 | 13 | B C D F... |
| 7 | 3 | 23 | B C D G... |
| 8 | 1 | 33 | B Q C D |
| 9 | ... | 47 | C D |
| ... | ... | ... | ... |

Table 7.1 – LCP, SA and actual suffixes ordered lexicographically.

“A B C D” : 3

“B C D” : 3

The computation complexity consists of two factors: suffix array construction and fixed phrase extraction loop. The former is $\mathcal{O}(|C||D|_{\max})$ complexity. The latter is proportional to the input size, or more precisely, proportional to the number of words in the input. Overall, the complexity is $\mathcal{O}(|C||D|_{\max})$.

The algorithm also has modest space demands, since it only requires the following input: the suffix array and the longest common prefix array, both of which are proportional to the input size and store only pointers to the original input. During operation, the only data structure maintained is the output argument, the final set of fixed phrases.

The algorithm affords several opportunities for parallelization. For example, each part of the suffix array starts with a different letter, and can be processed independently. Additionally, instead of partially rescanning the suffix array when there is a change in the current LCP value, a new process/thread can be spawned that will be responsible for updating the occurrences of the bigger phrase.

7.3.5 Constructing a template from phrases

We now describe the process of building a template based on the induced fixed phrases. We design the template itself to be an ordered list of fixed and non-fixed phrases. We define the *coverage* of a given template/email pair to be the fraction of characters of the email that can be aligned with the fixed text present in the template. It is an important measure since it describes an email compression, i.e., number of characters that might be saved while typing or while representing an email through the fixed and non-fixed regions etc.

The main task is to align the fixed phrases with the following optimization criterion: choose a set of non-overlapping phrases that maximize the coverage over the email. Fundamentally the process of template building involves dynamic programming to align the fixed phrases of the template onto the email. In particular, we formulate the solution recursively as follows:

$$C[i] = \max\left(C[i + 1], \max_{\forall ph_{i+1}} (\text{len}(ph_{i+1}) + C[i + \text{len}(ph_{i+1}) + 1])\right),$$

where $C[i]$ is the optimal solution (coverage) at the i th position in the representative email, ph_{i+1} correspond to the matching phrase ph to a position $i + 1$. In order to allow fast look ups of the matched phrases for each position in the email, we first build a prefix tree (trie) out of the whole set of obtained fixed phrases and further use it as an index to find all the phrases that match given position.

When the alignments of the phrases to the email are found, the email is transformed into the sequence of the matched fixed phrases separated by the parts of the template which were not mapped to a fixed phrase, and are presumed to contain variable content. For example, if we have 2 fixed phrases: “hi”, “how are you” and the email is “hi John, how are you”, we would like to generate the template “hi _____ how are you”. Example templates generated using our techniques on the Enron corpus are presented in Table 7.2.

| |
|---|
| <i>Template 1:</i> The report name : _____ p / l _____ , published as of _____ / 2001 is now available for view on the website . |
| <i>Template 2:</i> We have received the executed Letter Agreement date _____ / 2 _____ 00 amending the Copy will be distributed . |
| <i>Template 3:</i> Please , put it on my _____ . Vince |

Table 7.2 – Examples of the templates created based on SA. _____ correspond to the non-fixed regions of the templates.

Overall, the quality of the templates is hard to evaluate and can be either performed by humans or automatically. Human evaluation is rather expensive and error prone. Therefore, we rely on the coverage metric as a reflectino of the cluster edit distance.

7.4 Experiments

In this section we describe the experimental setup, data sources and insights we have obtained from both the baseline and suffix array based approaches. We perform two main experiments to analyze the quality of the templates produced by the two presented methods. We conduct the same experiment with two separate corpora: a synthetically generated corpus and a real email corpus from the Enron Corporation. We continue to use the *coverage* over emails for quality assessment, and we introduce the *template entropy* measure, which is the proportion of fixed phrases that a template contains. More fixed phrases in the template indicate more variable content separating it, thus introducing more uncertainty which we aim to minimize. Using these measures and given an induced template, our goal is to maximize coverage and minimize template entropy. To test the robustness of the proposed approach we have performed multiple runs varying clustering

threshold θ and fixed phrase document frequency γ .

7.4.1 Synthetic corpus

We do not claim to have any direct control over the input clustering process, but it is important to know how sensitive the template creation algorithms are to cluster quality. We would like to test both methods against clusters of varying quality, so in order to perform these tests while minimizing confounding factors from the input, we use a synthetically generated corpus of emails, which is constructed as follows:

We automatically generated 3 independent sets of 100 emails each. For each set we create 10 clusters with 10 emails each. For a given cluster, we selected a set of predefined “fixed” phrases that we then separated in each email by randomly generated text that acted as the “non-fixed” portion of the email. For example, in a cluster with two predefined fixed phrases, we randomly chose phrases to put before, between and after the fixed phrases for each email in the cluster.

To ensure that uniform email size did not confound our results, we created another synthetic set where the cluster sizes were distributed normally instead of uniformly. Table 7.3 provides details on the generated corpora. The average coverage per template for both distributions of cluster sizes is about 80%, and the average template entropy is 3. The randomly generated emails had an average size of 260B, which is in the 92nd percentile of email sizes in the public corpus.

| Metric | Syn1 | | Syn2 | | Syn3 | | Mean | |
|--------------|------|-----|------|-----|------|-----|------|-----|
| | U | N | U | N | U | N | U | N |
| Cluster size | | | | | | | | |
| Coverage (%) | 75 | 85 | 82 | 79 | 79 | 82 | 79 | 82 |
| Entropy | 3 | 2.7 | 4 | 4 | 2.7 | 3.2 | 3.2 | 3.3 |
| Characters | 178 | 246 | 148 | 162 | 249 | 177 | 192 | 195 |

Table 7.3 – Average characteristics of the generated synthetic corpus. U and N are uniform and normal of the cluster sizes.

We characterize the templating performance by varying two parameters that are specified prior to processing as described in Section 7.3: the clustering threshold θ and the fixed phrase document frequency γ . We vary cluster threshold in order to test the robustness of the methods against the quality of the produced clusters. Similarly, we check the quality of the created templates with respect to the coverage of the phrase by varying fixed phrase document frequency. We found that with $\theta \leq 0.6$ all synthetic emails were grouped into one cluster, while for $\theta = 0.7, 0.8, 0.9$ average cluster sizes were 20, 11, 6 respectively. Since generation of the fixed and variable parts in the synthetic emails is performed randomly, multiple clusters may share similar fixed and variable parts, which might result in over- or under-clustering. $\theta = 0.8$ showed the closest cluster distribution to the expected results. The results are shown on Figure 7.2 using our two selected measures. The higher the clustering threshold is set, the more template coverage converges towards a steady value for both methods. Similar behaviour is shown for the template entropy; as cluster quality increases, templates are built over more homogeneous sets of emails and therefore

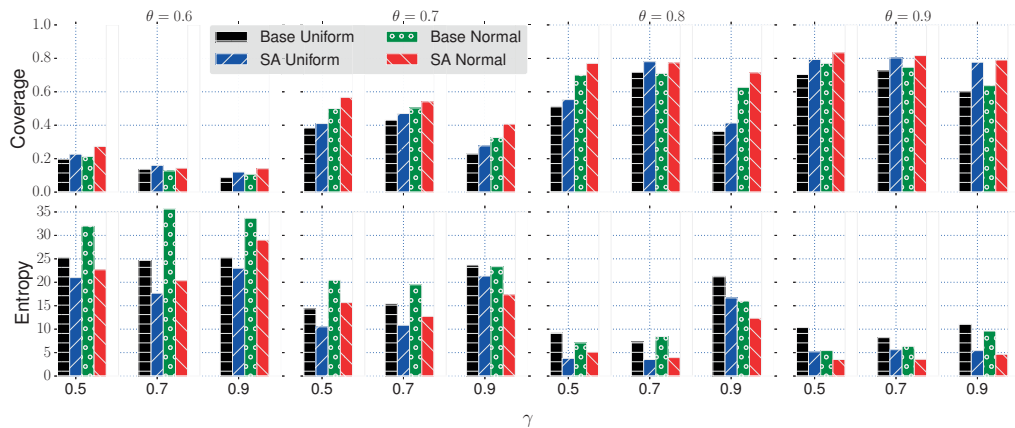


Figure 7.2 – Average template coverage and entropy for two synthetically generated corpora, i.e., equally and normally distributed cluster sizes. We present the coverage by varying clustering threshold θ and fixed phrase frequency γ . Suffix Array based approaches consistently show higher coverage (portion of the email marked as fixed) with lower Template Entropy (number of fixed-phrases per template).

have a more compact structure.

We can see the consistent improvement over the baseline approach both in terms of email coverage and template entropy. As shown on the synthetic corpus results Table 7.3, the expected coverage and entropy are 79% and 3.2 respectively for the first batch of generated emails. We can see that by varying both θ and γ , the suffix array based approach maintains dominance across all metrics. This consistently improved performance stems from the fact that the suffix array provides a higher quality set of candidate fixed phrases to use in template construction. Overall, having more phrases is beneficial when it comes to user profiling and better template construction even if the higher number of phrases requires an increase in extraction time.

7.4.2 Enron corpus

For the second series of experiments we used the real world publicly released email corpus from the Enron Corporation. The corpus contains a large database of over 600,000 emails generated by over 150 users, mostly from senior management of Enron. For the experiment we extracted all of the sent emails of the corpus users. This resulted in over 125,000 emails with more than 300 distinct email addresses⁶.

We preprocessed the corpus as described in Section 7.3. For the sake of evaluation, multiple statistics are collected during clustering, such as user outbox (sent mail) counts, cluster quality, and so on. We do this in order to provide a sense as to whether the clusters formed would be suitable for template induction. By exposing an explicit notion of the quality of the clusters, we avoid the possibility of an unknown feature acting as a confounding factor when evaluating the induced templates.

⁶We treat multiple corporate accounts of a user as separate entities since each account is used for a distinct purpose.

To measure quality of the clusters themselves, we use a variant of the “edit distance” (ED) measure. For an output cluster $C \in \mathbf{C}_{\text{out}}$, we calculate the ED of the cluster as:

$$\text{ED}(C) = \frac{1}{\text{chars}(C)} \sum_{D \in C} \text{CharED}(\text{REP}(C), D) \quad (7.2)$$

where $\text{chars}(C)$ is the sum of the character lengths of all member emails in C . We rely on this metric to emphasize the compression of the outbox and thus typing reduction of the repetitive parts. When a cluster has a body edit distance of less than 20% of its average content length and contains at least 5 emails, we call that cluster a “high-quality” (HQ) cluster. Based on this analysis and definition we obtain the following corpus statistics shown in Table 7.4.

| User deciles | Number of users | Total outbox size | Average outbox size | Average email length | Average cluster size for $\theta = 0.8$ | High quality clusters for $\theta = 0.8$ |
|-------------------|-----------------|-------------------|---------------------|----------------------|---|--|
| 0 | 79 | 79 | 1 | 284.81 | 1 | 0 |
| 20 | 19 | 50 | 2.63 | 372.21 | 1.97 | 0 |
| 30 | 28 | 310 | 11.07 | 1,352.31 | 3.22 | 0 |
| 40 | 31 | 1,372 | 44.26 | 589.07 | 5.15 | 2 |
| 50 | 32 | 2,885 | 90.16 | 641.81 | 5.63 | 2 |
| 60 | 32 | 5,886 | 183.94 | 483.96 | 4.8 | 7 |
| 70 | 31 | 10,809 | 348.68 | 618.29 | 6.16 | 11 |
| 80 | 33 | 23,377 | 708.39 | 409.21 | 4.56 | 53 |
| 90 | 31 | 80,336 | 2,591.48 | 570.87 | 7.84 | 102 |
| Aggregates | 316 | 125104 | 442 | 591 | 4.48 | 177 |

Table 7.4 – Description of the Enron sent mail corpus.

We break down the user sent mail volume into deciles to get a sense of how many users actually could be characterized as “high-volume” senders, as shown in Table 7.4. As the table shows, only active users that fall into the 40th percentile and higher by their sent mail volume are capable of producing templates that are of suitable quality, and as the decile increases, the number of HQ clusters grow superlinearly. The data in the table suggests that a user would need at least ~40 emails to produce any useful templates. Given the increasing prevalence of email for communication, this suggests that templating could be useful for virtually every current user of email.

The first two deciles did not have more than one email sent which is shown in the 4th column. Interestingly, the average size of a sent email is significantly greater in the third decile compared to others. This observation is driven by the data, i.e., users with a relatively low outbox size (~11 emails on average) tend to send mainly long annual reports. Moreover, the average size of the user outbox correlates with the tendency to write similar emails.

We also investigated whether the average edit distance in a cluster behaved as a function of the cluster size, as shown in Figure 7.3. Indeed, the greater the size of the produced cluster, the lower the chance to observe a small edit distance within the cluster. The two exceptions to this observation occur at cluster sizes 10 and 100, which consist of annual reports (within a cluster) that only vary in the month mentioned in the email.

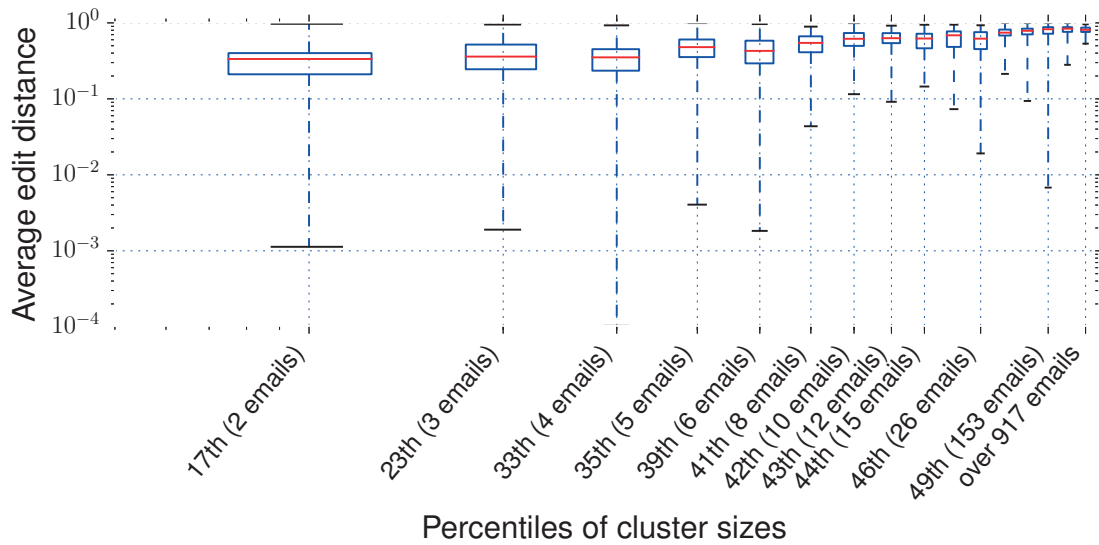


Figure 7.3 – The average ED as cluster size increases for $\theta = 0.8$

Using the edit distance as defined, we find that the edit distance for the bodies and subjects of the emails in a given cluster have a Pearson correlation coefficient of 0.31. As the edit distance increases for the subjects, the body edit distance tends to increase as well. The trends for subject vs. body are presented in Figure 7.4. Such information could be used in email clustering and cluster filtering, or for efficiency gains when performing clustering at large scale.

Similarly to the synthetic corpus experiments we performed template induction for the baseline and suffix array based approaches. As a result of clustering with $\theta = 0.8$ we obtained over 25,000 clusters. As expected, the higher the cluster threshold, the more clusters are produced. In particular, we observed strong positive correlation of 0.91 between θ and the number of clusters produced, while the synthetic data had a moderate positive correlation of 0.62. More prolific users (top deciles) tend to write more similar emails as described in Table 7.4. We tested the variations of the suffix array approach when limiting the size of the accepted fixed phrase (2, 3 words).

As can be seen from Figure 7.5, the suffix array based approach performs better than the baseline alternative in terms of coverage and entropy. We show only one pair of values for θ and γ here, but we observed similar behaviour for other values of these parameters. By varying the constraints on the phrase quality, we show that it is possible to balance between the coverage and the entropy of the template. For example, the most restrictive results (SA with phrases with a min of 3 words) have the least template coverage. However, this significantly reduces the number of fixed regions in the template. Overall, average entropy for various setups of our approach is ~ 4 , while maintaining the coverage within 60%. Considering an average email length of 600 characters and average word size to be ~ 10 , we obtain a ~ 35 word reduction in typing in an autocompletion setting for users with induced templates.

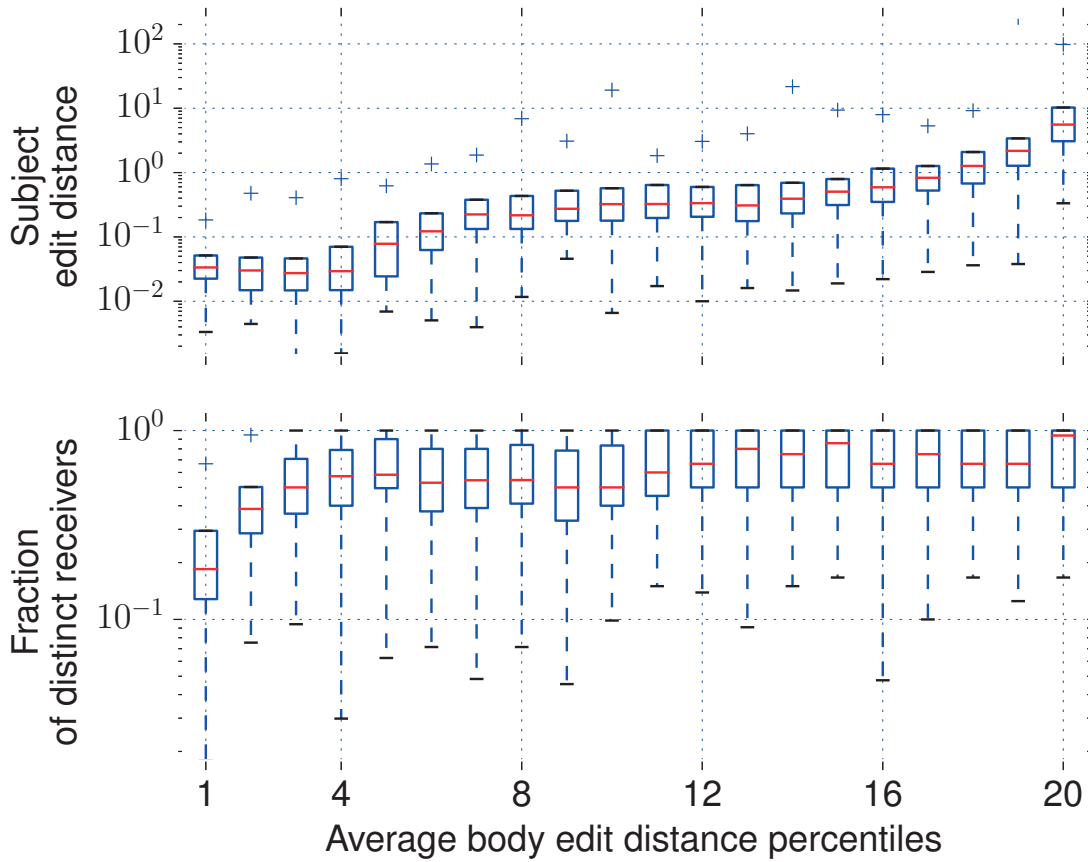


Figure 7.4 – Correlation between average body edit distance and average subject edit distance within cluster and fraction of distinct receivers (1 - all receivers are distinct) respectively.

Our findings indicate that using SA for template induction offers better performance than the baseline both in terms of email coverage and template entropy. As can be seen in Figure 7.2, where expected performance is $\sim 80\%$ and ~ 3 for coverage and entropy respectively, the suffix array based fixed phrase extraction shows better results for both metrics no matter the quality of the cluster. Similar behaviour is observed for the Enron corpus and shown in Figure 7.5.

Practical observations and recommendations for the choice of parameters γ and θ . *First*, experiments on both corpora, i.e., synthetic and ENRON, showed that clustering coefficient $\theta = 0.8$ balances well between the number of extracted clusters and inter-cluster edit distance of the emails. Additionally, we recommend to run templatization on the clusters with at least 4 emails. *Second*, by reducing γ , the entropy of the templatization reduces, i.e., the number of used fixed phrases to construct a template are relatively small. Thus, we consider that $\gamma > 0.85$ is too restrictive and we recommend to set the desired frequency of the fixed phrase (γ) between 0.6 - 0.8.

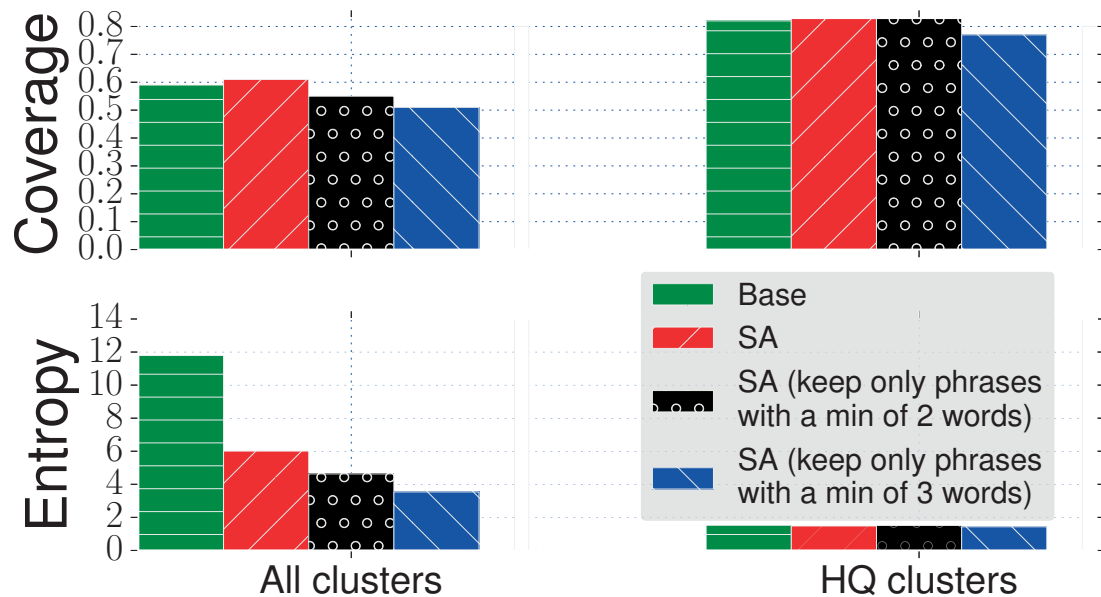


Figure 7.5 – Average template coverage and entropy for $\gamma = 0.8$ and $\theta = 0.8$ over Enron corpus all and high quality clusters.

7.4.3 Scalability analysis

To test the performance characteristics of both approaches we performed fixed phrase extraction over various cluster sizes. Even though this work focuses on the template extraction for a single user account the size of the outbox should not be neglected, since it could reach millions of emails. We kept each email size to be almost identical in size (approx. 1 KB each⁷). We varied the cluster size from 2 KB to 33 MB with the corresponding number of emails from 2 to 33,000. Figure 7.6 shows the execution time taken to create templates as the size of the input cluster increases. When the clusters are relatively small, the methods are equivalent in efficiency. However, the growth trends depicted in Figure 7.6 clearly show that the baseline approach takes longer to complete than the suffix array approach as the cluster size scales up. The baseline approach proves to be sensitive to both the number of emails in the cluster and email size variations within the cluster. The suffix array is agnostic to these variations due to the fact that the input is treated as a blob of information for which the suffix array is built and valid phrases are added to the result.

Additionally, one can easily see that the growth of the baseline is also superlinear - the baseline requires less than 200 seconds for 5K emails and 400 seconds for 15K emails, but requires over 1600 seconds for 30K emails. While the earlier segment has a slope of approximately 2/3 (2x time for 3x input), the next segment is closer to a slope of 2 (4x time for 2x input). This suggests that the slope will continue to grow as the input size increases. The suffix array approach shows a slight slope increase as well, but it is multiplicatively less than the baseline, making the SA approach more scalable.

⁷1 KB is an upper bound on the 95% of the emails send in our corpus.

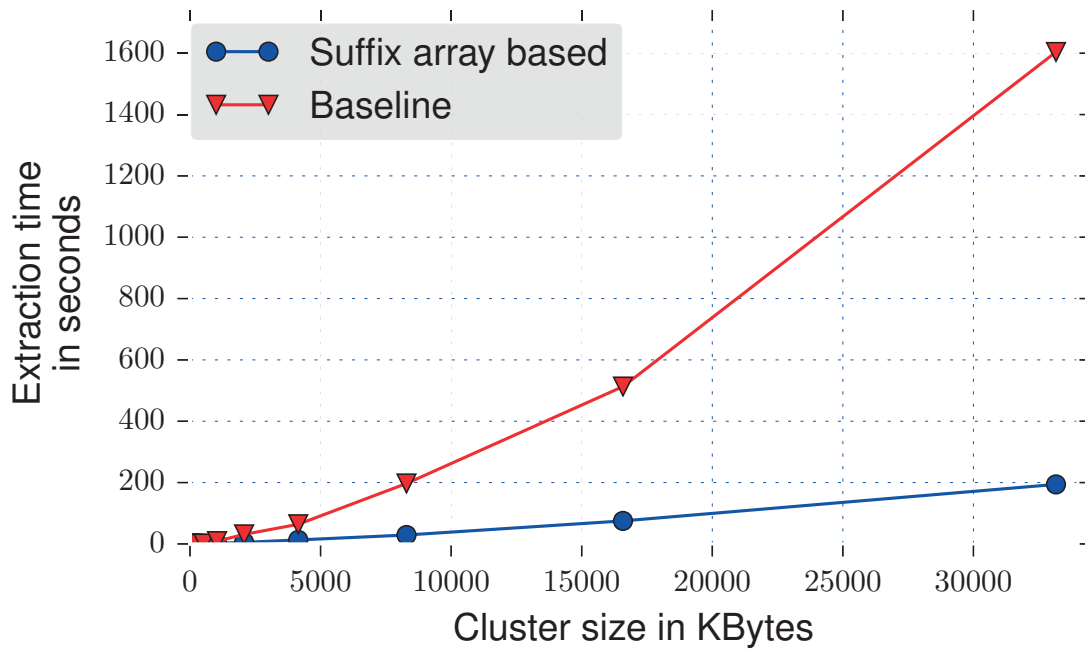


Figure 7.6 – A comparison of the increase in extraction size as the average cluster size increases.

7.5 Conclusions and Future Work

We have demonstrated the feasibility of performing high-quality plain text template induction, and have done so using a highly scalable solution. The experiments, performed on both a synthetic and organic email corpora, illustrate the efficacy of using suffix arrays to induce templates, even in the face of input clusters of varying quality.

We continued our investigation by then comparing two template induction algorithms. We have shown both in theory and in practice that using a suffix array is more effective than an out-of-the-box shingling baseline for template induction. The results of our investigation show that plain text documents can be templated more efficiently using suffix arrays: the baseline showed superlinear growth, while the suffix array’s growth is multiplicatively slower. Additionally, the templates induced using the suffix array encode more useful information than the greedy approach: across our experiments the suffix array templates provided consistently better coverage than the greedily-built templates. Our ancillary experiments suggested that if used for email autocompletion, the generated templates could on average save 35 words of typing when composing emails. Overall the presented work is salient for numerous applications, including optimizing the production of any textual content, extracting information from machine-generated content (imagine if the “user” sending the email is an algorithm written by an activists to mobilize people), and profiling composition behaviours of users.

In this work we only considered forming templates for individual users or bulk senders, however we would like to further explore across user clustering to induce templates as well as applica-

tions of the created templates for online activists. This could allow an even higher document compression and carry insights into the composition behaviours of users at an aggregate level.

Our initial implementation relied on a simple function to select the representative document of a cluster, but we would like to further explore alternative methods to find a “near-best” representation of a given cluster. Although we have shown that suffix array templates are more robust to cluster quality than greedily built templates, the clustering method affects the number of clusters produced, potentially creating many more clusters than there should be.

Finally, although we discussed and analyzed the efficiency of using the suffix array for template induction, our collection sizes are not real-world scale, and did not fully push the limits of our implementation. To be truly web-scale, our approach would need to work on billions of input documents, which will most likely present several efficiency challenges worth investigating. Both the clustering and induction phases could be improved by developing parallelized versions of both algorithms.

There are many more applications and use cases where our method would fit in the context of online activism. Below we name a few. *First*, descriptions of the online petitions could be generated or recommended to the activists automatically based on the earlier campaigns. *Second*, micro posts that mention online campaigns could be better auto-completed once a user specifies the campaign’s hashtag. *Third*, mailing lists of the activists can be assisted with the automatically generated templates of the replies or even questions. Finally, template induction could help in composing the reports of the activities and plans of the activists.

8 Conclusions

“Networked protests of the twenty-first century differ in important ways from movements of the past and often operate with a different logic.”

– Prof. Zeynep Tufekci, 2017, quoted from “Twitter and Tear Gas” p.XXIII [291].

8.1 Conclusions and Discussion

The Internet connectivity and Social media promise activists to provide a great medium to spread the idea to a wide number and range of people. Alas, there are, first, numerous gaps in the analysis of the online and offline actions performed by the activists on the Web, and, second, multiple flows in the current solutions, tools, and methodologies that prevent activists from efficient use of the existing platforms. In this thesis, we presented solutions to the issues related to the analysis, modelling, and facilitation of collective actions. We demonstrated the need for more efforts in optimizing and aiding the ways current activism is performed online and provided measurable, deterministic and interpretable models to shed more light on the mechanisms standing behind the collective actions.

Public Campaigns on Social Media and Beyond. First, we conducted an in-depth empirical and statistical analysis on over 100 online public campaigns to study their user engagement patterns. We proposed a novel categorization of the public campaigns by their goals and user engagements. We empirically showed that these types of campaigns employ different agendas, message patterns and user communities to spread their message. In particular, we observed that first degree neighbours are essential to spreading the messages, while the more diverse the central core of contributors is the more likely a campaign will gain a larger audience. Moreover, we saw how physical actions, official meetings, and calls for actions shape campaigns that focus on mobilising people and how online factual postings engage users into the awareness campaigns. Thus, our results make a case for more extensive, fine-grained and large-scale examining of the online campaigns.

Online Petitions and Influence of the External Sources. Second, we performed a cross-platform analysis of a collection of over 4,000 online petitions with their front page rankings and the corresponding social media posts. Our results suggest and measure the importance of the social media and external promotion on the popularity of the petition. We also proposed a novel model that combined multiple external influences and intrinsic nature of the human behaviour. The proposed model was shown to outperform multiple baselines concerning both short and long term prediction. The dynamics of the user engagement varies across various topics, however, explicit inclusion and modelling of the external factors enhances the prediction.

Efficient Document Filtering. Further, we showed that using a general, topic-specific pattern to filter micro posts during events leads to collections with the highest precision and recall while having a minimal training set. We showed how mined patterns that describe a particular topic are maintained and build using semantically homogeneous clusters of the input training example. We argued that our approach is well generalizable to different topics and can handle a large stream of messages and accurately extract occurrences of mined patterns. As a result, our approach had high precision and recall due to the several factors. First, semantic representation of the items—synsets—guarantees more “loose” topic representation. Second, the semantic similarity metric is used to tune the precision of the method. Third, increase or decrease of the support threshold results in decrease or increase of the recall respectively (vice versa for the precision). Despite being run on the historical data, our method could be applied to the unseen short text extraction. If synsets are general enough and capture closeness of the situational information, such as locations, perpetrator names or organisations, etc., our method would generalise even better to unseen events and messages.

Templatization of the Unstructured Repetitive Content. Finally, we devised an algorithm to extract in an efficient and deterministic manner a set of repetitive, most representative phrases that a user utilise when constructing a message. We showed how a document template could be created from a set of one’s fixed phrases and how the quality of extracted phrases influences the quality of the constructed template. Moreover, we performed an extensive evaluation of the proposed approach and showed that our algorithm outperforms the baseline. We saw that about 1% of the whole email user base produces a lot of repetitive content. Thus, efficient document templatization could result in significant storage reduction, i.e., a single template has to be stored once for a group of similar messages and only variable parts have to be kept on a document level. Concerning the fixed phrase extraction, despite personalising the autocompletion, it also can be used for the information filtering, as some of the spam, irrelevant or abusive messages that contain particular phrases can be blocked or removed automatically.

Bigger picture. In the thesis, online campaigns and e-petitions are studied separately in Chapter 3 and Chapter 5. However, a diverse set of online campaigns’ actions might be relevant to the case of e-petitions. In particular, promotion of the petitions can be made not only through the social media (in the form of tweets that calls for an action - signature), but also through the advertisements on the official meetings and conferences, or through the offline gatherings, protests, workshops, competitions. On the one side, while building a successful campaign,

significant volumes of informational and promotional documents are created. Therefore, our framework for the template induction and frequent phrase extraction described in Chapter 6 could automate and speed up the process of the message composition. On the other side, to ensure a better quality of the extracted templates, documents could be pre-filtered using the methodology explained in Chapter 6. Both of the facilitation approaches, when combined, could result in more accurate filtering and creation of either relevant (such as particular topics of interest, structured descriptions of the event or information) or irrelevant (such as spam, scum, semantically different texts) information.

On the way to campaign's success. Below we present a list of recommendations and observations for the activists on what could be considered as the best practices to increase campaigns' chances of the success. *First*, we saw that most of the successful petitions experience increased user participation during the first initial days, thus, it is important to prepare the material and the meticulous plan on how to engage more people and explore the possibilities to using multiple channels to convey the idea of the petitions as early as possible. *Second*, we recommend the activists to establish the connection with the petition platforms' owners and request them to feature a petition or a campaign on the front page. Front page showed to have the strongest effect on the user gain compared to social media. *Third*, we saw some evidence that users are less likely to post about sensitive matters (LGBTQA, women rights, abuse) on social media. Therefore, other means to promote and spread the information shall be found for such topics. *Fourth*, depending on the expected user engagement or defined goal of the campaign, different actions could help to (1) better shape the engagement or (2) convince users about your ideas or problems. In particular, we saw the correspondence between (a) mobilisation campaigns and official/governmental meetings and calls for actions; (b) awareness campaigns and physical actions and scientific or news publications; (c) ever-growing campaigns and links to the news media and physical actions. *Fifth*, duplicate or near-duplicate content does not collect many retweets, thus, diversifying the ways to present campaign's ideas and engage more diverse and active users leads to higher users participation.

This interdisciplinary work provides a consistent study of the online activism in the form of online campaigns and petitions, as well as, proposes several tools to empower and facilitate information spreading. *Overall*, it is clear that at the current stage it is nearly impossible to create an agenda of a movement that would ensure its success. However, with further advancements that are described below, consistent information sharing and activist's "good" intentions, we could reach the point where collective actions for the social good are very likely to reach their goals.

8.2 Future Work

The research on the collective actions in social media has been mainly focused on analyzing single instances of the activism. At the same time, non-for-profit computing is somehow neglected areas of applied computing [291]. In Chapter 2, we conducted a comprehensive survey that focuses on a variety of aspects from types and directions of digital activism to modelling and

predicting a user participation and popularity of the online items, as well as, methods and tools to facilitate and improve the outcome of the collective actions. The review of the existing solutions and techniques sheds light on the list of still unresolved or scarcely addressed issues. Below we highlight a few future directions that will be important to address in respect to (1) the semantic analysis of the information produced by the activists, including genuine and false flag statements, (2) tools and methodologies that are able to facilitate planning, maintaining and communication processes during the collective actions.

Analytics of the Collective Actions. To a large extent, the needs of the activists are rather practical and often short-term. In the context of good intentions, genuine, concise, fast, censorship-free and targeted spread of the information is of uttermost importance.

Organizational stability. Contrary to conventional movements and gatherings, online activism often lacks a well-established leader that advocates and promotes an idea or mobilizes people. On social media, anyone can express their opinion and, thus, cultivate an environment with multiple authorities that might have different stances on the issue. While this might not necessarily invalidate the efficiency and popularity of an action, it could have long-term consequences regarding follow-up activity [291].

Misinformation and “False Flags”. Another issue when it comes to the multi-leader environment is misinformation. It can be spread by any sources, and sometimes even credible accounts can be impersonated, thereby threatening the course of the action. Even though there exists a growing corpus of research in this field. e.g., [46, 73, 116], there is still a lot to be done. In particular, the following questions remain unclear: *What is the role of social media in shaping user’s opinion or manipulating users’ interests? What is the long-term implication of any bias as well as its elimination? Which methodologies and algorithms can be generalized to the domain of online activism?* . There is rapidly growing need to automate and facilitate identification and verification of the false information. By utilizing combined, global, common sense, and possible domain specific models, it could be possible to overcome and prevent misunderstanding and spread of the misleading information.

Fact Aggregation. In a similar context to the previous point, there is a need to further aggregate and compress facts about advocacy. First, redundant, “not-important” and near-duplicate information must be accurately replaced by a single instance of the occurrences; second, truthful or factual information has to be identified, and finally, a comprehensive summary should be constructed. Those are the steps that still require attention from the research community regarding accuracy and reproducibility. This research line depends on many factors, including the extent to which semantic representation and its derivatives can be used to solve consistent fact aggregation; accepted trade-offs between precision and recall; summarization from the multilingual sources, etc.

Censorship and Privacy. One of the major concerns on the web is privacy. Because a majority of the activism misaligns with the established political or social beliefs, the identity of some

activists can be potentially dangerous to reveal. For instance, Facebook has a policy of real name, thus, making it challenging and even menacing to gather people around sensitive issues. In the same context, even though social media soften the censorship on the information, still some platforms might not support “radicalized” or “controversial” discussion [291]. We need to create a widely accepted anonymized platform for online activism. Leveraging an extensive research on security and privacy, we could alleviate the censorship issues, and thus, enable activists to stand for their rights.

Consistent Methodologies and Guidelines. Before offering managerial tools and platforms for the activists on how to develop a campaign, it is still not well studied what are the main stages (in time and location) and actions that are to be used and are effective for a collective action. What are the major features that distinguish a successful advocacy with a failed one? How the user involvement in a campaign changes over time? How to improve and facilitate public interest on a particular topic? Is there a way to generalize the methodologies between the first of activism? Do practices that are used for the popular topics apply to the ones that are more sensitive, private and, therefore, unpopular? How to prevent malicious and false ideas to spread? These are the questions that remain unanswered and require effort in each direction.

Tools and Methods. In spite of existing limitations, we believe that if handled with meticulous care, collective actions can be partially automated and empowered by the developed tools and methodologies. In particular, three main directions require prompt attention: *first*, tools to “consume” the information, e.g., announcement/document aggregation, stance filtering, summarization, etc., *second*, tools and automatic assistants that effectively produce and distribute textual content, announcement, e.g., auto-completion tools, automatic identification of the channels and the schedules to spread the information, and *finally*, planning and agenda recommendation for creating a new campaign.

User-friendly and Generic Summarization. Despite the heterogeneity of the research on content summarization, clear, user-friendly and quality information aggregation system is lacking. The main challenges that prevent this tool to exist are diversity and ambiguity of the textual representation, multi-language sources, credibility of the information or a source, bots, impersonation, high volume of the information and, thus, lack of online scalability.

Sources of the Information Propagation. Although social media has led to a paradigm shift for awareness advocacy as it increases the speed, the effectiveness and the outreach of public campaigns, many activists still fail to reach beyond the communities for which they advocate [59]. Therefore, exploring the ways to surpass an existing “echo chambers” is important to learn how to alter the message appeal to a broad mass of people. Moreover, despite the recent advancement in influence maximization, this line of research is yet far from being integrated into the tools used by the web users. In particular, studies about activism on less popular and socially accepted issues, such as human rights, women rights, LGBTQ issues, etc., are rather scarce. As we explained in Chapter 5 not all topics are equally covered by social media.

Campaign Scheduling. In Chapter 2, we have seen that not many researchers study the actual set of actions that contribute to a online activism. In Chapter 3, we have established the large-scale study of coarse-grained actions that are performed by the environmental activists. However, what are the most effective actions for other types of issues remains unclear; what are the fine-grained activities and what is their efficiency are still unresolved questions. Moreover, the process of the starting, maintaining and reflecting on a campaign is not sufficiently explored to automate it in any way.

Risks. Despite considering only “good” intentions of the activists, we also acknowledge that this work enhanced the knowledge of the adversary. In particular, the analysis made in the previous chapters can be used for many malicious purposes. For example, (1) spamming of the users; (2) modifying and enhancing spam messages to make them less detectable; (3) discovering users that have particular preferences or disagreements and connecting them; (4) artificially raising the hype around some socially unacceptable campaigns or petitions, and many others. Moreover, it must also be noted that a better understanding of social media can also be used by malevolent users, e.g., to “hijack” well-intended initiatives, to spread misinformation about particular topics, etc. One recent example of such behaviour is the proliferation of fake news around US presidential campaign, where a thorough understanding of the online communities was used to spread “alternative” facts anonymously.

Bibliography

- [1] S. Abbar, Y. Mejova, and I. Weber. You tweet what you eat: Studying food consumption through twitter. *arXiv preprint arXiv:1412.4361*, 2014.
- [2] H. Abdelhaq, C. Sengstock, and M. Gertz. Eventweet: Online localized event detection from twitter. *Proc. VLDB Endow.*, 6(12):1326–1329, Aug. 2013.
- [3] D. Aberdeen, O. Pacovsky, and A. Slater. The learning behind gmail priority inbox. In *NIPS Workshop on Learning on Cores, Clusters and Clouds*, 2010.
- [4] L. A. Adamic and N. Glance. The Political Blogosphere and the 2004 U.S. Election. *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, pages 36–43, 2005.
- [5] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.
- [6] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13*, pages 607–616, New York, NY, USA, 2013. ACM.
- [7] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Göker, I. Kompatsiaris, and A. Jaimes. Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282, Oct 2013.
- [8] N. Ailon, Z. S. Karnin, E. Liberty, and Y. Maarek. Threading machine generated email. In *Proc. of the 6th ACM International Conference on Web Search and Data Mining*, pages 405–414, 2013.
- [9] M. Akbari, X. Hu, N. Liqiang, and T.-S. Chua. From tweets to wellness: Wellness event detection from twitter streams. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16*, pages 87–93. AAAI Press, 2016.
- [10] A. I. Alberici and P. Milesi. Online discussion, politicized identity, and collective action. *Group Processes & Intergroup Relations*, page 1368430215581430, 2015.

Bibliography

- [11] R. Alghamdi and K. Alfalqi. A survey of topic modeling in text mining. *International Journal of Advanced Computer Science and Applications*, 6(1), 2015.
- [12] R. B. Alley, J. Marotzke, W. D. Nordhaus, J. T. Overpeck, D. M. Peteet, R. A. Pielke, R. T. Pierrehumbert, P. B. Rhines, T. F. Stocker, L. D. Talley, and J. M. Wallace. Abrupt climate change. *Science (New York, N.Y.)*, 299(5615):2005–10, Mar. 2003.
- [13] J. Alpert and N. Hajaj. We knew the web was big. *The Official Google Blog*, 21, July 25 2008.
- [14] N. Alsaedi, P. Burnap, and O. Rana. Sensing real-world events using arabic twitter posts. In *International AAAI Conference on Web and Social Media, ICWSM'16*, 2016.
- [15] T. Althoff. *Online Actions with Offline Impact : How Online Social Networks Influence Online and Offline User Behavior*. 2017.
- [16] F. Amato, G. Cozzolino, A. Mazzeo, and S. Romano. *Malicious Event Detecting in Twitter Communities*, pages 63–72. Springer International Publishing, Cham, 2016.
- [17] J. An, D. Quercia, and J. Crowcroft. Recommending Investors for Crowdfunding Projects. *Arxiv - Computers & Society*, pages 261–269, 2014.
- [18] I. Androutsopoulos, J. Koutsias, K. V. Chandrinou, and C. D. Spyropoulos. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 160–167, 2000.
- [19] E. Anduiza, C. Cristancho, and J. M. Sabucedo. Mobilization through online social networks: the political protest of the indignados in spain. *Information, Communication & Society*, 17(6):750–764, 2014.
- [20] A. Arasu and H. Garcia-Molina. Extracting structured data from web pages. In *Proc. of the ACM SIGMOD International Conference on Management of Data*, pages 337–348, 2003.
- [21] Y. Artzi, P. Pantel, and M. Gamon. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT '12*, pages 602–606, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [22] J. Aslam, F. Diaz, M. Ekstrand-Abueg, R. McCreadie, V. Pavlu, and T. Sakai. Trec 2014 temporal summarization track overview. Technical report, DTIC Document, 2015.
- [23] C. Atkin and R. Rice. *Theory and principles of public communication campaigns*, 2012.

- [24] L. Backstrom, J. Kleinberg, L. Lee, and C. Danescu-Niculescu-Mizil. Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13*, pages 13–22, New York, NY, USA, 2013. ACM.
- [25] R. Bandari, S. Asur, and B. A. Huberman. The pulse of news in social media: Forecasting popularity. *CoRR*, abs/1202.0332, 2012.
- [26] P. Bao, H.-W. Shen, J. Huang, and X.-Q. Cheng. Popularity prediction in microblogging network: A case study on sina weibo. pages 177–178. International World Wide Web Conferences Steering Committee, 2013.
- [27] N. Barbieri, F. Bonchi, and G. Manco. Who to follow and why: Link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 1266–1275, New York, NY, USA, 2014. ACM.
- [28] J. A. Bargh, K. Y. McKenna, and G. M. Fitzsimons. Can you see the real me? activation and expression of the “true self” on the internet. *Journal of social issues*, 58(1):33–48, 2002.
- [29] G. Baruah, M. D. Smucker, and C. L. Clarke. Evaluating streams of evolving news events. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15*, pages 675–684, New York, NY, USA, 2015. ACM.
- [30] H. Bast and I. Weber. Type less, find more: Fast autocompletion search with a succinct index. In *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 364–371, 2006.
- [31] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. *ICWSM '11*, pages 438–441, 2011.
- [32] R. Bekkerman. Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora. *Computer Science Department Faculty Publication Series, University of Massachusetts, Amherst*, (218), 2004.
- [33] P. Belleflamme, T. Lambert, and A. Schwienbacher. Crowdfunding: Tapping the right crowd. *Journal of Business Venturing*, 29(5):585 – 609, 2014.
- [34] S. Bird, S. Barocas, K. Crawford, F. Diaz, and H. Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. 2016.
- [35] E. Blanzieri and A. Bryl. A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review*, 29(1):63–92, 2008.
- [36] D. Blei and J. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

Bibliography

- [37] D. M. Blei. Probabilistic topic models. *Communication ACM*, 55(4):77–84, Apr. 2012.
- [38] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman. Challenges of computational verification in social multimedia. In *Proceedings of the 23rd International Conference on World Wide Web, WWW '14 Companion*, pages 743–748, New York, NY, USA, 2014. ACM.
- [39] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016.
- [40] D. Boyd and K. Crawford. Critical Questions for Big Data. *Information, Communication & Society*, 15(5):37–41, 2012.
- [41] L. Bravo and M. Lenzerini, editors. *Proceedings of the 7th Alberto Mendelzon International Workshop on Foundations of Data Management, Puebla/Cholula, Mexico, May 21-23, 2013*, volume 1087 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.
- [42] H. Cai, V. W. Zheng, F. Zhu, K. C. Chang, and Z. Huang. From community detection to community profiling. *CoRR*, abs/1701.04528, 2017.
- [43] G. Caruana and M. Li. A survey of emerging approaches to spam filtering. *ACM Computing Surveys*, 44(2):1–27, 2012.
- [44] C. Castillo. *Big Crisis Data: Social Media in Disasters and Time-Critical Situations*. Cambridge University Press, 2016.
- [45] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck. Characterizing the life cycle of online news stories using social media reactions. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '14*, pages 211–223, New York, NY, USA, 2014. ACM.
- [46] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, pages 675–684, New York, NY, USA, 2011. ACM.
- [47] C. Castillo, M. Mendoza, and B. Poblete. Predicting information credibility in time-sensitive social media. *Internet Research*, 23(5):560–588, 2013.
- [48] L. E. Celis, A. Deshpande, T. Kathuria, and N. K. Vishnoi. How to be fair and diverse? *CoRR*, abs/1610.07183, 2016.
- [49] M. Cha, F. Benevenuto, H. Haddadi, and K. Gummadi. The world of connections and information flow in twitter. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 42(4):991–998, July 2012.
- [50] A. Chadwick. The internet, political mobilization and organizational hybridity: 'deanspace', moveon.org and the 2004 us presidential campaign. 2005.

- [51] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009.
- [52] W. Chen, L. V. Lakshmanan, and C. Castillo. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177, 2013.
- [53] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 199–208, New York, NY, USA, 2009. ACM.
- [54] X. Chen, Y. Xia, P. Jin, and J. Carroll. Dataless text classification with descriptive lda. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2224–2231. AAAI Press, 2015.
- [55] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, New York, NY, USA, 2014. ACM.
- [56] J. Choi and W. B. Croft. Temporal models for microblogs. *CIKM '12*, pages 2491–2494, New York, NY, USA, 2012. ACM.
- [57] H. S. Christensen. Political activities on the internet: Slacktivism or political participation by other means? *First Monday*, 16(2), 2011.
- [58] Z. Chu, I. Widjaja, and H. Wang. Detecting social spam campaigns on twitter. In *Applied Cryptography and Network Security*, pages 455–472. Springer, 2012.
- [59] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.
- [60] D. Contractor and T. A. Faruque. Understanding election candidate approval ratings using social media data. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13 Companion, pages 189–190, New York, NY, USA, 2013. ACM.
- [61] C. Corley and R. Mihalcea. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, EMSEE '05, pages 13–18, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [62] G. V. Cormack. Email spam filtering: A systematic review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, 2007.
- [63] G. Corrado. Computer, respond to this email. *The Google Research Blog*, November 3 2015.
- [64] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Bibliography

- [65] S. Counts, M. De Choudhury, J. Diesner, E. Gilbert, M. Gonzalez, B. Keegan, M. Naaman, and H. Wallach. Computational social science. *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW Companion '14*, 323(February):105–108, 2014.
- [66] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, 13(1):21–27, Sept. 2006.
- [67] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008.
- [68] K. Crawford and M. Finn. The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, 80(4):491–502, 2015.
- [69] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang. Discover breaking events with popular hashtags in Twitter. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 1794–1798, New York, NY, USA, 2012. ACM.
- [70] L. Cui, X. Zhang, X. Zhou, and F. Salim. *Topical Event Detection on Twitter*, pages 257–268. Springer International Publishing, Cham, 2016.
- [71] L. A. Dabbish and R. E. Kraut. Email overload at work: An analysis of factors associated with email strain. In *Proc. of the 20th Conference on Computer Supported Cooperative Work*, pages 431–440, 2006.
- [72] T. P. Davies, H. M. Fry, A. G. Wilson, and S. R. Bishop. A mathematical model of the london riots and their policing. *Scientific reports*, 3:1303, 2013.
- [73] D. Di Castro, L. Lewin-Eytan, Y. Maarek, R. Wolff, and E. Zohar. Enforcing k-anonymity in web mail auditing. In *Proc. of the 9th International Conference on Web Search and Data Mining*, 2016.
- [74] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12*, pages 536–544, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [75] Y. Ding. Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4):498–514, 2011.
- [76] X. Dong, D. Mavroudis, F. Calabrese, and P. Frossard. Multiscale event detection in social media. *Data Min. Knowl. Discov.*, 29(5):1374–1405, Sept. 2015.
- [77] M. Eaton. Manufacturing community in an online activist organization. *Information, Communication & Society*, 13(2):174–192, 2010.

- [78] V. Etter, M. Grossglauser, and P. Thiran. Launch Hard or Go Home! *COSN '13*, pages 177–182, 2013.
- [79] G. C. N. Farida Vis. To tackle the spread of misinformation online we must first understand it, 2014.
- [80] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.
- [81] M. Fernandez, L. S. G. Piccolo, D. Maynard, M. Wippoo, C. Meili, and H. Alani. Talking climate change via social media: Communication, engagement and behaviour. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 85–94, New York, NY, USA, 2016. ACM.
- [82] E. Ferrara, R. Interdonato, and A. Tagarelli. Online popularity and topical interests through the lens of instagram. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 24–34, New York, NY, USA, 2014. ACM.
- [83] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Commun. ACM*, 59(7):96–104, June 2016.
- [84] K. Filippova and Y. Altun. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1481–1491, 2013.
- [85] K. Filippova and K. B. Hall. Improved video categorization from text metadata and user comments. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 835–842, New York, NY, USA, 2011. ACM.
- [86] K. Filippova and M. Strube. Dependency tree based sentence compression. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 25–32, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.
- [87] K. Filippova, M. Surdeanu, M. Ciaramita, and H. Zaragoza. Company-oriented extractive summarization of financial news. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 246–254, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.
- [88] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.
- [89] D. Freelon, C. D. McIlwain, and M. D. Clark. Beyond the hashtags:# ferguson,# black-livesmatter, and the online struggle for offline justice. 2016.

Bibliography

- [90] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.
- [91] R. J. Gallagher, A. J. Reagan, C. M. Danforth, and P. S. Dodds. Divergent discourse between protests and counter-protests: #blacklivesmatter and #alllivesmatter. *CoRR*, abs/1606.06820, 2016.
- [92] S. Gao, J. Ma, and Z. Chen. Modeling and predicting retweeting dynamics on microblogging platforms. *WSDM '15*, pages 107–116, New York, NY, USA, 2015. ACM.
- [93] D. Gayo-Avello. No, you cannot predict elections with twitter. *IEEE Internet Computing*, 16(6):91–94, Nov 2012.
- [94] T. Ge, L. Cui, B. Chang, Z. Sui, and M. Zhou. Event detection with burst information networks. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3276–3286, 2016.
- [95] P. Gerbaudo. Tweets and the Streets: Social Media and Contemporary Activism. page 216, oct 2012.
- [96] T. Giorgino. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7):1–24, 8 2009.
- [97] T. Giorgino. Computing and visualizing dynamic time warping alignments in R: The DTW package. *Journal of Statistical Software*, 31(1):1–24, 2009.
- [98] S. Girtelschmid, A. Salfinger, B. Pröll, W. Retschitzegger, and W. Schwinger. Near real-time detection of crisis situations. In *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 247–252, May 2016.
- [99] K. Glasgow and C. Fink. *Hashtag Lifespan and Social Networks during the London Riots*, pages 311–320. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
- [100] K. Glasgow and C. Fink. Hashtag lifespan and social networks during the london riots. In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction, SBP'13*, pages 311–320, Berlin, Heidelberg, 2013. Springer-Verlag.
- [101] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce, EC '12*, pages 623–638, New York, NY, USA, 2012. ACM.
- [102] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi. On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*, pages 1799–1808, New York, NY, USA, 2015. ACM.

- [103] S. Gonzalez-Bailon and N. Wang. Networked discontent: The anatomy of protest campaigns in social media. *Available at SSRN 2268165*, 2013.
- [104] S. González-Bailón, J. Borge-Holthoefer, and Y. Moreno. Broadcasters and hidden influentials in online protest diffusion. *American Behavioral Scientist*, 57(7):943–965, 2013.
- [105] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno. Assessing the bias in samples of large online networks. *Social Networks*, 38:16 – 27, 2014.
- [106] P. A. Grabowicz, M. Babaei, J. Kulshrestha, and I. Weber. The road to popularity: The dilution of growing audience on twitter. *CoRR*, abs/1603.04423, 2016.
- [107] P. A. Grabowicz, N. Ganguly, and K. P. Gummadi. Distinguishing between topical and non-topical information diffusion mechanisms in social media. *CoRR*, abs/1603.04425, 2016.
- [108] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz. Social features of online networks: The strength of intermediary ties in online social media. *PLOS ONE*, 7(1):1–9, 01 2012.
- [109] C. Granger. Some recent development in a concept of causality. *Journal of Econometrics*, 39(1):199 – 211, 1988.
- [110] M. Grbovic, G. Halawi, Z. Karnin, and Y. Maarek. How many folders do you really need?: Classifying email into a handful of categories. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14*, pages 869–878, New York, NY, USA, 2014. ACM.
- [111] A. Guille and C. Favre. Mention-anomaly-based event detection and tracking in twitter. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 375–382, Aug 2014.
- [112] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Rec.*, 42(2):17–28, July 2013.
- [113] Q. Guo, F. Diaz, and E. Yom-Tov. Updating users about time critical events. In *Proceedings of the 35th European Conference on Advances in Information Retrieval, ECIR'13*, pages 483–494, Berlin, Heidelberg, 2013. Springer-Verlag.
- [114] A. Gupta and P. Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media, PSOSM '12*, pages 2:2–2:8, New York, NY, USA, 2012. ACM.
- [115] C. Hachenberg and T. Gottron. Locality sensitive hashing for scalable structural classification and clustering of web documents. In *Proc. of the 22nd ACM International Conference on Information & Knowledge Management*, pages 359–368, 2013.

Bibliography

- [116] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: from discrimination discovery to fairness-aware data mining. KDD Tutorial, 2016.
- [117] S. A. Hale, H. Margetts, and T. Yasseri. Petition growth and success rates on the UK no. 10 Downing street website. In *WebSci '13*, pages 132–138, New York, NY, USA, 2013. ACM.
- [118] M. Halupka. Clicktivism: A systematic heuristic. *Policy and Internet*, 6(2):115–132, 2014.
- [119] Q. He, K. Chang, and E.-P. Lim. Analyzing feature trajectories for event detection. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 207–214, New York, NY, USA, 2007. ACM.
- [120] X. He, M. Gao, M.-Y. Kan, Y. Liu, and K. Sugiyama. Predicting the popularity of web 2.0 items based on user comments. In *Proceedings of the 37th International ACM SIGIR Conference on Research; Development in Information Retrieval*, SIGIR '14, pages 233–242, New York, NY, USA, 2014. ACM.
- [121] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li. Rolx: Structural role extraction and mining in large graphs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1231–1239, New York, NY, USA, 2012. ACM.
- [122] A. Hermida, F. Fletcher, D. Korell, and D. Logan. Share, like, recommend: Decoding the social media news consumer. *Journalism Studies*, 13(5-6):815–824, 2012.
- [123] L. E. Hestres. Preaching to the choir: Internet-mediated advocacy, issue public mobilization, and climate change. *New Media & Society*, page 1461444813480361, 2013.
- [124] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [125] L. Hong, G. Convertino, and E. H. Chi. Language matters in twitter: A large scale study. In *ICWSM*, 2011.
- [126] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 57–58, New York, NY, USA, 2011. ACM.
- [127] S. Hong and D. Nadler. Does the early bird move the polls?: The use of the social media tool 'twitter' by u.s. politicians and its impact on public opinion. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, dg.o '11, pages 182–186, New York, NY, USA, 2011. ACM.
- [128] D. Hoornweg. *Cities and climate change: responding to an urgent agenda*. World Bank Publications, 2011.

- [129] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad. Opening closed regimes: what was the role of social media during the arab spring? 2011.
- [130] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.
- [131] Z. Hu, J. Yao, B. Cui, and E. Xing. Community level diffusion extraction. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 1555–1569, New York, NY, USA, 2015. ACM.
- [132] S.-W. Huang, M. M. Suh, B. M. Hill, and G. Hsieh. How Activists Are Both Born and Made. *CHI '15*, pages 211–220, 2015.
- [133] W. Huang, W. Chen, L. Zhang, and T. Wang. *An Efficient Online Event Detection Method for Microblogs via User Modeling*, pages 329–341. Springer International Publishing, Cham, 2016.
- [134] Z. Huang, A. Olteanu, and K. Aberer. Credibleweb: A platform for web credibility evaluation. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 1887–1892, New York, NY, USA, 2013. ACM.
- [135] E. Hyvönen and E. Mäkelä. Semantic autocompletion. In *Proc. of the 1st Asian Semantic Web Conference*, pages 739–751, 2006.
- [136] G. Iñiguez, J. Török, T. Yasseri, K. Kaski, and J. Kertész. Modeling Social Dynamics in a Collaborative Environment. *arXiv*, page 17, 2014.
- [137] A. Jackoway, H. Samet, and J. Sankaranarayanan. Identification of live news events using twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pages 25–32. ACM, 2011.
- [138] M. Jenders, G. Kasneci, and F. Naumann. Analyzing and predicting viral tweets. In *Proceedings of the 22nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 657–664, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
- [139] S. Jiang, Y. Hu, C. Kang, T. Daly, Jr., D. Yin, Y. Chang, and C. Zhai. Learning query and document relevance from a web-scale click graph. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 185–194, New York, NY, USA, 2016. ACM.
- [140] F. Jin, R. P. Khandpur, N. Self, E. Dougherty, S. Guo, F. Chen, B. A. Prakash, and N. Ramakrishnan. Modeling mass protest adoption in social network communities using geometric brownian motion. *KDD '14*, pages 1660–1669, New York, NY, USA, 2014. ACM.

Bibliography

- [141] Z. Jin, J. Cao, H. Guo, Y. Zhang, Y. Wang, and J. Luo. Rumor detection on twitter pertaining to the 2016 U.S. presidential election. *CoRR*, abs/1701.06250, 2017.
- [142] H. Jing. Sentence reduction for automatic text summarization. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, pages 310–315, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.
- [143] M. Joglekar, T. Rekatsinas, H. Garcia-Molina, A. G. Parameswaran, and C. Ré. Exploiting features for data source quality estimation. *CoRR*, abs/1512.06474, 2015.
- [144] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.
- [145] A. Jungherr and P. Jürgens. The political click : political participation through e-petitions in Germany by. *Internet, Politics, Policy 2010: An Impact Assessment*, 0(September):1–31, 2010.
- [146] K. Y. Kamath and J. Caverlee. Spatio-temporal meme prediction: Learning what hashtags will be popular where. In *Proceedings of the 22Nd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '13, pages 1341–1350, New York, NY, USA, 2013. ACM.
- [147] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukács, M. Ganea, P. Young, and V. Ramavajjala. Smart reply: Automated response suggestion for email. *CoRR*, abs/1606.04870, 2016.
- [148] S. Katragadda, S. Virani, R. Benton, and V. Raghavan. Detection of event onset using twitter. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1539–1546, July 2016.
- [149] C. Kedzie, K. McKeown, and F. Diaz. Predicting salient updates for disaster summarization. In *ACL*, 2015.
- [150] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.
- [151] D. Y. Kenett, F. Morstatter, H. E. Stanley, and H. Liu. Discovering social events through online attention. *PLOS ONE*, 9(7):1–7, 07 2014.
- [152] T. Kenter and M. de Rijke. Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1411–1420, New York, NY, USA, 2015. ACM.
- [153] J. Kim and J.-G. Lee. Community detection in multi-layer graphs: A survey. *SIGMOD Rec.*, 44(3):37–48, Dec. 2015.

- [154] A. P. Kirilenko and S. O. Stepchenkova. Public microblogging on climate change: One year of Twitter worldwide. *Global Environmental Change*, 26:171–182, may 2014.
- [155] S. Kiritchenko and S. Matwin. Email classification with co-training. In *Proc. of the Conference of the Center for Advanced Studies on Collaborative Research*, pages 301–312, 2011.
- [156] R. Kobayashi and R. Lambiotte. Tideh: Time-dependent hawkes process for predicting retweet dynamics. *CoRR*, abs/1603.09449, 2016.
- [157] G. Koneswaran and D. Nierenberg. Global farm animal production and global warming: impacting and mitigating climate change. 2008.
- [158] S. Kong, Q. Mei, L. Feng, F. Ye, and Z. Zhao. Predicting bursts and popularity of hashtags in real-time. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 927–930, New York, NY, USA, 2014. ACM.
- [159] J. Krumm and E. Horvitz. Eyewitness: Identifying local events via space-time signals in twitter feeds. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '15, pages 20:1–20:10, New York, NY, USA, 2015. ACM.
- [160] A. Kulkarni and T. Pedersen. Name discrimination and email clustering using unsupervised clustering and labeling of similar contexts. In *Proc. of the 2nd Indian International Conference on Artificial Intelligence*, pages 703–722, 2005.
- [161] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 591–602, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
- [162] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2335–2338, New York, NY, USA, 2012. ACM.
- [163] A. Kupavskii, A. Umnov, G. Gusev, and P. Serdyukov. Predicting the audience size of a tweet. In *International AAAI Conference on Web and Social Media*, ICWSM'13, 2013.
- [164] N. Kushmerick. *Wrapper induction for information extraction*. PhD thesis, University of Washington, 1997.
- [165] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 957–966. JMLR.org, 2015.

Bibliography

- [166] A. H. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira. A brief survey of web data extraction tools. *ACM Sigmod Record*, 31(2):84–93, 2002.
- [167] H. Lakkaraju, J. J. McAuley, and J. Leskovec. What’s in a name? understanding the interplay between titles, content, and communities in social media. 2013.
- [168] K. Lee, D. Palsetia, R. Narayanan, M. M. A. Patwary, A. Agrawal, and A. Choudhary. Twitter trending topic classification. ICDMW ’11, pages 251–258, Washington, DC, USA, 2011. IEEE Computer Society.
- [169] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. pages 251–260. ACM, 2012.
- [170] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto. Dynamical classes of collective attention in twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW ’12, pages 251–260, New York, NY, USA, 2012. ACM.
- [171] K. Lerman and T. Hogg. Using a model of social dynamics to predict popularity of news. In *Proceedings of the 19th International Conference on World Wide Web*, WWW ’10, pages 621–630, New York, NY, USA, 2010. ACM.
- [172] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’09, pages 497–506, New York, NY, USA, 2009. ACM.
- [173] D. D. Lewis and K. A. Knowles. Threading electronic mail: A preliminary study. *Information Processing & Management*, 33(2):209–217, 1997.
- [174] K. Lewis, K. Gray, and J. Meierhenrich. The structure of online activism. 1:1–9, 2014.
- [175] C. Li, A. Sun, and A. Datta. Twevent: Segment-based event detection from tweets. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM ’12, pages 155–164, New York, NY, USA, 2012. ACM.
- [176] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’16, pages 165–174, New York, NY, USA, 2016. ACM.
- [177] C. Li, J. Xing, A. Sun, and Z. Ma. Effective document labeling with very few seed words: A topic model approach. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, CIKM ’16, pages 85–94, New York, NY, USA, 2016. ACM.
- [178] H. Li, X. Ma, F. Wang, J. Liu, and K. Xu. On popularity prediction of videos shared in online social networks. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM ’13, pages 169–178, New York, NY, USA, 2013. ACM.

- [179] H. Li, D. Shen, B. Zhang, Z. Chen, and Q. Yang. Adding semantics to email clustering. In *Proc. of the 6th International Conference on Data Mining*, pages 938–942, 2006.
- [180] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. Tedas: A twitter-based event detection and analysis system. In *Proceedings of the 2012 IEEE 28th International Conference on Data Engineering, ICDE '12*, pages 1273–1276, Washington, DC, USA, 2012. IEEE Computer Society.
- [181] Y.-R. Lin, D. Margolin, B. Keegan, A. Baronchelli, and D. Lazer. Bigbirds never die: Understanding social dynamics of emergent hashtag. *arXiv preprint arXiv:1303.7144*, 2013.
- [182] S. W. Linderman and R. P. Adams. Discovering latent network structure in point process data. In *ICML*, pages 1413–1421, 2014.
- [183] R. Lindner and U. Riehm. Broadening Participation Through E-Petitions? An Empirical Study of Petitions to the German Parliament. *Policy & Internet*, 3(1):Art. 4, 2011.
- [184] B. Liu, X. Li, W. S. Lee, and P. S. Yu. Text classification by labeling words. In *Proceedings of the 19th National Conference on Artificial Intelligence, AAAI'04*, pages 425–430. AAAI Press, 2004.
- [185] Y. Liu and S. Chawla. Social media anomaly detection: Challenges and solutions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*, pages 2317–2318, New York, NY, USA, 2015. ACM.
- [186] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and danah boyd. The arab spring! the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication*, 5(0), 2011.
- [187] Z. Ma, A. Sun, and G. Cong. Will this #hashtag be popular tomorrow? In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '12*, pages 1173–1174, New York, NY, USA, 2012. ACM.
- [188] Y. Maarek. Web mail is not dead!: It's just not human anymore. In *Proceedings of the 26th International Conference on World Wide Web, WWW '17*, pages 5–5, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.
- [189] W. Magdy and T. Elsayed. Adaptive method for following dynamic topics on twitter, 2014.
- [190] J. Mahmud and H. Gao. Why Do You Spread This Message? Understanding Users Sentiment in Social Media Campaigns. *ICWSM '2014*, pages 607–610, 2014.
- [191] S. Mansfield-Devine. Anonymous: serious threat or mere annoyance? *Network Security*, 2011(1):4 – 10, 2011.

Bibliography

- [192] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, New York, NY, USA, 2011. ACM.
- [193] D. Marom, A. Robb, and O. Sade. Gender dynamics in crowdfunding (kickstarter): evidence on entrepreneurs, investors, deals and taste-based discrimination. 2016.
- [194] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.
- [195] M. Mathioudakis and N. Koudas. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1155–1158, New York, NY, USA, 2010. ACM.
- [196] J. Matusitz. *Terrorism and communication: A critical introduction*. Sage Publications, 2012.
- [197] R. McCreadie, C. Macdonald, and I. Ounis. Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 301–310, New York, NY, USA, 2014. ACM.
- [198] R. McCreadie, C. Macdonald, and I. Ounis. Crowdsourced rumour identification during emergencies. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 965–970, New York, NY, USA, 2015. ACM.
- [199] R. McCreadie, C. Macdonald, I. Ounis, M. Osborne, and S. Petrovic. Scalable distributed event detection for twitter. In *2013 IEEE International Conference on Big Data*, pages 543–549, Oct 2013.
- [200] D. Metzler, C. Cai, and E. Hovy. Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 646–655, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [201] T. Mikolov, G. Corrado, K. Chen, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12, 2013.
- [202] G. A. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
- [203] S. min Kim and E. Hovy. Crystal: Analyzing predictive opinions on the web. In *In EMNLPCoNLL 2007*, 2007.

- [204] T. Mitra and E. Gilbert. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '14*, pages 49–61, New York, NY, USA, 2014. ACM.
- [205] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.
- [206] L. Mosca and D. Santucci. Petitioning online: The role of e-petitions in web campaigning. *Political Campaigning on the Web*, 121, 2009.
- [207] S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd International Conference on World Wide Web, WWW '14*, pages 913–924, New York, NY, USA, 2014. ACM.
- [208] C. Neustaedter, A. Brush, and M. A. Smith. Beyond “from” and “received”: Exploring the dynamics of email triage. In *Extended Abstracts of the Proc. of the ACM Conference on Human Factors in Computing Systems*, pages 1977–1980, 2005.
- [209] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [210] D. T. Nguyen and J. E. Jung. Real-time event detection for online behavioral analysis of big social data. *Future Generation Computer Systems*, 66:137 – 145, 2017.
- [211] D. T. Nguyen and J. J. Jung. Real-time event detection on social data stream. *Mob. Netw. Appl.*, 20(4):475–486, Aug. 2015.
- [212] P. Norris. *The handbook of comparative communication research*. 2012.
- [213] Y. Ogata and K. Katsura. Immediate and updated forecasting of aftershock hazard. *Geophysical Research Letters*, 33(10), 2006. L10305.
- [214] A. OLTEANU. *Probing the Limits of Social Data: Biases, Methods, and Domain Knowledge*. PhD thesis, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE, 2016.
- [215] A. Olteanu, C. Castillo, N. Diakopoulos, and K. Aberer. Comparing Events Coverage in Online News and Social Media : The Case of Climate Change. *ICWSM '15*, pages 288–297, 2015.
- [216] A. Olteanu, C. Castillo, F. Diaz, and E. Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. 2016.
- [217] A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. *ICWSM '14*, 2014.

- [218] A. Olteanu, S. Peshterliev, X. Liu, and K. Aberer. Web credibility: Features exploration and credibility prediction. In *Proceedings of the 35th European Conference on Advances in Information Retrieval, ECIR'13*, pages 557–568, Berlin, Heidelberg, 2013. Springer-Verlag.
- [219] A. Olteanu, I. Weber, and D. Gatica-Perez. Characterizing the demographics behind the #blacklivesmatter movement. *CoRR*, abs/1512.05671, 2015.
- [220] I. Ortiz, S. L. Burke, M. Berrada, and H. Cortés. World protests 2006-2013. 2013.
- [221] N. Panagiotou, I. Katakis, and D. Gunopulos. *Detecting Events in Online Social Networks: Definitions, Trends and Challenges*, pages 42–84. Springer International Publishing, Cham, 2016.
- [222] B. Pang and S. Ravi. Revisiting the predictability of language: Response completion in social media. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12*, pages 1489–1499, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [223] P. Pantel and D. Lin. Spamcop: A spam classification & organization program. In *Proc. of the AAAI Workshop on Learning for Text Categorization*, pages 95–98, 1998.
- [224] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3), May 2012.
- [225] W. Pearce, K. Holmberg, I. Hellsten, and B. Nerlich. Climate Change on twitter: Topics, Communities and Conversations about the 2013 IPCC Working Group rep. *PLoS ONE*, 9(4):1–11, 2014.
- [226] M. Peng, J. Zhu, X. Li, J. Huang, H. Wang, and Y. Zhang. Central topic model for event-oriented topics mining in microblog stream. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, pages 1611–1620, New York, NY, USA, 2015. ACM.
- [227] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*, pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [228] D. Pighin, M. Cornolti, E. Alfonseca, and K. Filippova. Modelling events through memory-based, open-ie patterns for abstractive summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 892–901, 2014.
- [229] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, aug 2008.

- [230] H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 365–374, New York, NY, USA, 2013. ACM.
- [231] R. Procter, F. Vis, and A. Voss. Reading the riots on twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214, 2013.
- [232] R. Prokofyev. *Entity-Centric Knowledge Discovery for Idiosyncratic Domains*. PhD thesis, University of Fribourg, 2016.
- [233] J. Proskurnia, K. Aberer, and P. Cudré-Mauroux. Please sign to save... : How online environmental petitions succeed. In *Social Web for Environmental and Ecological Monitoring, Papers from the 2016 ICWSM Workshop, Cologne, Germany, May 17, 2016*, 2016.
- [234] J. Proskurnia, M.-A. Cartright, L. Garcia-Pueyo, I. Krka, J. B. Wendt, T. Kaufmann, and B. Miklos. Template induction over unstructured email corpora. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1521–1530, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.
- [235] J. Proskurnia, P. Grabowicz, R. Kobayashi, C. Castillo, P. Cudré-Mauroux, and K. Aberer. Predicting the success of online petitions leveraging multidimensional time-series. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 755–764, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.
- [236] J. Proskurnia, R. Mavlyutov, C. Castillo, K. Aberer, and P. Cudre-Mauroux. Efficient document filtering using vector space topic expansion and pattern-mining: The case of event detection in microposts. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, CIKM '17. ACM, 2017.
- [237] J. Proskurnia, R. Mavlyutov, R. Prokofyev, K. Aberer, and P. Cudre-Mauroux. Analyzing large-scale public campaigns on twitter. In *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA*, pages 225–243. Springer International Publishing, 2016.
- [238] H. Purohit, Y. Ruan, D. Fuhry, S. Parthasarathy, and A. Sheth. On understanding the divergence of online social group discussion. In *International AAAI Conference on Web and Social Media*, ICWSM'14, 2014.
- [239] D. Quercia, H. Askham, and J. Crowcroft. TweetLDA. In *Proceedings of the 3rd Annual ACM Web Science Conference on - WebSci '12*, pages 247–250, New York, New York, USA, June 2012. ACM Press.

Bibliography

- [240] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *Proceedings of the 21st International Conference on World Wide Web, WWW '12*, pages 599–608, New York, NY, USA, 2012. ACM.
- [241] M. W. Ragas and S. Kioussis. Intermedia agenda-setting and political activism: Moveon.org and the 2008 presidential election. *Mass Communication and Society*, 13(5):560–583, 2010.
- [242] S. Ravi. On-device machine intelligence, 2017. [Online; accessed 12 Feb 2017].
- [243] A. Ritter, C. Cherry, and W. B. Dolan. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*, pages 583–593, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [244] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12*, pages 1104–1112, New York, NY, USA, 2012. ACM.
- [245] M.-A. RizoIU and L. Xie. Online Popularity under Promotion: Viral Potential, Forecasting, and the Economics of Time. In *11th International AAAI Conference on Web and Social Media*, page 10, 2017.
- [246] M.-A. RizoIU, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck. Expecting to be HIP: Hawkes Intensity Processes for Social Media Popularity. In *World Wide Web 2017, International Conference on*, pages 1069–1078, Perth, Australia, 2017.
- [247] B. Robinson, R. Power, and M. Cameron. A sensitive twitter earthquake detector. In *Proceedings of the 22Nd International Conference on World Wide Web, WWW '13 Companion*, pages 999–1002, New York, NY, USA, 2013. ACM.
- [248] J. Rogstadius, M. Vukovic, C. A. Teixeira, V. Kostakos, E. Karapanos, and J. A. Laredo. Crisistracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, 57(5):4:1–4:13, Sept 2013.
- [249] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, pages 695–704, New York, NY, USA, 2011. ACM.
- [250] Y. Rong, Q. Zhu, and H. Cheng. A model-free approach to infer the diffusion network from event cascade. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16*, pages 1653–1662, New York, NY, USA, 2016. ACM.

- [251] S. D. Roy, T. Mei, W. Zeng, and S. Li. Towards cross-domain learning for social video popularity prediction. *IEEE Transactions on Multimedia*, 15(6):1255–1267, Oct 2013.
- [252] Y. Ruan and S. Parthasarathy. Simultaneous detection of communities and roles from large networks. In *Proceedings of the Second ACM Conference on Online Social Networks, COSN '14*, pages 203–214, New York, NY, USA, 2014. ACM.
- [253] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh. Extracting situational information from microblogs during disaster events: A classification-summarization approach. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15*, pages 583–592, New York, NY, USA, 2015. ACM.
- [254] D. Ruths and J. Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014.
- [255] M. Sadri, S. Mehrotra, and Y. Yu. Online adaptive topic focused tweet acquisition. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16*, pages 2353–2358, New York, NY, USA, 2016. ACM.
- [256] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, pages 851–860, New York, NY, USA, 2010. ACM.
- [257] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Inf. Process. Manage.*, 33(2):193–207, Mar. 1997.
- [258] E. T. K. Sang and J. Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [259] S. Sarawagi. Automation in information extraction and integration. In *Tutorial of the 28th International Conference on Very Large Databases*, 2002.
- [260] A. Schmidt, A. Ivanova, and M. S. Schäfer. Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries. *Global Environmental Change*, 23(5):1233–1248, 2013.
- [261] E. Schubert, M. Weiler, and H.-P. Kriegel. Signitrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 871–880, New York, NY, USA, 2014. ACM.
- [262] R. Schwartz, R. Reichart, and A. Rappoport. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL 2015*, 2015.
- [263] A. Segerberg and W. L. Bennett. Social media and the organization of collective action: Using twitter to explore the ecologies of two climate change protests. *The Communication Review*, 14(3):197–215, 2011.

Bibliography

- [264] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau. Hierarchical neural network generative models for movie dialogues. *CoRR*, abs/1507.04808, 2015.
- [265] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. *CoRR*, abs/1503.02364, 2015.
- [266] H. Shen, D. Wang, C. Song, and A. Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. *CoRR*, abs/1401.0778, 2014.
- [267] H.-W. Shen, D. Wang, C. Song, and A.-L. Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. In *Proc. of AAAI*, pages 291–297. AI Access Foundation, 2014.
- [268] S. Shields and G. Orme-Evans. The impacts of climate change mitigation strategies on animal welfare. *Animals*, 5(2):361–394, 2015.
- [269] B. Shulman, A. Sharma, and D. Cosley. Predictability of popularity: Gaps between prediction and understanding. *CoRR*, abs/1603.09436, 2016.
- [270] P. Slovic. Informing and educating the public about risk. In P. Slovic, editor, *The perception of risk*, pages 182–198. Earthscan, London, England, 2000. Cited by 0000.
- [271] P. Sobkowicz, M. Kaschesky, and G. Bouchard. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4):470 – 479, 2012. Social Media in Government - Selections from the 12th Annual International Conference on Digital Government Research (dg.o2011).
- [272] S. Solomon, G.-K. Plattner, R. Knutti, and P. Friedlingstein. Irreversible climate change due to carbon dioxide emissions. *Proceedings of the national academy of sciences*, pages – 0812721106, 2009.
- [273] Y. Song and D. Roth. On dataless hierarchical text classification. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI’14, pages 1579–1585. AAAI Press, 2014.
- [274] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *CoRR*, abs/1506.06714, 2015.
- [275] B. A. Sparks, H. E. Perkins, and R. Buckley. Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior. *Tourism Management*, 39:1 – 9, 2013.
- [276] E. S. Spiro and C. L. DuBois. Waiting for a retweet: Modeling waiting times in information propagation. 2012.

- [277] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference, ACSAC '10*, pages 1–9, New York, NY, USA, 2010. ACM.
- [278] X. Sun and H. Lin. Topical community detection from mining user tagging behavior and interest. *J. Am. Soc. Inf. Sci. Technol.*, 64(2):321–333, Feb. 2013.
- [279] W. E. Sunday. How misinformation spreads on the internet, 2017.
- [280] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10:10–10:29, Mar. 2013.
- [281] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, Aug. 2010.
- [282] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15 Companion*, pages 977–982, New York, NY, USA, 2015. ACM.
- [283] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications*, 5(1):1–20, 2014.
- [284] J. Till. The broken middle: The space of the london riots. *Cities*, 34:71 – 74, 2013. Urban Borderlands.
- [285] R. Tinati, L. Carr, W. Hall, and J. Bentwood. Identifying communicator roles in twitter. In *Proceedings of the 21st International Conference on World Wide Web, WWW '12 Companion*, pages 1161–1168, New York, NY, USA, 2012. ACM.
- [286] A. Tonon, P. Cudré-Mauroux, A. Blarer, V. Lenders, and B. Motik. Armatweet: Detecting events by semantic tweet analysis. In *ESWC*, pages 138–153, 2017.
- [287] M. Tremayne. Anatomy of protest in the digital era: A network analysis of twitter and occupy wall street. *Social Movement Studies*, 13(1):110–126, 2014.
- [288] Z. Tufekci. "Not This One": Social Movements, the Attention Economy, and Micro-celebrity Networked Activism. *American Behavioral Scientist*, 57(7):848–870, 2013.
- [289] Z. Tufekci. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014.
- [290] Z. Tufekci. Social movements and governments in the digital age: Evaluating a complex landscape. *Journal of International Affairs*, 68(1):1–XVI, Fall 2014. Copyright - Copyright Journal of International Affairs Fall 2014; Document feature - ; Last updated - 2015-01-12; CODEN - JINABJ.

Bibliography

- [291] Z. Tufekci. *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press, 2017.
- [292] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welp. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM '10*, pages 178–185, 2010.
- [293] M. M. Uddin, M. Imran, and H. Sajjad. Understanding types of users on twitter. *CoRR*, abs/1406.1335, 2014.
- [294] O. Vinyals and Q. V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015.
- [295] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.
- [296] F. Vis. Twitter as a reporting tool for breaking news. *Digital Journalism*, 1(1):27–47, 2013.
- [297] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *Proceedings of the 35th European Conference on Advances in Information Retrieval, ECIR'13*, pages 356–367, Berlin, Heidelberg, 2013. Springer-Verlag.
- [298] C. Wang, X. Zhao, Y. Zhang, and X. Yuan. *Online Hot Topic Detection from Web News Based on Bursty Term Identification*, pages 393–397. Springer International Publishing, Cham, 2016.
- [299] Y.-C. Wang, M. Joshi, W. W. Cohen, and C. P. Rosé. Recovering implicit thread structure in newsgroup style conversations. In *Proc. of the 2nd International Conference on Weblogs and Social Media*, pages 152–160, 2008.
- [300] A. Weiler, M. Grossniklaus, and M. H. Scholl. Survey and experimental analysis of event detection techniques for twitter. *The Computer Journal*, 60(3):329, 2017.
- [301] P. Weiner. Linear pattern matching algorithms. In *Proc. of the 14th Annual Symposium on Switching and Automata Theory*, pages 1–11, 1973.
- [302] J. B. Wendt, M. Bendersky, L. Garcia-Pueyo, V. Josifovski, B. Miklos, I. Krka, A. Saikia, J. Yang, M.-A. Cartright, and S. Ravi. Hierarchical label propagation and discovery for machine generated email. In *Proc. of the 9th International Conference on Web Search and Data Mining*, to appear, 2016.
- [303] J. Weng and B.-S. Lee. Event detection in twitter. *ICWSM*, 11:401–408, 2011.
- [304] L. Weng, F. Menczer, and Y. Ahn. Virality prediction and community structure in social networks. *CoRR*, abs/1306.0158, 2013.
- [305] D. Westerman, P. R. Spence, and B. Van Der Heide. Social media as information source: Recency of updates and credibility of information. *Journal of Computer-Mediated Communication*, 19(2):171–183, 2014.

- [306] Wikipedia. Hashtag activism, 2015. [Online; accessed 1 May 2015].
- [307] C. Wilson and A. Dunn. The arab spring| digital media in the egyptian revolution: Descriptive analysis from the tahrir data set. *International Journal of Communication*, 5(0), 2011.
- [308] D. Winseck. Weak links and wikileaks: How control of critical internet resources and social media companies' business models undermine the networked free press. In *Beyond WikiLeaks*, pages 166–177. Springer, 2013.
- [309] W. Xie, F. Zhu, J. Jiang, E. P. Lim, and K. Wang. Topicsketch: Real-time bursty topic detection from twitter. In *2013 IEEE 13th International Conference on Data Mining*, pages 837–846, Dec 2013.
- [310] H. Xu, M. Farajtabar, and H. Zha. Learning granger causality for hawkes processes. *CoRR*, abs/1602.04511, 2016.
- [311] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang. Modeling user posting behavior on social media. SIGIR '12, pages 545–554, New York, NY, USA, 2012. ACM.
- [312] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 177–186, New York, NY, USA, 2011. ACM.
- [313] Yasserli, Taha, Hale, Scott, and Markets, Helen. Modeling the rise in Internet-based petitions. *Journal of Information Technology & Politics*, 9(4):453–470, 2012.
- [314] W. L. Youmans and J. C. York. Social media and the activist toolkit: User agreements, corporate interests, and the information infrastructure of modern social movements. *Journal of Communication*, 62(2):315–329, 2012.
- [315] R. Yu, H. Qiu, Z. Wen, C. Lin, and Y. Liu. A survey on social media anomaly detection. *SIGKDD Explor. Newsl.*, 18(1):1–14, Aug. 2016.
- [316] R. Zafarani, M. A. Abbasi, and H. Liu. *Social Media Mining: An Introduction*. Cambridge University Press, New York, NY, USA, 2014.
- [317] M. J. Zaki, S. Parthasarathy, and W. Li. A localized algorithm for parallel association mining. In *Proceedings of the ninth annual ACM symposium on Parallel algorithms and architectures*, pages 321–330. ACM, 1997.
- [318] H. Zamani and W. B. Croft. Estimating embedding vectors for queries. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pages 123–132, New York, NY, USA, 2016. ACM.
- [319] L. Zhang, J. Zhu, and T. Yao. An evaluation of statistical spam filtering techniques. *ACM Transactions on Asian Language Information Processing*, 3(4):243–269, 2004.

Bibliography

- [320] S. Zhang and S. Vucetic. Semi-supervised discovery of informative tweets during the emerging disasters. *CoRR*, abs/1610.03750, 2016.
- [321] W. Zhang, A. Ahmed, J. Yang, V. Josifovski, and A. J. Smola. Annotating needles in the haystack without looking: Product information extraction from emails. In *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2257–2266, 2015.
- [322] X. Zhang, X. Chen, Y. Chen, S. Wang, Z. Li, and J. Xia. Event detection and popularity prediction in microblogging. *Neurocomputing*, 149, Part C:1469 – 1480, 2015.
- [323] J. Zhao, J. Wu, and K. Xu. Weak ties: Subtle role of information diffusion in online social networks. *Phys. Rev. E*, 82:016105, Jul 2010.
- [324] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1513–1522, New York, NY, USA, 2015. ACM.
- [325] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong. Detecting spammers on social networks. *Neurocomputing*, 159:27 – 34, 2015.

Julia Proskurnia
PhD Student

julia@proskurnia.in.ua

Education

EPFL, Switzerland

PhD in Computer Science (EDIC)

2014–2017

Key words Profiling, Modelling and Facilitating Online Activism.

Distributed Information Systems Lab, Supervisor: Prof. Karl Aberer

KTH Royal Institute of Technology, Sweden

MSc in Distributed Computing (EMDC)

2011–2013

Subjects taken: Advanced Topics in Distributed Computing, Implementation in Distributed Computing. *Notable*

Projects: Gossip Learning on Social Networks. *Master Thesis:* Genium Data Store: Real Time, Low Latency, Reliable, Consistent, Scalable Distributed Data Store for Nasdaq OMX

Universitat Politecnica de Catalunya, Spain

MSc in Distributed Computing (EMDC)

2011–2013

Subjects taken: Distributed and Networked Systems, Parallel Programming Models and Algorithms, Security in Information Technology Systems, Decentralized Systems, Scalable Distributed Systems. *Notable Projects:* Smith-Waterman Algorithm Parallelization - An application written in C++ that parallelizes a serial algorithm using MPI library; Group Membership and Leader Election using ZooKeeper - An implementation with Java of these two distributed systems' primitives.

National Technical University of Ukraine Kiev Polytechnic Institute (NTUU KPI), Ukraine

Systems Analyst, MSc Summa cum laude

2009–2011

Master Thesis: Decision making system for portfolio investment in uncertain conditions. *Subjects taken:* Programs developing and testing, Modern programming technologies, Computer information systems developing, Cryptology, Primary course in intellectual systems projection, Software support of computers networks, System analysis of World Economy, Fuzzy models and methods in intellectual decision making systems.

National Technical University of Ukraine Kiev Polytechnic Institute (NTUU KPI), Ukraine

BSc in Computer Science Summa cum laude

2005–2009

Subjects taken: Mathematical Analysis, Discrete Mathematics, Programming and Algorithmic Languages, Probability Theory and Mathematical Statistics, Numerical Methods, OOP, Computer Networks, Decision making theory in complex system, DB, Statistical Analysis of Economic Processes.

Working Experience

Google Inc.

Software Engineering Intern

Jun-Nov 2015

Gmail comprehension. Personalization of smart reply.

Ecole Polytechnique Federale de Lausanne

Researcher in Distributed Computing Lab

July-Dec 2013

Design a video streaming framework over byzantine highly dynamic infrastructure.

Nasdaq OMX, Stockholm, Sweden

Master thesis internship in Core Development team

Jan–Jun 2013

Designing, developing, testing a distributed data store based on reliable total order multicast abstraction. The data store provides the following properties: real-time, low-latency, reliable, fault-tolerant, consistent and scalable.

Universitat Oberta de Catalunya, Barcelona, Spain

Visiting Researcher in Department of Distributed Computing and Optimisation Summer 2012

AlfaBank Ukraine, Kyiv, Ukraine

Economist

2010–2011

Decision making support. Developing mathematical models of bank processes. Developing reports for the retail business.

City administration of Kyiv, Ukraine

Analyst

2009–2010

General planning of the city development. Statistical data processing. Companies assessment. Profitability of projects.

Megaland, Ukraine

Sales Analyst

2007–2008

Sales forecast. Planning of the cash flow. Reporting development. Price formation. Management inventory.

Honors and awards

GHC Facebook Scholarship Recipient **2016**

Doctoral fellowship, EDIC program in EPFL **2013-2014**

EMEA Google Anita Borg *Scholarship* Winner **April 2012**

Awarded "Faculty Pride 2011" **March 2011**

Awarded with named scholarship by Parliament and deputy Andrievsky **Jul 2010-Jan 2011**

1st place at the National level of European BEST Engineering Competitions **Aug 2010**

Awarded for the high achievements in educational and scientific activities at Institute of Applied System Analysis, the best student of the faculty **May 2010**

Named Scholarship of Academician Daleckiy **Feb-Jun 2010**

Participations

Book Reviews

Apache Kafka, October 2013, PACKT Publishing, ISBN : 1782167935

Cassandra Data Modeling and Analysis, Summer 2014, by PACKT Publishing

Publications/Conferences

Proskurnia J., Mavlyutov, R., Castillo C., Aberer K., Cudre-Mauroux P. *Efficient Document Filtering Using Vector Space Topic Expansion and Pattern-Mining*. CIKM'17.

Proskurnia J., Grabowicz P., Kobayashi R., Castillo C., Cudre-Mauroux P., Aberer K. *Predicting the Success of Online Petitions Leveraging Multidimensional Time-Series*. WWW'17.

Proskurnia J., Cartright M. , Garcia-Pueyo L., Krka I., Wendt J., Kaufmann T., Miklos B. *Template Induction Over Plain Text Corpora*. WWW'17.

Proskurnia J., Cudre-Mauroux P., Aberer K. *Please Sign to Save...: How Online Environmental Petitions Succeed*. ICWSM, SWEEM'16.

Proskurnia J., Mavlyutov R, Prokofyev R.: *Analyzing Large-Scale Public Campaigns on Twitter*. SocInfo'16.

Cabrera, G., Juan, A., Lazaro, D., Marques, J., Proskurnia, I.: *A simulation-optimization approach to deploy Internet services in large-scale systems with user-provided resources* Simulation: Transactions of the Society for Modeling and Simulation International (indexed in ISI SCI, 2012 IF = 0.692, Q3). ISSN: 0037-5497. 2014

Proskurnia Iu. S., Bruzgys Z., Girdzijauskas S. *Gossip learning with linear models on fully distributed data over clustered graphs*. 15-th International Conference SAIT 2013.

Proskurnia Iu. S., Marques J.M. *Large-scale Decentralized Storage Systems used by Volunteer Computing*. 14-th International Conference SAIT 2012.

Proskurnia Iu. S. *Research of the Differential Evolution method for neural network TSK and FOTSK learning. Analysis of Forecast Direction method for FOTSK and GMDH in forecasting problems*. 13-th International Conference SAIT 2011.

Proskurnia, I.S., Grivko B.S. *Analysis and optimization of investor portfolio in fuzzy conditions*. Mathematical and computer modelling 2010, Ukraine.

Proskurnia I.S., Grivko B.S. *Researching of modification FOTSK of neural network TSK in forecasting problems*. Information Models of Knowledge 2010. 470, 177-185.

Grivko B.S., Proskurnia I.S. *Using of forecasting methods for portfolio making in fuzzy conditions*. 12-th International Conference SAIT 2010.

Proskurnia I.S., Grivko B.S. *Using Adaptive Kalman filtering for prediction sales dynamics*. New Technologies 2009, No 2(24). 162, 76-81.

Languages

Russian – native language

Ukrainian – native language

English – speak fluently and read/write with high proficiency

Lithuanian, French – speak, read/write at basic level

