EPFL

Thèse n° 8991

The Role of Compromised Accounts in Social Media Manipulation

Présentée le 21 octobre 2022

Faculté informatique et communications
Laboratoire de systèmes d'information répartis
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Tugrulcan ELMAS

Acceptée sur proposition du jury

Prof. R. West, président du jury
Prof. K. Aberer, directeur de thèse
Prof. S. Zannettou, rapporteur
Prof. M. Humbert, rapporteur
Prof. A. Bosselut, rapporteur

École polytechnique fédérale de Lausanne

2022

"Don't believe everything you read on the internet."
— Sun Tzu

Dedicated to truth and trust... And to my family :)

Acknowledgements

First and foremost, I would like to thank my advisor, Karl Aberer. He gave me perhaps the greatest gift of my life: a job that I'm passionate about. Thanks to his vision and his advice that directed me in the right direction, I could build my own research stream and enjoyed the process. Karl never enforced me to optimize for papers, citations, grants, etc. but encouraged me to accomplish something great that satisfy academic curiosity.

Secondly, I would like to thank my de-facto co-advisor Rebekah Overdorf whom I met in my second year. She taught me how to do research in practice by involving in the process. Thanks to her, I learned how to develop and present my ideas. She also took her time to help with my grant proposals which I'm so grateful for.

I would like to thank my thesis committee: Robert West, Mathias Humbert, Antoine Bosselut, and Savvas Zannettou. They took their time to read this thesis and gave me a lot of valuable feedback.

I sincerely appreciate having Carmela Troncoso as my mentor in EPFL. I e-mailed her for almost every job opportunity I have in my last year to seek advice, to which she replied with great attention and details that helped me to put my academic career on track.

Many thanks to my colleagues in LSIR: Chantal, for keeping up with my recent e-mail flood after corona is over and conferences turned to in-person, Remi, for helping me with my work and translation of the thesis abstract, Marco for helping me in wrapping up my latest project, Jeremie, for saving my oral examination, Thang, for accompanying me in fun trips out of town, and Panayiotis, for helping me numerous times and accompanying me in Athens.

I was lucky to work with many brilliant students who helped me to write awesome papers, including but not limited to Ahmed Furkan Özkalay, Alfonso Amayuelas, Thomas Romain Ibanez, Alexandre Hutter, Mathis Randl, Youssef Attia. It was a pleasure! I especially thank Kristina Hardi who became my first collaborator, and first and last office mate in EPFL.

In my Ph.D. journey, I'm accompanied by awesome mates I'm thankful for: my dorm-mates Jennifer, Francesco, and Elizabeth, my history-lover mate Stefan, my book-lover mate Sena, my yoga-lover mate Eda, and my ski-lover

mate Sina. I especially thank Ozan for being my buddy for the first two years and it was so sad that he left early. Special thanks to the epic EPIC presidents, Karen and Vinitra, for organizing awesome activities.

I thank my friends from my home university who are still beside me remotely: Abdullah, for being the best Ph.D. buddy overseas, Sevgi, for the short Erasmus experience in Leuven, Melisa, for the exciting remote working adventures, Mehmet, for the amazing vacation in Kayseri, Ozancan, for the road trip to Andorra, Haluk, for being my buddy in Istanbul, and Ömert, for letting me crash his couch after a late night concert in Zurich. Special thanks to Zafer and Ali for telling me the best-kept secrets of China and Australia.

Next, the vaccination lovers! I first thank Onur, who is the first person I know in EPFL. I'm very glad that I met him, he helped me with my studies, career, bureaucracy, and everything. Later, he multiplied and introduced me to Nathalie and Omnia, whom I'm delighted to know :) Many thanks to Arzu and Ayberk for all the quality time in their home or outside. I thank Serdar for the fun coffee breaks in gas stations during our trips. Thanks to Utku and Gülseren for establishing the Norman Gaming Center. Thanks to Ezgi for the crazy fasıl nights in Eaux-Vives. Finally, thanks Anıl for forming the EPFL oriental band with me!

Uh, and the Mamas mafia! I thank Yiğit for the late-night smackdown experience, Ezgi for sponsoring çiğköftes on my birthday, Alparslan, for being my closest friends in terms of geographical coordinates, Ekin, for being the second syllable of Mallinedonette, Rojda, for doing my boring homework, and Ayyüce, for bearing my trolls the whole time. Special thanks to Ata for keeping up with my fast-paced travel itineraries!

Finally...

My brothers Abdullah and Emrah. You became my role models because you were both studying, socializing, and traveling a lot. In the end, I traveled more ;). I might not have become a decent computer scientist if you had not taught me how to play Street Fighter II: The New Challengers on Sega MegaDrive so thank you very much! I also thank my extended family: Gül, Safire, Duru, Kemal, and Kutay. You are all beautiful people! :) I wish you happy lives with your families.

My father Erturan. You always wanted me to be a doctor. I am glad that I made your dreams come true :)

My sweet mother Aynur. I would not have come this far if you had not spent your nights reading The Little Match Girl to me when I was sick.

Lausanne, September 23, 2022 T. J.

Abstract

In recent years we have seen a marked increase in disinformation including as part of a strategy of so-called hybrid warfare. Adversaries not only directly spread misleading content but manipulate social media by employing sophisticated techniques that exploit platform vulnerabilities and avoid detection. It is getting increasingly important to analyze social media manipulation to better understand, detect and defend public dialogue against it.

In this thesis, we contribute to the research on social media manipulation by describing and analyzing how adversaries employ compromised social media accounts. We begin by providing a background of social media: we describe the mechanisms and the influence of the platforms to better understand why the adversaries target them. We then give a detailed overview of social media manipulation, and the techniques to detect and counter it. Next, we discuss our contributions in this thesis: 1) an extensive *analysis* of an attack on social media algorithms using compromised accounts, 2) a study of the *implications* of compromised bots for bot research through the characterization of retweet bots, 3) a *detection* method to find compromised accounts that are later repurposed.

Firstly, we uncover and analyze a previously unknown, ongoing astroturfing attack on the popularity mechanisms of social media platforms: ephemeral astroturfing attacks. In this attack, a chosen keyword or topic is artificially promoted by coordinated and inauthentic activity to appear popular. Crucially, this activity is removed as part of the attack which facilitates using compromised accounts that are still managed by their original owners. We observe such attacks on Twitter trends and find that these attacks are not only successful but also pervasive. We detected over 19,000 unique fake trends promoted by over 108,000 accounts. Trends astroturfed by these attacks account for at least 20% of the top 10 global trends. We created a Twitter bot to detect the attacks in real-time and inform the public.

Secondly, we study the implications of compromised accounts to bot research. We do this by characterizing retweet bots that have been uncovered by purchasing retweets from the black market. We determine that those accounts were compromised as they observe anomalous behavior, share spam, and self-state that they are hacked. We then analyze their differences from human-controlled accounts. From our findings on the nature and life-cycle of retweet bots, we point out several inconsistencies between the retweet bots used in

this work and bots studied in prior works. Our findings challenge some of the fundamental assumptions related to bots and in particular how to detect them.

Thirdly, we define, describe, and provide a detection method for *misleading repurposing*, in which an adversary changes the identity of a potentially compromised social media accounts via, among other things, changes to the profile attributes in order to use them for a new purpose while retaining their followers. We propose a methodology to flag repurposed accounts that uses supervised learning on data mined from the Internet Archive's Twitter Stream Grab. We found over 100,000 accounts that may have been repurposed. We also characterize repurposed accounts and found that they are more likely to be repurposed after a period of inactivity and deleting old tweets. We also provide evidence that adversaries target accounts with high follower counts to repurpose and some make them have high follower counts by participating in follow-back schemes. We present a tool to root out accounts that became popular and repurposed later.

Our work is significant in presenting how breaches of user security jeopardize platform security and public dialogue. Furthermore, it enhances the knowledge of how the bots and troll accounts work and aid platforms and researchers in building new solutions.

Keywords: *social media manipulation, compromised accounts, bots, trolls, social media security, disinformation, social media*

Résumé

Ces dernières années, nous avons assisté à une nette augmentation de la désinformation, notamment dans le cadre d'une stratégie de guerre dite hybride. Les adversaires ne se contentent pas de diffuser directement des contenus trompeurs, mais manipulent les médias sociaux en employant des techniques sophistiquées qui exploitent les vulnérabilités des plateformes et évitent d'être détectés. Il devient de plus en plus important d'analyser la manipulation des médias sociaux pour mieux comprendre, détecter et défendre le dialogue public contre elle.

Dans cette thèse, nous contribuons à la recherche sur la manipulation des médias sociaux en décrivant et en analysant comment les adversaires utilisent des comptes de médias sociaux compromis. Nous commençons par présenter le contexte des médias sociaux : nous décrivons les mécanismes et l'influence des plateformes pour mieux comprendre pourquoi les adversaires les ciblent. Nous donnons ensuite un aperçu détaillé de la manipulation des médias sociaux, ainsi que des techniques permettant de la détecter et de la contrer. Ensuite, nous discutons de nos contributions dans cette thèse : 1) une *analyse* approfondie d'une attaque sur les algorithmes des médias sociaux utilisant des comptes compromis, 2) une étude des *implications* des bots compromis pour la recherche sur les bots par la caractérisation de bots de retweet, 3) une méthode de *détection* pour trouver les comptes compromis qui sont ensuite réutilisés.

Tout d'abord, nous découvrons et analysons une attaque d'astroturfing en cours, jusque-là inconnue, sur les mécanismes de popularité des plateformes de médias sociaux : les attaques d'astroturfing éphémères. Dans cette attaque, un mot-clé ou un sujet choisi est artificiellement promu par une activité coordonnée et non authentique pour apparaître populaire. Surtout, cette activité est supprimée dans le cadre de l'attaque, ce qui facilite l'utilisation de comptes compromis qui sont toujours gérés par leurs propriétaires d'origine. Nous observons de telles attaques sur les tendances Twitter et constatons que ces attaques sont non seulement réussies mais aussi omniprésentes. Nous avons détecté plus de 19 000 fausses tendances uniques promues par plus de 108 000 comptes. Les tendances issues de ces attaques d'astrosurfing éphémères représentent au moins 20% des 10 principales tendances mondiales. Nous avons créé un bot Twitter pour détecter les attaques en temps réel et informer le public.

Deuxièmement, nous étudions les implications des comptes compromis sur la recherche de robots. Pour ce faire, nous caractérisons les robots de retweet qui ont été découverts en achetant des retweets sur le marché noir. Nous déterminons que ces comptes ont été compromis car ils observent un comportement anormal, partagent du spam et déclarent qu'ils ont été piratés. Nous analysons ensuite leurs différences par rapport aux comptes contrôlés par l'homme. À partir de nos conclusions sur la nature et le cycle de vie des robots de retweet, nous soulignons plusieurs incohérences entre les robots de retweet utilisés dans ce travail et les robots étudiés dans des travaux antérieurs. Nos résultats remettent en question certaines des hypothèses fondamentales liées aux bots et en particulier comment les détecter.

Troisièmement, nous définissons, décrivons et fournissons une méthode de détection de la réaffectation trompeuse, dans laquelle un adversaire change l'identité d'un compte de médias sociaux potentiellement compromis en modifiant, entre autres, les attributs du profil afin d'utiliser le compte à une nouvelle fin tout en conservant ses followers. Nous proposons une méthodologie pour signaler les comptes réaffectés qui utilise l'apprentissage supervisé sur des données extraites du Twitter Stream Grab d'Internet Archive. Nous avons trouvé plus de 100 000 comptes susceptibles d'avoir été réaffectés. Nous caractérisons également les comptes réaffectés et constatons qu'ils sont plus susceptibles d'être réaffectés après une période d'inactivité et de suppression d'anciens tweets. Nous fournissons également des preuves que les adversaires ciblent les comptes ayant un nombre élevé de followers pour les réaffecter et que certains d'entre eux leur donnent un nombre élevé de followers en participant à des systèmes de follow-back. Nous présentons un outil permettant d'éliminer les comptes qui sont devenus populaires et ont été réaffectés par la suite.

Notre travail est important pour montrer comment les atteintes à la sécurité des utilisateurs compromettent la sécurité de la plateforme et le dialogue public. En outre, il améliore la connaissance du fonctionnement des bots et des comptes de trolls et aide les plateformes et les chercheurs à créer de nouvelles solutions.

Mots-clés: *manipulation des médias sociaux, comptes compromis, bots, trolls, sécurité des médias sociaux, désinformation, médias sociaux*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Social media is one of the primary means of communication today. People worldwide use social media platforms to socialize, share photos, watch cat videos, keep up with their favorite celebrities, follow the news, and even discuss politics. The widespread usage of social media empowers platforms to influence the public, and such influence has attracted adversaries who aim to exploit them for malicious purposes. Adversaries can spread their harmful narratives and make them reach the masses in seconds through social media. Lies travel faster than the truth, especially when they breed on social media platforms. Indeed, we have seen a marked increase in disinformation, including as part of a strategy of so-called hybrid warfare in recent years. The disinformation campaigns often targeted topics such as elections, climate change, and vaccines, which are critical to public health [258].

Adversaries manipulate social media to bring their malicious campaigns to public attention. To do so, they often employ sophisticated techniques to exploit platforms' vulnerabilities while avoiding detection by internal and external investigators. For instance, they deploy automated accounts, colloquially named "bots", at scale to promote certain users, posts or narratives by inflating the popularity metrics (e.g., like counts) of these entities [68]. The platforms may use these metrics as a proxy for reputation and further amplify those with inflated metrics, e.g., suggest them to other users [271]. Additionally, they exploit policy-related vulnerabilities of the platforms. For instance, an external researcher collected Facebook data of users using their friends' permission (instead of the users themselves.) The research acquired the data of 87 million users by only getting collection permission from 270 thousand users. The data company named Cambridge Analytica used this data to profile voters and sway elections through targeted advertisements. The incident is named the "Cambridge Analytica scandal" and encouraged many countries to take precautions against illegal data collection and usage, such as GDPR [241].

A growing body of research on social media manipulation focuses on the detection of, analysis of, and counter-measures against the manipulations. The studies initially

concentrated on less sophisticated strategies such as employing spam bots that aggressively extend their network and share advertisements [34, 313]. However, in recent years, researchers reported that adversaries adopt sophisticated strategies such as social media bots that imitate humans [114] and mix automated behavior with human behavior (named as a cyborg) [59]. Such strategies create a challenge for automated systems and even for humans to differentiate malicious users from legitimate users. Thus, adversaries employing those strategies are more successful at staying under the radar and manipulating social media [69].

Nevertheless, while the current research demonstrates that adversaries use malicious social media accounts with human-like behavior, they fell short of explaining *how* they do it. More work is needed to understand how adversaries create human-like social media accounts, how they use them, and their impacts so that the platforms and researchers can build counter-measures against such accounts. The thesis contributes to a critical aspect of the problem: adversaries compromise legitimate social media accounts and use them as bots, which is one of the strategies to have human-like bots. It is also the first work that shows the adversaries take partial control of social media accounts but let their human owners still use them so that they can confuse the bot detection systems. Moreover, it proposes a detection method for compromised accounts that are later repurposed, which are difficult to detect due to data limitations.

1.2 Thesis Statement and Contributions

The overarching goal of this thesis is to present the role of compromised accounts in social media manipulation. We summarize this with the following thesis statement:

> **Thesis Statement** *The current research assumes that adversaries employ accounts created for malicious purposes to manipulate social media. This thesis proposes that they also compromise legitimate accounts, which we must counter with appropriate detection systems and defenses.*

To accomplish our goal, we present 1) an extensive *analysis* of how adversaries use compromised accounts to manipulate social media using the attacks on Twitter trends as a case study, 2) the *implications* of compromised bots on the bot research by studying the retweet bots' characteristics, 3) a *detection* methodology to find compromised accounts in the wild that are repurposed. For each step, we state our research problems and contributions.

Analysis— Analysis of Attacks Employing Compromised Accounts: The Case of Ephemeral Astroturfing Past studies analyzed bots that attack social media in a coordinated manner. However, to the best of our knowledge, no one studied how the adversaries employed the compromised accounts in such attacks. Using the ephemeral astroturfing attacks on Twitter trends as a case study, we present an extensive analysis of how compromised accounts were used in the attacks, the success of these attacks, and their implications. We tackle the following research questions:

- How do the ephemeral astroturfing attacks on Twitter trends work?
- How do the adversaries use compromised accounts in these attacks?
- How to detect these attacks and the bots used in the attacks?
- What are the implications of these attacks on Twitter trends, platform security, and society as a whole?
- What are the possible counter-measures against these attacks and the exploitation of compromised accounts?

By tackling these questions, we 1) present the first case study of social media manipulation using compromised accounts, 2) defined and described ephemeral astroturfing attacks for the first time, 3) performed the first large-scale analysis of Twitter trend manipulation, 4) found that over 19,000 fake trends were created by these attacks 5) fake trends make up 47% of local Twitter trends in a specific country and 20% of global Twitter trends 6) detected 108,000 bots which is the largest bot dataset reported in a single paper to the best of our knowledge 7) present a counter-measure by detecting the attacks in real-time and inform the public.

We present this study in detail in Chapter 3. It was published in:

> Elmas, Tuğrulcan, Rebekah Overdorf, Ahmed Furkan Özkalay, and Karl Aberer. "Ephemeral astroturfing attacks: The case of fake twitter trends." In 2021 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 403-422. IEEE, 2021.

The first author initiated the project, performed the analysis, created the visualizations, and wrote down the results. The second author contributed with analysis ideas and helped with the writing and supervision. The third author helped with data annotation, created two plots, and proposed an extra feature to analyze. The fourth author supervised the project.

The §3.11 of this chapter is accepted as a poster to Truth and Trust 2022. The author of this thesis initiated the project, performed the analysis, created the visualizations, and wrote down the results. The thesis director supervised the project.

Implications— Implications of Compromised Accounts on Bot Research: The Case of Retweet Bots: Research on social media bots often define their assumptions before data collection, annotation, or detection methodology. However, the validity of such assumptions is not well established in the literature. We particularly tackle the assumption that the adversaries use sophisticated strategies to pass social media bots as humans and show that they also use compromised accounts to pass as humans instead. To do that, we use a dataset of bots used for paid retweets. We characterize those accounts by comparing their behaviors with human accounts. Our research questions are as follows:

- Where do these retweet bots come from? Were they created for this purpose or compromised?
- What is the lifetime of these retweet bots?
- How do the retweeters in our dataset act differently from human users?
- Are there any differences between the bots examined in this work and those found in prior studies?

In answering these questions, we 1) present the first study focusing on retweet bots exclusively; 2) characterize retweet bots, providing evidence that some are mass-compromised and used aggressively; 3) challenge fundamental assumptions about the nature of bot accounts, such as account age and over-activity; and 4) we discuss implications on and challenges in bot detection.

We present this study in detail in Chapter 4. It was published in:

> Elmas, Tuğrulcan, Rebekah Overdorf, and Karl Aberer. "Characterizing Retweet Bots: The Case of Black Market Accounts." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 171-182. 2022.

The first author initiated the project, performed the analysis, created the visualizations, and wrote down the results. The second author helped with the writing and supervision. The third author supervised the project.

Detection— Detecting Repurposed Compromised Accounts in the Wild: The Case of Misleading Repurposing: As compromised accounts pose a problem to social media and the web in general, several studies tackled the problem of their detection [96, 161, 162]. However, they are limited to detecting accounts whose data before getting compromised still exists, or the account is used for a malicious purpose. On social media platforms such as Twitter, adversaries can compromise accounts, delete their past data and repurpose it for a legitimate purpose. On Twitter, such a change goes unnoticed by the accounts' followers; thus, popular accounts can get compromised and sold over the market. We name this strategy "misleading repurposing". We tackle detecting misleading repurposing, which has not been studied before to the best of our knowledge. We also propose a visualization tool so that researchers can investigate such repurposed accounts. We answer the following research questions:

- How does the misleading repurposing work?
- What are the characteristics of the accounts that are repurposed?
- How to detect misleading repurposing using public data?
- What are the implications of and counter-measures against misleading repurposing?

By answering these questions, we 1) introduce the concept of misleading repurposing and suggest a definition, 2) present the first large-scale study of misleading repurposing, 3) establish a hand-labeled ground-truth dataset of repurposed accounts using datasets published by Twitter, 4) study the characteristics repurposed accounts 5) propose a visualization tool to study repurposed accounts.

We present this study in detail in Chapter 5. It was resubmitted to the ICWSM 2023 after minor revisions. Here is the preprint version:

> Elmas, Tuğrulcan, Rebekah Overdorf, and Karl Aberer. "Misleading repurposing on twitter." arXiv preprint arXiv:2010.10600 (2022).

The first author initiated the project, performed the analysis, created the visualizations, and wrote down the results. The second author helped with the writing and supervision. The third author supervised the project.

The §5.11 of this chapter was accepted as a talk to Truth and Trust 2021. Two master's students in EPFL developed the tool presented in the section as a semester project, Thomas Ibanez and Alexandre Hutter. The author of this thesis initiated the project, provided the data and the methodology, and has written down the results. Dr. Rebekah Overdorf and Dr. Karl Aberer supervised the project.

Our caveat is that in this thesis, we tackle the research problems using data from a single social media platform, Twitter, and used a single case study in some cases. While the results may generalize to other datasets and social media platforms in theory, our goal is not to come up with observations and approaches that apply to every context. We are rather interested in explaining the role of compromised accounts using a specific context (i.e., a social media platform or a dataset) to better understand how such accounts can be used to manipulate social media.

1.3 Thesis Outline

We now introduce the organization of this thesis. We first provide the background where we provide an in-depth analysis of the current work that motivates this thesis. We then extensively lay out the three main contributions towards presenting the role of compromised accounts in social media manipulation in chapters 4, 5, and 6. We conclude by summarizing contributions, discussing future work, and giving recommendations. Fig. 1.1 summarizes the thesis outline. Here is the detailed outline of the thesis:

Chapter 2 puts the thesis in the context of current research and motivates it. We specifically survey the inner workings of media and social media and discuss why they are targets. We then provide an overview of social media manipulation, its detection, and counter-measures. We also motivate using Twitter as the context for our study. In each section, we highlight our contributions to the body of research we review.

Chapter 3 presents an extensive analysis of a new attack, ephemeral astroturfing attack on Twitter trends, in which compromised accounts were used as bots to promote harmful

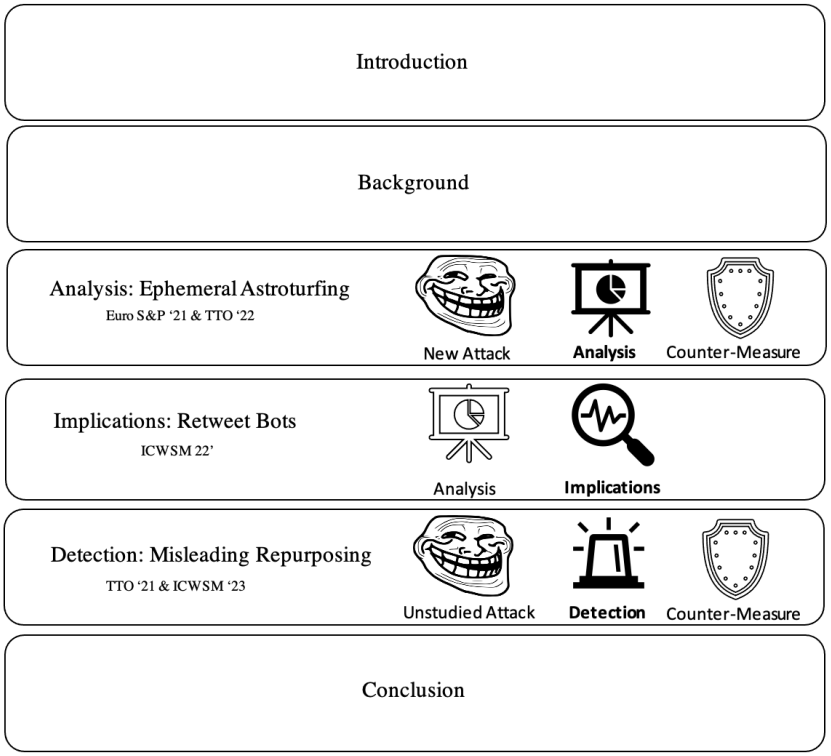

Figure 1.1: The overview of the thesis outline. Each chapter has two to three contributions. The main contributions are bolded.

narratives on top of Twitter. It also presents a counter-measure that detects attacks in real-time and informs the public.

Chapter 4 presents a study on the implications of the compromised bots on bot research. To do that, it analyzes the characteristics of compromised retweet bots. It also describes the current studies' fall-backs.

Chapter 5 presents an unstudied attack which is repurposing compromised accounts, and a methodology to detect such compromised accounts. It also proposes a visualization tool for researchers to root out such accounts as a counter-measure.

Chapter 6 concludes the thesis with a summary of our contributions, a discussion of future work, and recommendations.

Chapter 2

Background

This chapter provides background information to put our study into context. This chapter is not an exhaustive literature review on social media studies but a brief overview to understand the motivations of our work. As a starting point, we describe the traditional media and its similarities to social media. We then describe social media and its impact, which makes it vulnerable to manipulation. Next, we define and describe social media manipulation, which consists of messages being spread, the techniques used, and the vulnerabilities exploited. We then discuss detection and counter-measures against the manipulations. We lastly discuss our motivations to focus on Twitter in this thesis.

2.1 The Traditional Media

In this section, we first define media, describe its mechanics, and explain why it is a target for manipulation. We then discuss the similarities and differences with social media.

Preliminaries Media is "the means of communication, as radio and television, newspapers, magazines, and the internet, that reach or influence people widely."[1] The media plays the role of a mediator in politics. They provide the people with news and opinions in democratic societies and shape their political beliefs. In doing so, they apply their own principles and assessments and thus, transform the politics [219].

Mechanics Media do gate-keeping through editorial decision-making. They decide which events and issues are covered in the news and brought to the public attention [251]. They set the agenda: rank the importance of news over others and thus, influence the people's ideas of what comprises the most pressing issues of society [61]. When providing the news, they may pay particular attention to an aspect of the issue over the others, which is named "framing" [107]. Their decisions on selection, ranking, and framing of news signal their underlying principles in the editorial process. Consistent patterns of favoring one side against another in editorial decision-making may signify news bias [108].

[1] https://www.dictionary.com/browse/media

For instance, a newspaper may give more coverage to news related to immigrants, make them appear in the headlines, and frame them as a security issue, which may show that they are biased against the immigrants.

Manipulation In a democratic environment, the media may be the primary target to influence the public and sway elections due to its ability and efficiency in spreading opinions. Thus, adversaries may employ stealthy techniques to gain a favorable image of themselves, known as media manipulation [65]. They can have more and/or positive media coverage, attack their opponents, or spread their frames and narratives over the news [212]. For instance, adversaries plant fake news on a website or a less credible news outlet and then alert legitimate news outlets to report this new information, which is named information laundering [193].

Connection to Social Media Social media inherits all those properties of the traditional media: it can too do gatekeeping, set the agenda, frame particular narratives, and be biased. However, social media works in different ways: the public itself contributes to the news production, the editorial process is weak or non-existent, and the algorithms play a significant part in what is brought to the public attention. Consequently, it also has the potential to influence the public, which makes it vulnerable to manipulation. We now describe the concept of social media in detail and how it influences the public.

2.2 The Social Media and Its Influence

In this section, we first define social media and describe its mechanics. We also provide a brief overview of its impact on the public, which makes it a target for manipulation.

Preliminaries Social media is defined as "websites and other online means of communication that are used by large groups of people to share information and to develop social and professional contacts"[2]. Such websites, which we refer to as social media platforms, are widely used worldwide. As of 2022, Facebook has 2.9 billion, Instagram has 1.48 billion, and Twitter has 436 million active users[3].

Mechanics A typical social media platform consists of users linked to each other by unidirectional or bidirectional relationships, which makes it a social network. In the traditional media, the information is produced by the news agency and flows to consumers in a top-down manner. However, on social media, the information can flow from anyone in any direction. Additionally, the editorial process is often very weak or non-existent: the information produced by social media users is not reviewed beforehand. Platforms generally remove information retrospectively after it is already posted and flagged by another user [271]. Thus, the information can spread quickly on social media as it can be shared by anyone and does not get reviewed, making the platforms targets for those

[2]https://www.dictionary.com/browse/social-media
[3]https://datareportal.com/reports/digital-2022-global-overview-report

who want to influence the public. The adversaries craft fake news that inspires fear, disgust, and surprise in people, which makes them spread faster than true news on social media [293].

Furthermore, the abundance of information of any quality encourages platforms to apply additional mechanics such as algorithms, design decisions, and platform policies to do gate-keeping and curate information. Those mechanics can further influence public behavior. In 2010, Facebook showed that a single tweak on the user interface could affect the voting decision of hundreds of thousands. They conducted a field experiment to study the effect of social contagion on voter turnout. In the experiment, the platform displayed a social message to people in the treatment group, showing a random set of friends who declared that they voted in the elections. The researchers indicate that this treatment itself increased voter turnout by 340,000 voters, which represented 0.14% of the voting population at the time [39]. This is an excellent example of the extent of platforms' influence.

Impact The mechanics of social media have an impact on the public. Such an impact can be positive. For instance, Moyer et al. [202] showed that popular posts on Reddit have increased page-views on relevant Wikipedia articles. However, the previous work points out their detrimental effects as well. For instance, platforms expose the users to belief-reinforcing information to maximize engagement and effectively trap them in a bubble that the opposite opinion cannot pass through, which is named "filter bubbles". [43, 125, 213] Filter bubbles make people vulnerable to social media manipulation. This is because if a group of people cannot receive a particular opinion on an issue (e.g., vaccines are safe), they can easily be deceived by those who promote the counter and the potentially malicious narrative (e.g., vaccines are bad). The research on filter bubbles encouraged researchers to study mitigating their effects [122, 123, 124, 178, 294]. For instance, in our work, we propose recommending polarized topics to celebrities so that they can moderate the discussion and burst filter bubbles [97]. Such research encouraged social media platforms to acknowledge the presence of filter bubbles [139] and put effort into mitigating them [287].

Manipulation The fact that information can flow freely and quickly on social media, platforms' vulnerable mechanics and influence attract adversaries. They aim to exploit the platforms for malicious purposes using social media manipulation techniques. We now define and describe social media manipulation in detail.

2.3 Social Media Manipulation

In this section, we define and describe social media manipulation. Precisely, we present the typical threat model for social media manipulation, which consists of a message, a vulnerability, and a technique. We then describe and briefly survey these three elements.

Preliminaries We broadly define social media manipulation as "*employing a series of techniques that exploit the vulnerabilities of social media platforms to disrupt the public dialogue.*". The techniques may consist of sophisticated strategies and tools such as automation and can be employed for campaigns such as political propaganda [304]. In a typical social media manipulation threat model, an *adversary* attempts to bring a (potentially harmful) *message* to the public attention to have a detrimental *impact* the public dialogue. In doing so, they may use sophisticated *techniques* and exploit the *vulnerabilities* of the platform, i.e., attack the platform. In response, the social media platform can build *counter-measures* to prevent the attack (i.e., completely prevent it from being possible) or to mitigate it (i.e., reduce the likelihood or impact of the attack) [306]. The *impact* may be misleading the public on critical issues such as health. It may affect society as a whole, e.g., decreasing voter turnout. It may jeopardize the integrity of the platform and the company and result in a financial loss. It may also defame individuals or organizations if it is a targeted attack such as a smear campaign. The *adversary* may be a techno-savvy person, a social media agency that manages the reputation of its clients [102], or even the government actors [132], depending on the scale and the goals of the operation. We now provide two example threat models that follow this structure.

Example 1 Many social media platforms allow the usage of multiple accounts. They also publicly measure the approval or disapproval of posts by allowing accounts to vote for others (e.g., liking, retweeting, upvoting). As such, the same person can create multiple accounts to gain extra votes, which is a vulnerability. An adversary seeking to propagate offensive content to bully somebody may exploit this vulnerability. The impact may be the defamation of the target. Most social media platforms prevent this by disallowing dislikes. Youtube recently mitigated the issue by hiding the number of dislikes from the public [260]. Reddit allows and publicly displays downvotes but mitigates the attack by suspending accounts if they are used by the same person and vote the same posts, which it names "vote manipulation" [141].

Example 2 People may make mistakes while creating content on social media. They may want to edit their posts after creating them. Some social media platforms let users edit their content retrospectively. However, this feature can be abused. An adversary seeking to influence the public can compromise a popular account and stealthily change its old posts with malicious content, which is a vulnerability. The impact may be misleading the public into thinking that the account posted the content as it is. Those endorsing the original content may be victims of the edit if it comprises malicious content. Twitter prevents this attack by not allowing users to edit their posts. Facebook, however, lets users edit their posts but mitigates the problem by keeping track of the edits publicly. To the best of our knowledge, social media manipulation by retrospective edits has never been studied in the literature. However, in 2020, FireEye reported a similar attack in which Russian hackers compromised news websites in Poland, Lithuania, and Latvia and replaced existing news articles with fake ones [117]. This poses a

threat for the people citing or endorsing the original article if they did not retrospectively remove their reference after the attack.

We will now describe three elements of the typical threat model, messages, techniques, and vulnerabilities, in more detail.

2.3.1 Messages

The messages that the adversaries aim to spread may be any harmful content that disrupts the public dialogue. They may consist of disinformation to mislead the public, biased interpretation (frames) of the issues to influence the public or offensive content that inspires hate or incite violence. We now provide a brief overview of these types of messages.

Disinformation and Fake News Adversaries stealthily craft false information to mislead the public on purpose, which is named disinformation, or fake news if they mimic mainstream news [144]. As social media facilitates information spreading, adversaries use it as a breeding ground for fake news. They create "infodemic"s, which are the states of the undisciplined spread of disinformation, unverified information, and conspiracy theories [324]. We observed this especially during elections [112], riots [109] and the current pandemic [113] where coordinated groups [210] and/or automated accounts (bots) [283] have participated in malicious propaganda campaigns to influence public opinion through fake news. We also observe that adversaries disseminate fake news using fake trends and discuss it in detail in chapter 3.

Bias and Frames Social media manipulation can manifest itself not only by spreading disinformation and harmful narratives but also by consistently sharing biased information and framed narratives. For instance, some social media users consistently support only one political entity and indiscriminately criticize others on every subject, named "seminar users" [78]. Partisans frame issues aligned with the frames their favorite political party holds [194]. Coordinated groups can promote a specific frame for political gains and may even switch between them. For instance, the Istanbul Convention was an international treatment to protect women from domestic abuse. In our work, we found that coordinated groups of divorced men initially campaigned against the convention focusing on one-sided custody of children. However, they tactically re-framed their campaigns and cited religious motivations to gain broader support, and eventually succeeded as Turkey withdrew from the convention citing the same reasons [99].

Offensive Content Targeted messages that intend to harm others, such as bullying, spamming, hate speech, threats of violence, or sexual harassment, are considered offensive content [263] The lack of an editorial process makes this type of content more prevalent on social media when compared to traditional media even if the platforms take precautions. Thus, there is a growing body of research to detect, analyze and mitigate offensive content on social media. Here, we focus on hate speech because, in addition to individuals, it has the potential to affect society as a whole.

Hate speech is defined as "offensive discourse targeting a group or an individual based on inherent characteristics - such as race, religion or gender - and that may threaten social peace."[4] Researchers study hate speech on social media from different angles to better understand the phenomena. They found that hate speech target people's behavior (e.g., slow people), race (e.g., black people) and sexual orientation (e.g., gay people) [200] and religion [106] in some situations such as after extremist violence [207].

Adversaries may manipulate social media to disseminate hate speech to disrupt the public. For instance, Albadi et al. found that adversaries use bots to spread hate speech on polarizing topics such as Israel/Palestine and Yemen on Arabic Twittersphere [13]. In our work in Chapter 3, we observed that adversaries employ bots to push the slogan "Syrians Get Out" (*#SuriyelilerDefolsun*) to the top of Twitter trends in Turkey. The campaign was successful, and the slogan received the attention of the public, the media, and social science studies.

2.3.2 Vulnerabilities

Like every website and application, social media platforms may have security vulnerabilities that adversaries exploit for malicious goals. Since our focus is on using compromised accounts to manipulate social media, we provide a brief survey of both the vulnerabilities risking user security and the vulnerabilities risking civic integrity in this section. We also discuss their connection, which is how breaches to user security and compromised accounts create vulnerabilities on the platforms that impact the public.

Vulnerabilities Risking User Security Platform vulnerabilities may lead to adverse effects that put users' security and privacy at high risk, resulting in data breaches that expose user data such as emails and passwords. For instance, in March 2019, Facebook experienced a data breach due to storing passwords in plaintext in their internal servers [217]. Software bugs also may lead to abusing the platform's mechanisms for profit. For instance, Twitter used to allow posting tweets using SMS without an extra authentication. Adversaries took advantage of this and spoofed victims' phone numbers to send tweets on their behalf [2]. In 2010, a Turkish high-school student discovered a bug on Twitter. He was able to force people to follow him by simply tweeting "Accept" followed by the profile handle of the victim. The bug went viral, and those who exploited it forced their favorite celebrities to follow them. Twitter fixed the bug after this incident [175].

Vulnerabilities Risking Civic Integrity Adversaries manipulate social media to disrupt civic processes such as elections and harm society. As such, they may exploit vulnerabilities that may not have immediate adverse effects on the individual but may be exploited to manipulate the public, which is the main focus of this thesis. The most famous example of such vulnerability has led to Cambridge Analytica Scandal. In this incident, the adversaries developed an app that allowed them to collect their users' data

[4] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

who gave informed consent. However, the vulnerability of Facebook also allowed them to collect the data of the users' friends, who did not consent. Thus, the adversaries could collect the data of 87 million users by only getting consent from 270 thousand users. They used this data to model the psychological profile of users so that they could exploit their vulnerabilities through micro-targeting for political manipulation [231].

Platforms may not prevent some vulnerabilities if they are not aware of them or if they estimate that the likelihood of the exploit or its impact is low. Consequently, such vulnerabilities may occasionally be used for social media manipulation and influence the public. For instance, Facebook does not allow multiple accounts and only allows people to use their real names and pictures. However, a Facebook user can create and manage multiple "Pages" (e.g., a fan page for an artist). Pages' actions are limited, but they can comment on other pages' posts. State-sponsored actors from Azerbaijan exploited this feature and created multiple pages disguised as personal profiles (i.e., the pages' names were in name surname form, and they had profile images of people). They used these pages to comment on other pages in a coordinated manner to "astroturf", i.e., give the impression of public support for government narratives. Facebook removed these accounts and reported its strategy [204]. However, as of 2022, it is still possible to comment on other pages as a page, and the adversaries can still employ the same strategy. Our work in chapters 3 and 5 are analyses of such vulnerabilities. In chapter 3, we analyze a vulnerability that we discovered and disclosed to Twitter. They acknowledged the vulnerability but did not take precautions to prevent it. In chapter 5, we perform a large-scale vulnerability previously reported by researchers who work with Twitter but have not been studied in academic work. Twitter still did not fix this vulnerability as of 2022.

Vulnerabilities Posed By Compromised Accounts Compromising accounts, colloquially named "hacking", is to control or take over accounts. The attackers can compromise accounts by stealing their credentials or executing malicious scripts on the victims' devices [93]. Compromised accounts are both the impacts of security vulnerabilities of the platforms and also vulnerabilities themselves. This is because such accounts generally build trust relationships with the platform and the public before being compromised [93]. Adversaries may abuse this trust to manipulate social media. One of the recent and famous examples of such abuse is the hijacking of Twitter accounts in July 2020. In this event, attackers compromised 130 popular and verified accounts of people and organizations such as Elon Musk, Bill Gates, Joe Biden, Barack Obama, Binance, Apple, and Uber. The attackers posted tweets on their behalves, asking users to send bitcoin to a cryptocurrency wallet they designated, promising they would send the double amount in return. The attackers made 110,000$ before Twitter removed their tweets and blocked tweets from verified accounts as a temporary counter-measure. The adversaries carried out the attack by gaining access to administrative tools through social engineering (i.e., targeting Twitter employees to obtain their credentials) [155]. This thesis contributes to

the literature on the vulnerabilities posed by and the impact of compromised accounts by introducing their role in social media manipulation.

2.3.3 Techniques

The adversaries use certain techniques to exploit the platform vulnerabilities. In our work, we introduce two new techniques: ephemeral astroturfing attack (Chapter 3) and misleading repurposing (Chapter 5). These techniques use a combination of pre-existing techniques: compromising accounts, coordination, content deletion, managing automated accounts, and managing misleading accounts. Therefore, in this section, we focus on and provide a brief overview of those specific techniques.

Compromising Accounts Compromising accounts let the adversaries create or engage with social media content on legitimate users' behalves. It is a better alternative to creating new accounts from scratch due to the trust the compromised accounts built. As such, compromised accounts would be less likely to be suspended, may be used for longer, and may amplify the message more efficiently due to their popularity and reputation than a newly created account with no history. Adversaries can compromise accounts by obtaining their credentials. They can do it by phishing (e.g., through fake login pages disguised as legitimate websites), data breaches, or other hacking techniques (e.g., dictionary attacks). Some adversaries create Ponzi schemes: they promise free engagements to users in exchange for using their accounts to provide free engagements to newer users [256]. In our work, we also observe a Ponzi scheme where the adversaries provided free followers to legitimate Twitter accounts but exploited the users to create fake trends in exchange [102]. Adversaries can claim partial control of the account: they use the account parallel to the owner. They can also claim full control of the account and deny the original owner from using the account while using it as it is. Moreover, they can repurpose the account after the takeover: sell it to somebody else or keep it but change its identity or purpose. We study a case of partial control in Chapter 3, full control in Chapter 4, and repurposing in Chapter 5 in detail.

Coordination Adversaries employ coordination to scale the size of the manipulation. This is because if many users coordinate towards conveying a message, both the public and the algorithms may recognize the message as popular, credible, or a result of collective action [201]. Adversaries can do this by directly posting the same message from multiple accounts or indirectly amplifying the message by coordinated engagements. The social media algorithms may amplify such messages thinking that they may instigate even more engagements and increase the user screen time, resulting in more ad clicks [271]. For instance, adversaries use multiple accounts to artificially increase votes on Reddit threads of their choice to boost their visibility and suppress the visibility of other legitimate posts [49]. This is because Reddit ranks the posts by the number of votes, which may be because it assumes those posts keep the users continue scrolling on Reddit. More visibility also encourages other users to create more engaging content

for Reddit [230]. Employing multiple (potentially fake) personas (e.g., social media accounts) in a coordinated manner to give the impression of public support while masking the sponsor is named "astroturfing" [165]. We cover it in detail in Chapter 3.

Some users coordinate among themselves to boost each others' posts and have a win-win situation named "reciprocity abuse" [88, 298]. In our work, we focus more on the accounts controlled by adversaries, which are fake or compromised. These accounts may be automated and operate on a large scale (bots) or controlled by humans but are misleading (trolls). We now provide an overview of such accounts.

Managing Automated Accounts (Bots) Adversaries often use automated fake accounts controlled by software, which are named "bots". Bots can automatically perform simple interactions such as following, sharing, and posting in-genuine content (e.g., advertisement), also known as spam. Their advantage is that they can easily scale, which makes them efficient tools in social media manipulation. The bots can be used for different primary functions and categorized according to those functions. For instance, fake follower bots inflate follower counts and boost the perceived popularity of other Twitter users [68]. Clients can purchase bots' services from black markets to amplify their messages [129]. Some Twitters users, even though they are not bots themselves, use bots to send automated follows so that they can get follows in exchange and look popular [308]. Chapter 4 presents the first study that exclusively focuses on retweet bots that inflate retweet counts.

Managing Misleading Accounts (Trolls) Adversaries also employ accounts that humans manage but act by extrinsic motivations (i.e., sponsored) in a coordinated manner. Platforms define such accounts as exhibiting "coordinated inauthentic behavior." [203, 279] They are also colloquially named as "trolls". The main characteristic of such accounts is misleading the public about their true identity and goals. For instance, Twitter detected, removed, and published the data of 2700 accounts that originated from Russia and were doing propaganda during the 2016 U.S. elections. The accounts adopted American names, reported their locations in the United States, and shared news stories supporting Russian narratives in a coordinated manner [321]. The National Security Council later investigated the issue and stated there is high confidence that Russians interfered with the 2016 U.S. elections. According to them, Russia aimed to erode public confidence in the U.S. democratic process and to help Donald Trump win the elections [64]. Such accounts can be combined with other techniques such as compromising accounts, which we cover in detail in chapter 5.

Malicious Deletions Users can retrospectively delete their social media content and engagements for various reasons, such as feeling remorse. The adversaries can abuse this feature and use it as a technique to hide their malicious activity. Collecting deleted data is challenging because platforms do not provide such content. Thus, social media manipulation employing deletions is currently understudied. Our work which we cover in chapter 3 in detail, is one of the first case studies on how deletions could be employed

to manipulate social media. We found that adversaries employ bots in a coordinated manner to push slogans to the top of Twitter trends. Twitter recognizes that slogan as a popular slogan and thus, displays it as a "trending topic". However, the trending algorithm does not take the deletions into account. Consequently, adversaries delete the bots' tweets and hide the inauthentic nature of the trending topic [102]. Torres et al. found that accounts involving in "follow trains" (aggressively promote users to involve in reciprocal followings) frequently delete their posts to hide their malicious activity [269]. Adversaries also use deletions to bypass Twitter's 2400 tweets per day limit [268] and to repurpose accounts [101, 321].

2.4 Detection

This section provides an overview of methods to detect social media manipulation. Detection is the first step to discovering the harmful content being propagated and its authors. Both the platforms and the researchers implement systems to detect social media manipulations or the accounts involved. The mechanisms depend on the data accessibility, scalability, and the specific goals of the detection. Thus, researchers and platforms may have different detection mechanisms. Platforms generally do not disclose their detection mechanisms as the adversaries may circumvent them. Thus, we only focus on detection mechanisms that the researchers disclose.

2.4.1 Detecting The Messages

Adversaries often disseminate potentially harmful messages when they manipulate social media. The content of the message can consist of disinformation and fake news, hate speech, and other types of offensive content and can be in the form of text, short text (e.g., tweets), images, and videos. Researchers employ different types of techniques to detect different types of content. We now briefly survey the types of contents and proposed detection techniques.

Detecting Fake News Fake news on social media has three components: the source, the content of the article, and the social media context [247]. Fake news detection techniques usually employ features extracted from a combination of these components. The source of the news aids classification as some sources are less credible, i.e., they have a reputation for producing low-quality content [248]. Linguistic features extracted from the content facilitate the detection of fake news in the form of text. These features may reflect the writing style of the article, such as the usage of assertive and factive verbs (e.g., "claim", "indicate") that capture the degree of certainty, report verbs (e.g., "deny) that emphasize the attitude towards the source of information and n-grams that signify subjectivity and bias [222]. Recent studies suggest that more advanced features such as state-of-the-art text representation models such as BERT improve fake news detection [82, 156, 228]. Visual features are also helpful in detecting fake news with visual components [48].

The third component, the social media context, can distinguish fake news by the adversarial behavior behind them. Adversaries employ malicious accounts to propagate the news by posting or sharing (e.g., retweeting) them. Those accounts may give away the suspicious nature of the news they share. Individual-level features (i.e., metadata of the user such as follower counts) [51], and group-level features [310] such as aggregated statistics of individual-level features [172, 187] and network features extracted from the community structure of the accounts facilitate the detection [330]. The stance of the account and the users engaging with their posts are also reliable indicators [250]

Detecting Bias Detecting bias enhances the analysis of social media manipulation as it helps to identify the ideology of the adversaries. Researchers study bias in terms of a strong positive stance towards a particular side in a controversial topic [184] or towards a political party or ideology [173]. Studies showed that both the textual features and the network features are effective in detecting biased. Darwish et al. propose an unsupervised approach to detect user stances on controversial topics by clustering on the user features such as retweets and hashtags [79]. Barbara et al. use Bayesian Ideal Point Estimation to learn users' political ideology from their followings [30]. Mendelsohn et al. use the latter approach to determine the ideology of the users who frame the immigrants as a threat to public order and found that they are generally conservative users in the context of U.S. [194]. Luceri et al. propose their own political ideology detection based on biased media outlets shared by the user. They then analyzed the differences between bots aligned with different parties [185]. In our work in chapter 3, we also analyze the bias of astroturfing bots by the stance of trends they promote and found that they do not have a strong bias towards a particular political party or ideology. This implies that creating fake trends was a business model.

Detecting Hate Speech Platforms detect hate speech to remove them while the researchers detect them to discover their targets, the authors, and their goals. They may also monitor hate speech and intervene when necessary [138]. Past studies tackled the problem using textual features. ElSherief et al. employ key phrases and hashtags to detect hate speech on social media [106]. Ruwandika et al. propose a supervised learning approach using Naive Bayes on text data modeled using Tf-idf [236]. Vidgen et al. use a combination of textual features such as word embeddings, presence of part of speech "conjunction", and entities depicting locations and organizations after running named entity recognition on the text to detect Islamophobic hate speech on social media [291]. Zannettou et al. state that detecting hate speech is still an open problem, and there is no general classifier that detects all kinds of hate speech due to the subjectivity of the definition of the term [323].

2.4.2 Detecting The Techniques

Detecting Coordination Platforms such as Facebook and Twitter consider coordination a necessary condition for platform manipulation and detect the accounts using

their coordination patterns. Since these mechanisms are black-box, the researchers propose their own methods to detect and study coordination. They use similar temporal activity [55] and common content [7] among the accounts as features. Different feature combinations can be helpful to detect different types of coordination such as account handle sharing, image coordination etc [211]. Detecting coordination helps researchers analyze the propaganda activities by adversaries and how they employ coordination for such activities during elections [147]. In our work, we detect coordination using anomaly detection on the temporal activity of the bots, which we cover in detail in chapter 3. Detecting coordination is closely related to detecting bots and trolls, which we will review now.

Detecting Bots Detecting bots is crucial to keeping the platform integrity. This is first because bots inflate the number of active users on the platform, which may make the business partners lose trust in the platform. More importantly, the malicious bots disrupt the public dialogue. Consequently, researchers propose detection methods to monitor the bots and understand their goals. They also detect them to clean the social media data from their posts when they study the public dialogue. Researchers detect bots exploiting their coordination and inauthentic behavior. While the former alone is effective in discovering a group of accounts, they may not predict the likelihood of being a bot for a given user. Thus, researchers propose per-user supervised classification methods that train on a known set of bots and humans. They use features based on the profile or tweet statistics [176, 197], content [299], network properties [15] or a combination of them [92, 170, 240] In chapter 3, we detect bots both by exploiting their malicious content deletion pattern and coordination with the other bots. Furthermore, in chapter 4, we analyze the characteristics of bots that were compromised accounts and their implications for such bot detection systems.

Detecting Trolls Detecting trolls is more challenging than detecting bots as there is no clear definition of trolls and the ground truth data is limited. Platforms do not explicitly state that they detect trolls. They rather frame them as accounts involved in "coordinated inauthentic activity" [203] and "coordinated harmful activity" [280]. The researchers colloquially name such accounts as trolls. They usually build detection systems to detect the type of accounts disclosed by the platforms. Addawood et al. propose to detect Russian trolls on Twitter using linguistic cues [8]. Luceri et al. propose to use the accounts' engagements [186]. Saeed et al. propose TrollMagnifier, which predicts the likelihood of an account being a Russian troll based on their profile statistics and their interactions with the known Russian trolls on Reddit [237]. Some trolls repurpose old accounts to do propagadanda [321] for which we propose a detection system in Chapter 5.

Detecting Compromised Accounts Platforms can detect compromised accounts to return them to their original owners or suspend them if the former is not possible [132].

Researchers may also benefit from detecting compromised accounts so that they can analyze how the adversaries use them. Several studies tackled the problem of compromised account detection using different approaches but with similar assumptions. Egele et al. propose to detect large-scale compromises and isolated high-profile compromises using statistical modeling and anomaly detection, assuming that the compromised accounts will observe a sudden shift in behavior [94]. Kaur et al. propose an authorship verification approach by measuring the textual similarity between two sets of social media posts created at different times. They assume that the compromised accounts will have a set of spam posts that will differ from the genuine posts [162]. Karimi et al. propose a classification methodology that extracts users' temporal behavior, textual content, and the social network. They consider an account compromised if it contains posts published by a hacker, in addition to posts by the legitimate owner and an announcement that the account was hacked [161]. All those studies assume that the hackers will retain the genuine tweets by the legitimate owner and use the accounts to post spam and observe other types of anomalous behavior. Some also assume that the original owner will regain access to the account. In our work in chapter 5, we tackle the problem where those assumptions do not hold: that the hacker may remove the past data and repurpose the account to use it like a legitimate user and never gives it back. We name behavior "misleading repurposing" and cover it in chapter 5 in more detail.

2.5 Counter-Measures

The platforms or researchers build counter-measures to prevent or mitigate the vulnerabilities and fight against social media manipulation. This can manifest itself in new policies and their enforcement. Increased transparency also mitigates the issue by raising public awareness. In addition to detection, platforms and researchers build technical solutions to counter or mitigate the impact of manipulations. We briefly survey these counter-measures in this section.

2.5.1 Policies

Platforms often state what is disallowed under their terms of service and community guidelines as a ground to moderate content on their platform. They define what constitutes platform manipulation and other abusive behavior.

The rise of disinformation and other harmful activity on social media encouraged platforms to implement new policies. For instance, Twitter used to allow people to coordinate and take collective action on their platform to make their voices heard, especially during crises like Arab Spring [271]. However, in 2020, people used Twitter to coordinate among themselves to spread conspiracy theories named "QAnon". Since those people might not be bots or state-sponsored trolls, their malicious behavior did not violate Twitter's terms of service. As a response, Twitter defined "coordinated harmful activity" and forbid people to coordinate among themselves to spread disinformation or

other kinds of harmful narratives. They then suspended those who spread the conspiracy theories on their platform [281].

However, platforms may have disparities in their policies, which adversaries may exploit to prefer one platform over another to manipulate. For instance, Facebook disallows significant changes to Page names and their subject matters, and Twitter has no such policy. Thus, adversaries can repurpose Twitter accounts and mislead the public about the past of the page, which we name "misleading repurposing." We cover this technique in chapter 5 in detail.

Governments also imply new policies and enforce platforms to comply with them. For instance, California put a new law into effect which requires "bots" to disclose their automated nature [63]. Correspondingly, Twitter launched "automation" labels and enforced bot accounts to self-state that they are automated and identify their developers [275]. In 2016, European Union put "General Data Protection Regulation" (GDPR) into effect. It enhances the users' control of their own social media data [284]. This instigated Facebook to facilitate users' download of their complete Facebook data using the "Download Your Information" tool, accessible through settings [195].

Some platforms may not fully comply with governments' laws and do not remove the posts and users on demand. They instead withheld (i.e., censored) the content from the users connecting from the state, which sends a legal request. We observed that the majority of the withheld users on Twitter are censored in Turkey, Russia, Germany, France, and India [98].

In some cases, platforms implement policies not to outright ban a type of content but take precautions to limit it or its spread. For instance, in 2019, Whatsapp began limiting the users to forward a message up to five times at once to counter the mass-forwarding of disinformation. Melo et al. performed a quantitative study to evaluate the solution's effectiveness and found that it is ineffective in blocking the propagation of misinformation campaigns in public groups [119].

2.5.2 Enforcement

Platforms enforce their terms of service through content moderation. They first detect violations using detection methods and/or human-based moderation such as freelancers [232], volunteers [179] in addition to reports by the users [67]. They then remove violating posts or suspend users who do not comply with their terms of service, which is named hard moderation [249].

Recently, the platforms have been criticized for acting as authorities who censor others, e.g., the users whose messages are of public interest, such as the U.S. president [16]. Therefore, platforms start to turn on what is called *soft moderation* [249]. For instance, Facebook, Twitter, Instagram, and TikTok puts labels on covid-related posts and redirect the user to official sources to get the most reliable information [180], instead of fact-checking and removing the posts right away. Twitter labels state-affiliated actors and news outlets [11]. They also put warnings on tweets that contain disinformation, e.g., Trump's tweets claiming election fraud [320]. Reddit quarantine subreddits that

promote what the platform classifies as "hoax" and display a warning upon entering them instead of banning the whole community [229].

In addition to warnings that increase user awareness, platforms also limit the activity and experience of some users. For instance, Twitter limits amplification to state-affiliated actors and media, in addition to labels [272]. They also apply quality filters to accounts with low-quality content (automated or spam) and decrease their visibility, e.g., their replies are only visible if the user clicks on "Show additional replies" [273]. Facebook, Google, Amazon, Spotify, and TikTok banned political ads on their platform during elections to prevent the spread of misinformation [86], while Twitter banned them indefinitely [274].

Platforms also proactively filter or remove the content immediately. For instance, Twitter keeps a dictionary of words they do not allow to be on the trending topics list, such as words related to porn [14]. Platforms detected and removed the videos of the Christchurch shooting immediately, although the adversaries attempted to avoid detection by distorting the video [66]. Our knowledge of such proactive measures by the platforms is currently limited due to having no access to data and ethical reasons.

2.5.3 Transparency

Social media manipulation became a weapon of state-sponsored actors who seek to interfere with other countries. This sparked interest in the public, the academy, and policymakers about the manipulative activity of foreign powers on social media platforms. Consequently, the platforms became more transparent about the manipulations from which they suffered. For instance, after the 2016 U.S. elections, Twitter initiated the Civic Integrity project, which provides increased transparency on "Information Operations". They provide public data on state-sponsored actors who manipulate the platform and report their strategies by collaborating with research institutions and universities [276]. Those reports enhance our understanding of social media manipulation techniques. For instance, Grossman et al. [132] analyzed state-sponsored actors published by Twitter and reported that they use a technique which we name "misleading repurposing". However, they did not extensively study it but instead motivated our study. The data Twitter provided became the ground truth in our work on misleading repurposing, which we cover in chapter 5 in detail.

2.5.4 Technical Solutions

Platforms build their own solutions to counter social media manipulation. They generally do not disclose them but briefly inform the public when they are under scrutiny. For instance, in 2020, Facebook disclosed that they deployed a model named SimSearchNet++ to match near-duplications of images that contain misinformation. They also state that they use LASER to compute the semantic similarity of the texts in images. While they published their methodology and results on LASER and made the model available to the public, they did not disclose SimSearchNet++ [145].

Although platforms take precautions to fight social media manipulation, their efforts are limited. This encourages researchers to build their own solutions to defend against social media manipulation. Some propose tools to monitor disinformation. For instance, Shao et al. propose Hoaxy, a platform that helps users track online misinformation by simultaneously monitoring news data from social media, news, and fact-checking websites [243]. Similarly, Shu et al. propose FakeNewsTracker [246] which detects fake news and visualizes its content using word clouds. We also create a Twitter bot to track fake trends on Twitter which we cover in chapter 3 in detail.

Some tools raise awareness for the authenticated users of the tool on their vulnerabilities. For example, WDTKAM [130] shows its users the personal information they disclose on the web through their posts on Twitter. Gao et al. [122] proposes a tool to show its users their own biases in order to mitigate selective exposure and burst filter bubbles. Others provide information about suspicious profiles. For instance, Botometer predicts the automation probability of a given user [80]. Birdspotter predicts the "botnets" of an account and its influence [225]. Evently [167] visualizes how users' content spreads online to show that information propagation works differently for bots and authentic users. We also introduce our own tool "WayPop" to help researchers root out accounts that are repurposed on Twitter in chapter 5.

Researchers also propose solutions for users that do not rely on the usage of external tools. For instance, Minaei et al. propose an approach based on obfuscation using noise injection. They defend against the adversaries who hunt for deleted social media posts that may be violating the users' privacy by injecting decoy deletions on the victims' behalf [198]. Zhang et al. [326] propose a data poisoning attack to defend users against unsolicited Twitter post recommendations "You might like", which may be amplified due to social media manipulation. Our Twitter bot in chapter 3 also provides a non-tool-dependent solution as it announces the fake trends on Twitter itself.

2.6 Twitter for Social Media Studies

Twitter is a social media platform where people socialize through microblogs limited to 280 characters called "tweets". Twitter has unique characteristics that make it different from other popular platforms such as Facebook and Linkedin. Firstly, Twitter social networks can be unidirectional: users follow other users to subscribe to their updates, while the latter may not do the same. Secondly, unlike Facebook, Twitter accounts are open to the public by default, meaning anybody may see their posts. Users can make their accounts private, but this heavily curbs their ability to socialize and thus, limits their Twitter experience. Thirdly, Twitter allows (and even encourages by services like TweetDeck) the usage of multiple accounts, anonymous accounts, and automated accounts. These features make Twitter a platform where people easily connect to and interact with each other and express their opinions freely since they do not have to reveal their identity. They also enhance the spread of information.

Studies showed evidence of social media manipulation in other social media platforms such as Reddit [224], Tiktok [168] and Youtube [177], and even multiple platforms at the same time [302]. However, most of the previous work focused on Twitter. This is first because Twitter facilitates data collection and analysis through official APIs and tutorials. Secondly, most of the Twitter accounts are public. A recent study found that only 4% of active accounts were private [163]. Twitter allows data collection from public accounts. Thirdly, the fact that the platform allows users of automated, anonymous, and multiple accounts may make it more vulnerable to manipulations. Consequently, many studies, such as ours, use Twitter to study social media and its manipulation. Tufekci criticizes the dominance of Twitter in social media studies, stating that Twitter's mechanisms may not translate to other platforms [270]. We acknowledge this limitation and discuss how findings may generalize to other platforms in our work.

Chapter 3

Analysis of Attacks Employing Compromised Accounts: Ephemeral Astroturfing

Ephemeral Astroturfing Attacks:
The Case of Fake Twitter Trends

Euro S&P 2021

This chapter presents an *analysis* of an attack that we uncovered, which employs compromised social media accounts: ephemeral astroturfing attack. The attack manipulates social media by targeting the popularity mechanisms of social media platforms. In this attack, a chosen keyword or topic is artificially promoted by coordinated and inauthentic activity to appear popular, and, crucially, this activity is removed as part of the attack. We observe such attacks on Twitter trends and find that these attacks are not only successful but also pervasive. We detected over 19,000 unique fake trends promoted by over 108,000 accounts, including not only fake but also compromised accounts, many of which remained active and continued participating in the attacks. Trends astroturfed by these attacks account for at least 20% of the top 10 global trends. Ephemeral astroturfing threatens the integrity of popularity mechanisms on social media platforms and by extension the integrity of the platforms.

3.1 Introduction

Mechanisms deployed by social media platforms to display popular content are a primary vector by which platforms increase engagement. Facebook's newsfeed algorithm; Reddit's "r/popular"; and Twitter's trending topics, "trends," are integral to both platform functionality and the underlying business model. These mechanisms are valuable because they determine which content is most visible to users. Twitter's *trends* can be equated to traditional advertising channels and can be useful for marketing [50], as

Twitter acknowledges by charging companies to promote their brands on trends for a day [115].

The integrity of such popularity mechanisms is integral to the social media ecosystem. Users expect that the popular content they are shown is the result of authentic activity on the platform, legitimate grassroots campaigns expect that their content will be fairly considered, and the platform expects that showing popular content increases engagement. Further downstream, advertisers expect that popularity mechanisms behave in a way to increase engagement and therefore revenue. Even further, those who use trends to study society and social media, i.e. researchers and journalists, expect that trends accurately reflect popular themes that are discussed by the public.

Since these popularity mechanisms carry so much influence and potential for revenue, they are an attractive target for adversaries who want their illicit content to be seen by many users. For instance, "like farms" are used to generate fake likes on Facebook to boost posts to the top of users' news feeds [75], and bots can be used on Reddit to artificially "upvote" posts to increase their visibility [49]. In the case of Twitter trends, adversaries, sometimes from abroad [220], boost disinformation and conspiracy theories to make them trend so that they are further amplified [4], as in the case of QAnon followers hijacking the trend #SaveTheChildren [233]. Due to this incident, many called on Twitter to stop curating trends by using the hashtag #UntrendOctober [62].

Attacks on popularity mechanisms rely on making inauthentic content or actions appear organic. Borrowing terminology used to refer to campaigns that fake grassroots organizing on social media, we call them "astroturfing" attacks. Once exposed, astroturfing attacks erode user trust in the platform. Gaining an understanding of these attacks is a crucial part of keeping the platforms safe for users and valuable for advertisers, thus preserving the business model of the platforms.

In this chapter, we provide an in-depth analysis of a new type of astroturfing attack that remains unstudied in the academic literature which we call *ephemeral astroturfing*. Ephemeral astroturfing differs from traditional astroturfing in that the actors hide the malicious activity while successfully executing an astroturfing attack, paradoxically aiming to make something more visible while making the content responsible for the visibility invisible. By removing any evidence of the attack, ephemeral astroturfing outperforms other approaches in three key ways: (i) it enables the use of active, *compromised* accounts as sources of fake interactions, accelerating the popularity; (ii) it evades detection by users, the platform, and academic studies; and (iii) it prevents users from reporting the malicious activity as spam, so traditional spam classifiers are unable to prevent future attacks.

We focus on fake Twitter trends as a case study to investigate ephemeral astroturfing attacks. Twitter is a popular platform for many critical discussions, including political debates, with appropriate data available to study: Twitter provides both deletion notices and trends through its official APIs. We observe that Twitter trends suffer from ephemeral astroturfing attacks both in Turkish local trends, affecting Turkey's 11.8

million active users, and global trends. Precisely, we find that ephemeral astroturfed attacks on Twitter trends started in 2015 and accounted for at least 47% of the top-5 daily trends in Turkey and at least 20% of the top 10 global trends. We find that Twitter does not consider whether a tweet has been deleted when determining which keywords should trend and thus is vulnerable to ephemeral attacks.

Ephemeral astroturfing is enabled by the current design of the algorithm that determines Twitter trends. Trends are refreshed every 5 minutes, taking as input tweets that have been published in some time interval. However, despite the importance of the integrity of the list of trends, the algorithm *does not check* whether those tweets are still available or have been deleted. This vulnerability can be expressed as a sort of Time-of-Check-Time-of-Use (TOCTOU) attack, by which at the moment that the data is "used" to determine a trend, it is different than when it was "checked" because it is deleted. In other words, this attack exploits a violation of the complete mediation principle when using security-critical inputs (tweets) to update a key asset for the platform.

Due to the severity of the attack, we notified Twitter (once in July 2019 and again in June 2020) and provided a detailed description of the attack and the accounts involved. After the first notification they acknowledged that the attacks do exist (July 2019), and after the second notification (June 2020) they replied that they would forward them to the relevant team to address. We have followed up since, but have not received any indication that they are progressing. The attacks on Twitter trends continue as of July 2022.

In summary, our contributions are the following:

1. We introduce and present an analysis of a social media manipulation technique that employs *compromised* accounts, which is novel.

2. We define and describe a new type of attack on the popularity mechanisms: *ephemeral astroturfing* (§3.3).

3. We uncover ephemeral astroturfing on Twitter trends as it occurs in-the-wild. We find that it has been ongoing since 2015 and that it has a strong influence on local trends i.e., we find more than 19,000 unique keywords that are the result of ephemeral astroturfing attacks (§3.4) which employed at least 108,000 bots; and on global trends, i.e., we find that at least 20% of the popular global trends during our study were the result of ephemeral astroturfing (§3.5). Our study is the *first large-scale analysis* of fake trends.

4. We study the ecosystem behind ephemeral astroturfing attacks on Twitter trends. We find that they rely on a mix of bots and compromised accounts (§3.6). We also find that there is a business model built around the attacks in (§3.7).

5. We discuss the implications on platform security and society, propose countermeasures, and identify barriers to deploying defenses in practice. (§3.8).

3.2 Related Work

3.2.1 Social Media Manipulation

The wide adoption of social media platforms has attracted adversaries aiming to manipulate users on a large scale for their own purposes. Such manipulation attacks span from targeted advertising assisted by mass data collection [45] to state-sponsored trolling [206], propaganda [264], spam [60, 131, 313, 325], popularity inflation [68], and hashtag hijacking [285]. Many of these manipulation attacks employ bots and bot-nets to execute since wide deployment is often a necessary component. We focus on this class of bot-assisted manipulation attacks. In our study, we observed political propaganda (not necessarily pro-government) and illicit advertisements that manifest themselves not through hashtag hijacking, as is often the case as well, but through direct trend manipulation.

Bots are becoming increasingly difficult to identify manually [118, 295] or automatically [72, 74]. Social bots are designed to mimic human users on social media [303]; they copy real identities (personal pictures, tweets), mimic the circadian rhythm of humans, gain followers by following each other, and mix malicious and hand-crafted tweets [309]. CyboHuman bots [59], *cyborgs*, humans assisted by bots [253], and *augmented humans* mix automation and human activity. In some cases, users register their accounts with malicious apps that make them part of a botnet. Ephemeral astroturfing attacks allow attackers to employ compromised users who continue using the account in parallel with the attackers, similar to [223]. Attackers hide from the legitimate user by deleting the attack tweets. Since these are otherwise benign users, they are likely to confuse supervised methods due to their dissimilarities to traditional bots. They would also confuse graph-based detection systems such as [153, 314] since they connect with other benign accounts. Although they are compromised, compromised account detection systems [31, 95, 96, 319] are not able to detect them if they do not account for deletions since the tweets that disclose compromisation are deleted.

Existing bot detection methods fall short of detecting the bot behavior we describe here as they rarely consider content deletion. Botometer [81] works on a snapshot of a profile and not on real-time activity, so it cannot detect the bot-like activity of accounts analyzed in this study since such activity is deleted quickly. Recently, Varol et al. [311] used content deletion as a bot feature but used a proxy to capture deletions: a high recent tweeting rate but a low number of tweets. This may capture the deletion of old tweets but not tweets deleted quickly. Debot [55] is based on keyword filtering by Twitter's Streaming, which does not give deletion notices and would not collect the relevant data if the attacked keyword is not be provided before the attacks, which is not possible for the keywords which trend only once. Chavoshi et al. [56] discovered a set of Turkish bots with correlated deletion activity to hide bot-like behavior. However, this study did not uncover whether these deletions were part of the astroturfing attacks we describe here. In our work, we classify the bot-created fake trends using their characteristic behavior: deletions and the generated content.

3.2.2 Astroturfing and Fake Trends

Although astroturfing by attacking trends using bots and manufacturing fake trends has been briefly reported on by the news media in both Saudi Arabia [6] and Turkey [17], it remains understudied in the academic literature. To the best of our knowledge, this work is the first to systematically study the mechanics of manipulating the Twitter trends feature on a large scale.

While not directly concerning the trending mechanism, previous works have analyzed *campaigns* that are artificially promoted [112, 164, 226, 290] or found evidence of manipulation of popular topics that are also trending by studying suspended and/or fake accounts and the overall temporal activity [154, 327]. They stopped short of studying malicious activity before keywords' reaching trends lists. In our work, we study the adversarial behavior that aims to push certain keywords to trends list directly, the behavior to evade the detection, and the accounts that are used for such operation.

3.2.3 Attack Detection

An ephemeral astroturfing attack is essentially a signal with an anomaly. Thus, detecting ephemeral astroturfing attacks is similar to tasks such as anomaly detection, outlier detection, event detection, and bursty keyword detection. We survey methods proposed for these tasks and test some of them in this chapter. Anomaly detection can be performed by modeling the regular activity. Breunig et al. [42] adopts this approach and proposes Local Outlier Factor which computes the degree of isolation of data points with respect to their neighborhood. On the other hand, Liu et al. [181] propose Isolation Forest which isolates anomalies without modeling normal data points. More complex methods integrate network activity into temporal activity for better performance. For instance, Miz et al. [199] propose anomaly detection to detect localized increases in temporal activity in a cluster of nodes.

Bursty keyword detection is a special case of anomaly detection that focuses on keywords of interest with sudden attention shifts within a period. They could represent new events such as disasters [262]. Their detection relies on techniques similar to anomaly detection. For instance, Guzman et al. [136] propose a fast and scalable online method that uses window variation on signals that represent word frequency. Data mining methods such as Apriori [10], Eclat [315], FP growth [137] may also assist bursty keyword detection.

Event detection generally consists of identifying bursty keywords to later group them together to represent events. A common and simple method is to use the relative popularity of a keyword during a time span when compared to its overall popularity [5, 196, 332]. Probabilistic methods include computing the magnitude of difference between the expected and the actual distribution of the volume of tweets [134] and computing the fitness of the distribution to exponential distribution [238]. Signal processing methods are also proven to be useful for bursty keyword detection and/or event detection. Zhao et al. [329] performs peak detection using an adaptive sliding window. Weng et al. [300]

propose EDCoW, which represents the word occurrence per time as signals using Discrete Wavelet Transformation.

3.3 Ephemeral Astroturfing

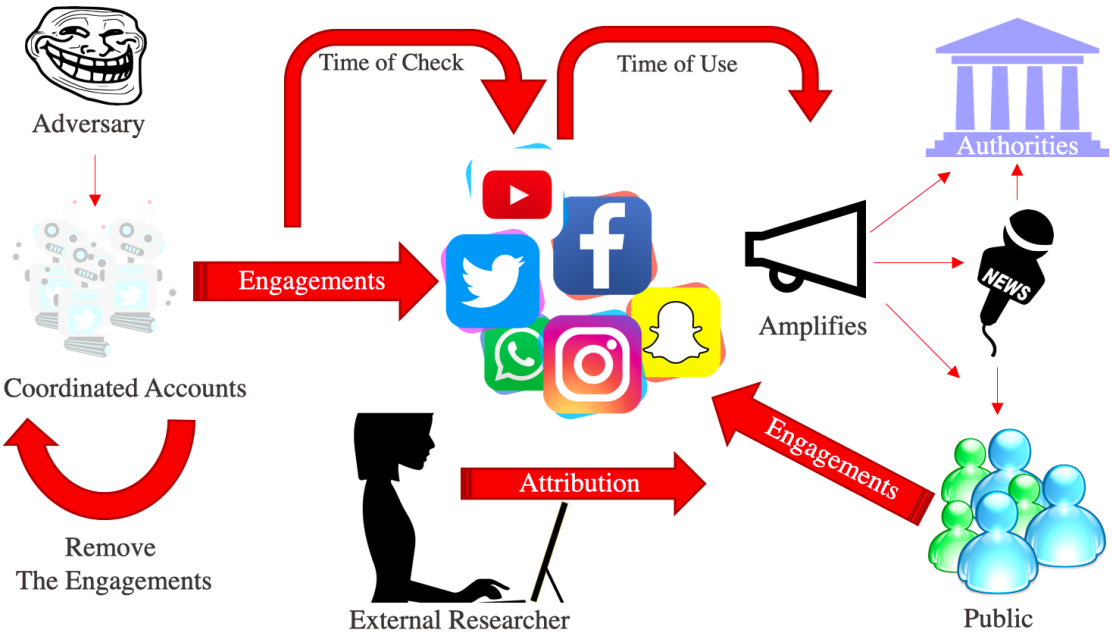

Figure 3.1: Summary of ephemeral astroturfing attack.

Attack Summary The goal of this attack is clear: make content popular through platform amplification. To do so, an adversary employs *coordinated accounts* to create *fake engagements*, which is an astroturfing attack. The social media platform *checks* for the engagements and recognize the content as popular and *amplify* it, e.g., show it on its main page as popular. However, the adversary *removes the engagements* to remain stealthy. They do it right before or when the platform uses the content to amplify as popular. The platform does not check for the engagements at that moment. Since the content is already amplified, the public may pick it up and create more engagements, which would keep the content popular. In this case, the platform continues to amplify the content and the public continues to discuss it, creating a *feedback loop*. In a successful attack, the news media may also pick up the content and may even bring it to the attention of authorities. An external researcher studying the content may *attribute* it to the public.Fig. 3.1 depicts the summary of the attack.

In theory, any social media platform with a popularity mechanism that has a period between the time of check and time of use and/or a period between consecutive checks may be vulnerable to ephemeral astroturfing. However, in this chapter, we focus on Twitter because it facilitates collecting popular content it amplifies and the engagements

promoting them. Additionally, we observe ongoing ephemeral astroturfing attacks on Twitter trends.

On Twitter, the popular content translates to a *target keyword* reaching the *trends lists*. An ephemeral astroturfing attack is executed by a number of accounts that are controlled by a single entity, which we refer to as *astrobots*. Each astrobot creates a tweet at roughly the same time. After a short period, these tweets are deleted. Alongside the target keyword, each tweet contains some pre-made or generated content that is enough to pass the spam filters of the platform (but not necessarily the Turing test). After an attack that renders a keyword trending successfully, other users adopt the new trend and post tweets that are not deleted. Fig. 3.2 shows the tweeting and deletion patterns of different astroturfing attacks with distinct non-adversarial behavior patterns.

Basis for the Model To model ephemeral astroturfing on Twitter trends, we look to the case of an attack that we observed in the wild that has been targeting Twitter trends in Turkey. To understand the attack, we created a honeypot account and signed it up for a free follower scheme that phishes users' credentials. We suspected that this scheme was being used to compromise accounts for ephemeral astroturfing attacks because the scheme was being advertised via ephemeral astroturfing. Our suspicions were confirmed when our account began tweeting and quickly deleting content containing keywords of about-to-be trends. Precisely, our astrobot account tweeted 563 times in 6 months before we exited the scheme. We now describe ephemeral astroturfing attacks on Twitter based on our observations.

Fig. 3.2 shows the tweeting and deletion patterns of different astroturfing attacks with distinct non-adversarial behavior patterns.

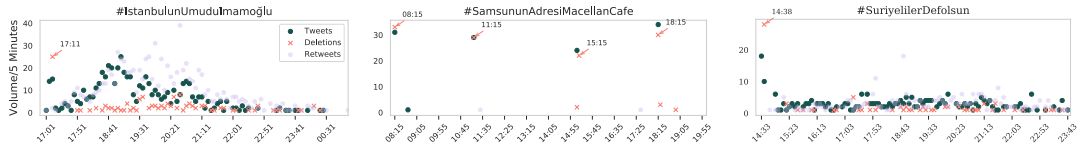

Figure 3.2: *#İstanbulunUmuduİmamoğlu* is a slogan associated with a candidate in the 2019 Istanbul election rerun. Note that although the hashtag is astroturfed by an attack initially (at 17:11), it was later adopted by popular users who got many retweets and drew the attention of the wider public. *#SamsununAdresiMacellanCafe* is an advertisement for a cafe, astroturfed to be seen in trends in Turkey. The hashtag did not receive attention from anyone other than astrobots: there are only coordinated tweets and deletions. *#SuriyelilerDefolsun* is a derogative slogan meaning "Syrians Get Out!". The hashtag grabbed the attention of the wider public due to its negative sentiment and sparked controversy in Turkey despite being astroturfed.

Attack Model

Let w be the target keyword. Let a set of posts that contain w be $T = \{t_0, t_1, ...t_n\}$, with creation time $p_{t_i} \in \mathcal{P} = \{p_0, p_1, ...p_n\}$, deletion time $d_{t_i} \in \mathcal{D} = \{d_0, d_1, ...d_n\}$. An attack $\mathcal{A}$ occurs when there is a T s.t.

1. Many posts: $|T| > \kappa$: at least κ posts involved,

2. Correlated Posting: $max(\mathcal{P}) - min(\mathcal{P}) < \alpha_p$: the posts are created within a window of size α_p,

3. Inauthentic Content: each post is comprised of w and a premade or generated content c that will pass the platform's spam filters,

4. Correlated Deletions: $max(\mathcal{D}) - min(\mathcal{D}) < \alpha_d$: the posts are deleted within a window of size α_d,

5. Quick Deletions: $d_{t_i} - p_{t_i} < \theta \quad \forall t_i \in T$: all posts are deleted within θ.

We leave the parameters (κ,α_p,α_d,θ, c) in the definition unspecified and later infer concrete values based on the instances of the attack that we detect.

To simulate trending behavior and confuse the algorithm which computes how popular the target keyword is, the attackers create many correlated posts in a short time window (rules 1 and 2). Any type of coordinated and/or bot activity has to pass the spam filters to evade detection and also to be considered in the platforms' metrics for popularity. These attacks are too large and coordinated to be executed at scale with handcrafted content, so the content must be pre-made or generated by an algorithm and therefore exhibit patterns in their content (rule 3). While recent advances in generating meaningful text make it more difficult for humans to spot such patterns, these advances have not reached the point of being able to create short texts related to a keyword for which it has no training data. Additionally, such arrangements are costly. These three points are common to all astroturfing attacks.

The ephemerality is captured by rules 4 and 5 in the attack model. Both appear to be the result of the attackers' tendency to quickly hide any trace of their attack from the public and the compromised accounts they employ. Additionally, deletions create a clean slate when users click on a trend, i.e., there will be no posts associated with the keyword when someone clicks on it on Twitter's trends list, so the attackers can post new content and be the first in the search results of the target keyword.

3.4 The Case of Fake Twitter Trends in Turkey

While astroturfing attacks are a global problem, we observe ephemeral astroturfing on a large scale in local trends in Turkey. Turkey has the 5$^{\text{th}}$ highest number of monthly Twitter users and a highly polarized political situation [171, 192]. The Turkish mainstream media has occasionally reported about the prevalence of fake trends there [17, 29, 77, 242], primarily sourced through interviews with attackers who manifest themselves as social media agencies. These agencies can be found via a simple Google Search for trend topic services and even advertise themselves using fake trends.

We inspected the attack tweets used to create fake trends reported by Turkish media [20, 26, 159]. We found a pattern in the structure of the deleted tweets: the content appears to be sourced from a lexicon of Turkish words and phrases, e.g., "to organize milk frost deposit panel." They do not have a sentence structure nor do they convey meaning and the verbs are always in the infinitive form. We call these tweets *lexicon*

tweets. Our honeypot account also tweeted and deleted such tweets while promoting fake trends.

In this section, we uncover a massive ephemeral astroturfing operation in Turkey. First, we inspect and annotate trends starting from 2013 and find the first instance of an ephemeral astroturfing attack. Next, we show the features of the attack concerning our attack model. Finally, we build a training set and train a classifier to find all trends that are associated with at least one attack.

3.4.1 Datasets

To study trends, we first need a trend dataset. We collect all trends in Turkey from an external provider[1]. This list contains every trending keyword since July 7, 2013. The trends are collected every 10 minutes and indexed only by date, not time. As such, we treat every date and keyword pair as a separate trend to account for keywords trending in multiple days. Second, we need tweets. To this end, we employ Archive's Twitter Stream Grab [1], which contained 1% of all Twitter data from September 2011 until September 2019 at the time of this analysis. This dataset contains deletions of tweets as well as the exact time the tweet is deleted. We verified that these deletions are due to authenticated users deleting tweets and not due to Twitter administrative actions by contacting Twitter. Our trend dataset does not contain the exact time a trend reaches trending but only the date. Therefore, for each trend, we associate tweets that contain the trend that is either posted on the same day that the keyword was trending or the day before to account for the keywords that were trending after midnight. (We later confirm that our results are robust to this design decision as most of the astroturfed trends do not have any previous discussions that stretch beyond a day earlier. See §3.5 for details.) We name this combined dataset the *retrospective* dataset.

3.4.2 Manual Annotation of Attacked Trends

The goal of the manual annotation task is to uncover which keywords were trending as the result of an ephemeral astroturfing attack and which were not. The annotators inspect trends, along with any tweets, deleted or otherwise, that contain the trending keyword.

We first filter out trends with less than 10 associated tweets so that we are left with those that have enough data to meaningfully assign a label. Of those that remain, we randomly select one trend per day, resulting in 2,010 *trend-date pairs* in total.

The annotators examined the first 10 tweets of each trend and their deletion times, if available. A trend was considered to be initiated solely by an ephemeral astroturfing attack if 1) the tweets were deleted quickly and 2) the content of the first 10 associated tweets have a describable pattern that indicates automation (e.g., lexicon, random characters, repeated text).

[1] http://tt-history.appspot.com

Note that constraining the annotation to only the first 10 tweets may hurt recall, i.e. we may miss the case where many tweets containing the target keyword are posted earlier in the same day of the attack so the attacked trend appears to be organic when only the first 10 tweets are considered. However, our observations and analyses in §3.5 show that this behavior is rare.

Two authors contributed to the annotation process evenly. One author additionally annotated a random sample of 200 trends. The annotation agreement on whether a trend was initiated by an ephemeral astroturfing attack or not was k = 0.88 (almost perfect agreement). We further annotated the tweets associated with each of the 182 trends (5,701 tweets, 5,538 with unique content) as if they are part of an attack (i.e. if they are created and deleted shortly together while having the same pattern in their content) or not. Additionally, both annotators created subsets of the "not ephemeral astroturfing" label for other types of astroturfing attacks (i.e. those which did not employ deletions). These attacks did not employ deletions so they are out of scope for this chapter.

We found that the first instance of a trend employing ephemeral astroturfing attacks was in June 2015 and by 2017 it had become mainstream. Overall we found 182 trends that were astroturfed by ephemeral astroturfing attacks using lexicon tweets. We did not observe any trends that are not promoted by lexicon tweets and still have the deletion patterns in our attack model.

3.4.3 Analysis of Annotated Trends

Time Window of Actions Per our definition, ephemeral astroturfing attacks post many attack tweets in a very small time window ($< \alpha_p$) and delete them in a very small time window ($< \alpha_d$). Fig. 3.3 shows how small this time window is for the attacks we labeled: except for a few outliers, both α_p and α_d are only a few minutes.

Lifetime of Attacks Ephemeral astroturfing attacks post many tweets and then delete them after a short period of time ($< \theta$). Fig. 3.4 shows the difference between the creation and deletion times of each attack tweet (i.e. lifetime, or how long a tweet "lives" before it is deleted) and the median lifetime of tweets per trend. Most have a very short median lifetime; however, some tweets are deleted after a few hours. This might be due to an error on the attackers' side (i.e. buggy code).

3.4.4 Classification to Uncover More Attacks

Next, we aim to automate the process of building a large-scale dataset of ephemeral astroturfing attacks in order to perform a large-scale analysis. We build a simple classifier based on the features of the annotated data and the tweets collected from our honeypot account (§3.3).

Lexicon Content Both our analysis of the annotated trends and our honeypot's tweets tell us that the ephemeral astroturfing attacks that we see in this case employ lexicon tweets, which are trivial to classify. We study the honeypot's tweets to derive the rules

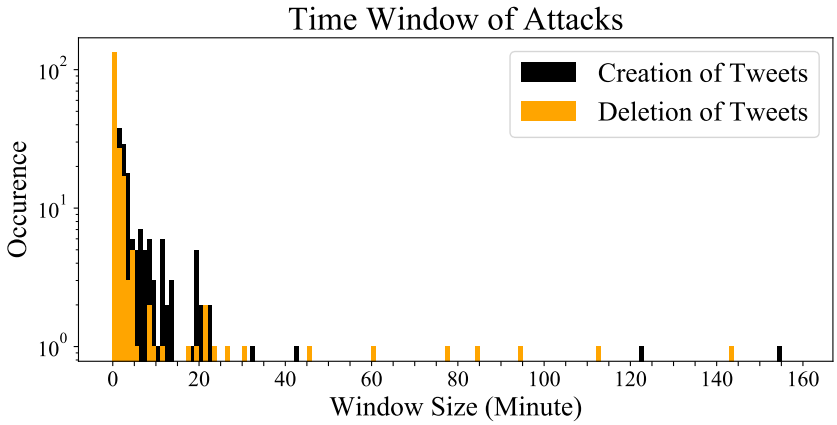

Figure 3.3: The size of the time window in which the attack tweets are created ($< \alpha_p$) is shown in blue. This shows the difference between the first and last tweet created containing the keyword for each trend. The size of the time window in which the attack tweets are deleted ($< \alpha_d$) is shown in orange. This shows the difference between the first and last tweet deleted containing the keyword for each trend. Most attacks occur in a very small time window.

for the lexicon classifier, since we are certain these tweets were sent by the attackers. We came up with the following rules and evaluated them on the 5,538 unique annotated lexicon tweets:

1. Only alphabetical characters except parenthesis and emojis. (99.4% of honeypot, 96.6% of annotated).

2. Beings with a lowercase letter. (99.4% of honeypot, 96.3% of annotated). False negatives were proper nouns from the lexicon.

3. Has between 2-9 tokens, excluding emojis. This range corresponds to the maximum and minimum number of tokens of the honeypot's tweets. In the annotation set, there were 5 lexicon tweets with only one token and 29 with more than 9. (100% of honeypot, 99.4% of annotated).

The combination of these rules yields a recall of 92.9% (5,147 / 5,538). To compute precision on deleted tweets, we ran the classifier on all of the deleted tweets in the sample of 2,010 trends: 17,437 tweets in total after dropping any duplicates (e.g., retweets). The classifier reported 370 lexicon tweets or a precision of 93.3%. Of the false positives, 336 were from before June 2015, indicating that they were used in astroturfing attacks that predate the rise of ephemeral astroturfing using lexicon tweets There were only 34 false positives after June 2015.

To corroborate the precision at scale, we show that lexicon tweets are common among deleted tweets associated with trends but rare otherwise. We classify all Turkish tweets in our retrospective dataset from June 2015. Fig. 3.5 shows that most lexicon tweets

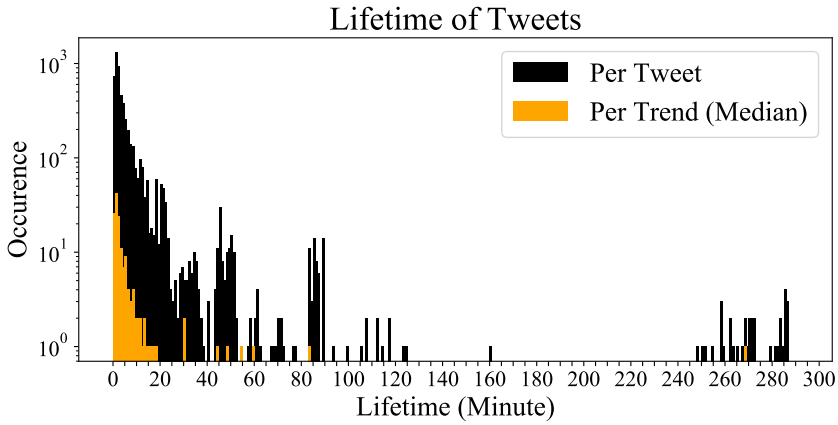

Figure 3.4: Lifetime, the difference between time of creation and deletion of the annotated lexicon tweets. Blue shows the lifetime of individual tweets, and orange shows the median of the lifetime of tweets per trend. Attackers delete the tweets in 10 minutes (in most cases) and the difference between two histograms suggests that sometimes they miss some tweets to delete.

associated with a trend are deleted, but very few lexicon tweets not associated with a trend are deleted.

Although lexicon tweets appear to be generated randomly using a lexicon of tweets, some occur more than once. Table 3.1 shows the most commonly repeated tweets (excluding the target keyword), their translations, and the number of times they occur in the data. We also observe that some words are so uncommon that even a native Turkish speaker may need to refer to a dictionary. This suggests that the attackers may be using infrequent words to pass Twitter's spam filters.

Table 3.1: The most frequent lexicon tweets found in the dataset.

Frequency	Tweet	Translation
77	tenkidi kaynaştırabilme siperisaika	critical to be able to boil lightning rod
64	yarım gün güzelleştirilme oyalayabilme	half day to be prettifiable to be able to distract
64	kargocu yan bakış azımsanma aforozlanma	deliveryman side view to be underestimated to be excommunicated
64	yemenici kalsiyum klorür yarım bağlaşım koyulaştırmak	hand-printed head scarve maker chloride half coupling to coagulate
62	örgütleme süt karlanmak panel	to organize milk frost deposit panel

Supplementary Annotations Our annotated dataset contains only 182 astroturfed trends, which is too few to train a classifier. As such, we perform a second phase of annotations to extend the dataset. We selected a random sample of 5 trends per day after Jan 1, 2017, (4,255 trends in total) and annotate whether or not they were part of a lexicon attack. As this task is much larger than the previous one, and because we now have more information about how these attacks operate, to speed up annotations we only considered deleted tweets associated with a trend. We look for a burst of lexicon tweets posted and then deleted. We found that the condition *at least four lexicon tweets* successfully differentiated attacked and organic trends, with only one instance of an

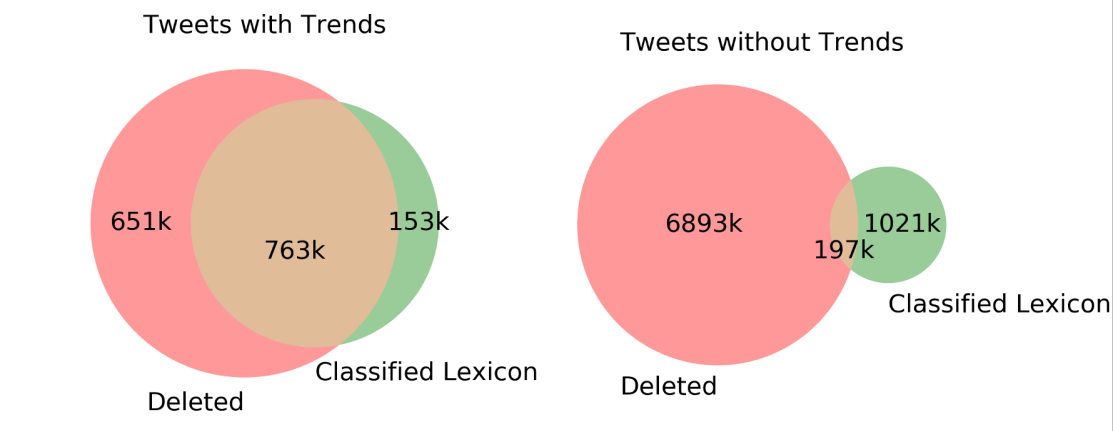

Figure 3.5: Venn Diagram of the retrospective dataset concerning deleted and lexicon tweets. Tweets that are classified as lexicon account for only 2.3% of all deleted tweets that are not associated with any trend (right diagram), but 53.1% of all tweets associated with a trend (left diagram). Further, 83.2% of all tweets that are classified as lexicon and associated with a trend are deleted.

attacked trend with fewer (3) lexicon tweets. Two annotators confirmed and corrected the labels. The resulting dataset contains 838 trends that were associated with at least one attack and 3,417 which were not which indicates a base rate of 19.7% for the attacks.

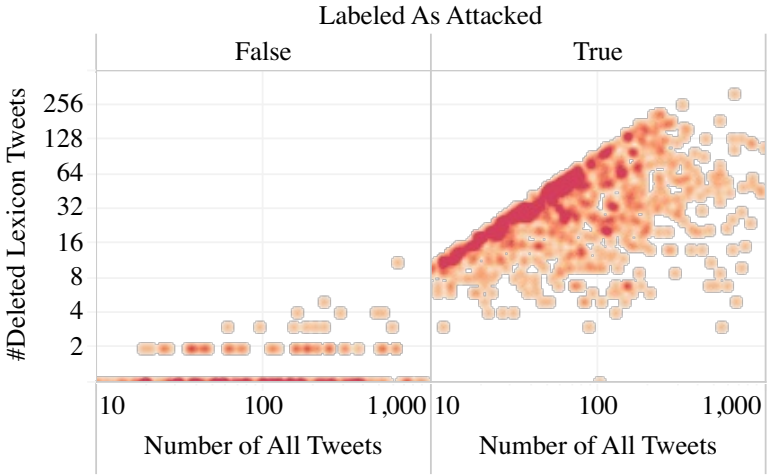

Figure 3.6: The number of deleted tweets classified as lexicon and number of all tweets per trend labeled as attacked (right) and other (left). Four deleted tweets classified as lexicon clearly separate the two classes.

Fig. 3.6 shows the results of our lexicon classifier on the deleted tweets. It can separate the positive and negative cases in most cases. The classification task is then to account for the few false positives and negatives.

Classification We sort the trends by date and use the first 80% as training data. The test data starts from trends in February 2019. The training set contains 648 positives

and 2,756 negatives while the test set contains 195 positives and 656 negatives. A simple decision tree that checks if there are at least 4 deleted lexicon tweets associated with a trend and if more than 45% of all lexicon tweets are deleted achieves a 99.7% 5-fold cross-validation score, 100% precision, 98.9% recall, and 99.4% F-score. The classifier can achieve such good results because it is classifying a very specific pattern that came to be due to a vulnerability in Twitter's trending algorithm. It is no surprise that the attackers have not changed their attack method since their current method is already very successful. Note that 4 lexicon tweets in the 1% sample maps to roughly 400 lexicon tweets in reality, a clear anomaly considering that lexicon tweets are rare.

This classifier found 32,895 trends (19,485 unique keywords) associated with at least one attack between June 2015 and September 2019. Most were created from scratch (astroturfed) but very few were promoted after they reached trending (see §3.5). We refer to these as *attacked trends* for the remainder of this chapter.

Classification of Astrobots Classifying any user who posted a tweet containing an attacked trend with a lexicon tweet deleted within the same day as an astrobot yields 108,682 astrobots that were active between June 2015 and September 2019. 44.9% of these users do not remain on the platform as of July 2020. Through the users/show endpoint of Twitter API, we found that **27,731** of these users are **suspended**. For the rest (21,106), we are given a user not found error (code 50). Those users may be deleted by the account owners. We leave a fine-grained classification of astrobots to future work.

Other Countries We manually examined temporal activity (i.e. the number of tweets and deletions per minute) associated with non-Turkish trends with more than 10 deletions but did not find any positive example. We additionally built a lexicon-agnostic classifier and ran it on all hashtags contained in non-Turkish hashtags but failed to find positives that we could reliably Thus, the remainder of the chapter will focus on ephemeral astroturfing attacks on Twitter trends in Turkey.

3.5 Attack Analysis

In this section, we analyze the trends associated with the attacks to first answer if the attacks cause or just promote the trends. We then measure the success of the attacks using various metrics. We also examine the other tactics the attackers may have employed by studying the time of the trends and the tweets' location information. Lastly, we show an anomaly in the volume field provided by Twitter which shows how many tweets talk about the associated trend and discuss what it may signify.

Part of this analysis requires a dataset of trends that contains their exact time and ranking. We were unable to find such a historical dataset; however, we collected a detailed dataset of the top 50 trends in real-time from Turkey and the world between June 18, 2019, and September 1, 2019, by sending API requests to Twitter every 5 minutes. We name this dataset the *real-time trends* dataset.

3.5.1 Causation or Promotion?

Our initial observation which we build our classification method upon is the enormous number of lexicon tweets being created and subsequently deleted before the new keywords reached the trend list. As our retrospective dataset does not contain the exact time a trend reaches trending, we could not use this information in our classification and only classify if a trend is attacked at some point. We now show that for the majority of the trends, the attacks are the only activity before the target keyword becomes trending, and thus, attacks cause the trend.

Using the real-time trends dataset, we first collect all tweets associated with each trend from the retrospective dataset, before the first time the trend reaches the top 50 trends list until the first time they dropped from the top 50 trends list. Twitter states that the trending algorithm shows what is popular now [277] which means that they take recency of tweets containing a trend as input. We did not see any major difference in the results when we only consider recent tweets, i.e. tweets created within an hour, and thus show results without accounting for the recency. We later found that this was because attack tweets were generally very recent, created within five minutes before the target keyword becomes trending (See §3.5.2.3).

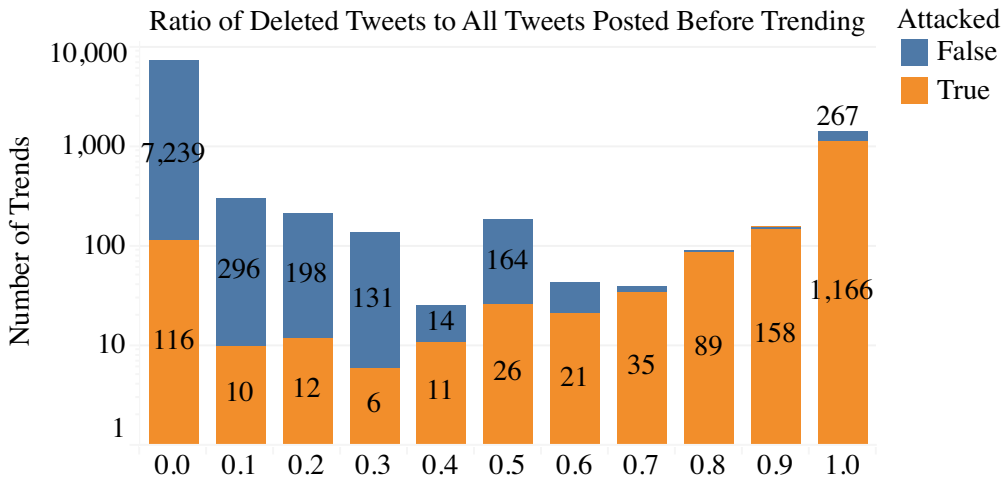

Figure 3.7: The histogram depicting the ratio of all tweets that are created and deleted to all tweets created before the trend enters the list. This ratio is overwhelmingly high for attacked trends while it is zero for the majority of non-attacked trends.

Fig. 3.7 shows the ratio of tweets deleted to all tweets for each trend. Strikingly, the attackers delete all their tweets **even before the target keyword reaches trending** in n = 1166 / 1650 (70.6%) cases. This demonstrates that Twitter's trending algorithm does not account for deletions. The attackers likely delete quickly because they aim to provide their clients with a clean slate once the keyword reaches trending. Additionally, the attackers may want to hide the fact that this is a fake trend from the public since people can see the attack tweets once the target keyword reaches trending by clicking on the keyword on the trends list. Very few non-attacked trends have a high percentage of

deletions. These trends have less than four tweets found in the retrospective data and as such, they are either false negatives or noise.

For 90.6% of the attacked trends, the tweets deleted within the same day make up at least half of the discussions. Our further analysis yields that these deleted tweets are indeed lexicon tweets. We examined the data of 155 attacked trends in which deletions make up less than 50% before they reach the trends list. We found that 24 attacked trends did not have any lexicon tweets, suggesting that they may be attacked at another time they were trending. For 56 trends, the lexicon tweets are deleted after the trend entered the list. Only for 37 trends, there were less than 4 lexicon tweets before the trend enters the list and there were many more lexicon tweets posted after the trend reached the list. These trends initially reached the list in a lower rank (median 32.5) but their highest ranks are in the top 10 with only 2 exceptions, suggesting that attacks promoted these trends rather than creating from scratch. The rest of the trends have prior lexicon tweets but also had some sort of other discussions. Thus, for at least 90.6% of the cases, the attacks create the trends from scratch while for only 3.7% of cases we can argue that the attacks are employed to promote a trend.

3.5.2 Success Metrics

3.5.2.1 Binary Success

For measuring success, we begin with the simplest metric: does the target keyword reach trends? We detect unsuccessful attacks by running our classifier on tweets associated with keywords that were not trending on that day. If the classifier yields positive, that would mean there was an ongoing unsuccessful attack. We only use hashtags as a proxy for trend candidates as it's computationally expensive to run our classifier on every n-gram in the data. We collect all hashtags and their tweets between June 2015 and September 2019 from the retrospective dataset. We found only 1085 attacked hashtags that did not make it to the trends on the same day or the day after. 169 of those hashtags trended another day. As the number of trends that are hashtags since 2015 June is 21030, we estimate that attacks are successful by 94.8% of the time. However, our results may be biased towards the attacks that employed sufficiently many bots with, which our classifiers can produce a positive estimate.

We consider two main reasons that an attack fails: 1) the attack is not strong enough to be selected as a trend (at least not stronger than the signals associated with organic trends) by the trending algorithm and 2) the attack is filtered by Twitter's spam filters. In the former case, per our attack definition, the failed attack may have fewer posts than the other candidate trends ($|T| < \kappa$), or the time window of the correlated tweets may be too wide ($max(\mathcal{P}) - min(\mathcal{P}) > \alpha_p$). In the case where the attack is filtered by Twitter's spam filters (as in [14]), we observe that some attacks include phone numbers (e.g., *#Whatsapp0***x***x***** *'s are digits), profanity (i.e. *#M****G*t*) or words related to porn (e.g., *#PornocuÖ******, which target an individual claiming he likes porn). There are also cases where the attackers made obvious mistakes

e.g., they intended to push "Ağaç Bayramı" (Tree Fest), but "Ağaç" (Tree) trended, or they typed the target keyword twice and tried to push *#53YıllıkEsaretYardımcıHizmet-#53YıllıkEsaretYardımcıHizmet* and failed because the keyword was too long or it has two hashes. Since the number of unsuccessful attacks is too low and we are limited to only 1% of tweets, it is nontrivial to find exactly why each attack was unsuccessful.

3.5.2.2 Rank

Another measure of success and an indicator that the attacks cause or help trends tremendously is the attacked trends' ability to climb higher and faster than other trends. Fig. 3.8 shows that the rank of trends when they reach the trends list for the first time follows a nearly uniform distribution. However, for the attacked trends, almost all rank in the top 10 with the majority ranking in the top 5 initially. This also shows that attackers' goal is to make the target keyword visible on the main page of Twitter or explore section on its app.

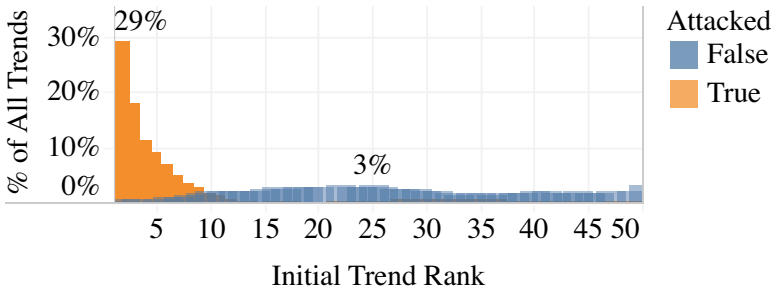

Figure 3.8: Histogram of the trends' initial rank for the attacked trends versus non-attacked trends. Attacked trends' usually rank in the top 5 with the majority ranking 1^{st}.

3.5.2.3 Speed

In addition to reaching higher, attacks also make a keyword reach trends faster than other trends. To measure this, we subtract the median time of tweets posted before the associated keyword reaches the trends list for the first time from the time it reaches trends which we name the **speed** of a trend. Fig. 3.9 shows that the speed of attacked trends is much higher and concentrated around 5 minutes which amounts to the time Twitter refreshes the trends list. This suggests that the attackers do not even start some sort of discussion before the target keyword, but just attack with enough bots to make it reach the trends suddenly.

3.5.2.4 Duration

Another measure of how well an attack succeeds is how long the attacked trends stay in the trends list. The attacked trends stay in the trends list for longer even when compared to non-attacked trends that also entered the trends in the top 10, as Fig. 3.10

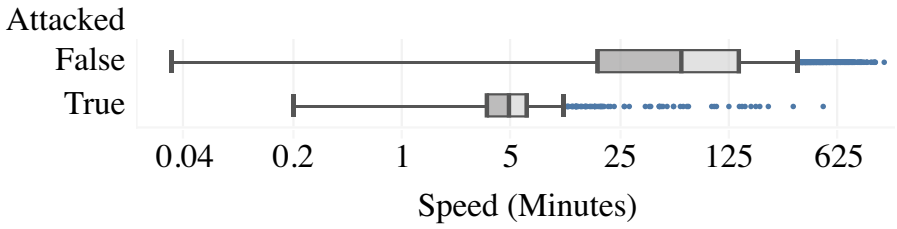

Figure 3.9: The speed of keywords reaching trending. Most of the attacked trends reach trending around just 5 minutes, very fast when compared to other trends (median: 63 minutes).

shows. The initial attack's strength may influence the length of the trend. However, additional actions may play a role in influencing the length of the trend. The attacks may be combined with an organic or an inorganic campaign or a mixture of two (as in #İstanbulunUmuduİmamoğlu in Fig. 3.2) or may capture the attention of the public which discusses the trend for an extended amount of time (as in #SuriyelilerDefolsun in Fig. 3.2) or the trend is promoted by subsequent attacks (as in #SamsununAdresiMacellanCafe in Fig. 3.2).

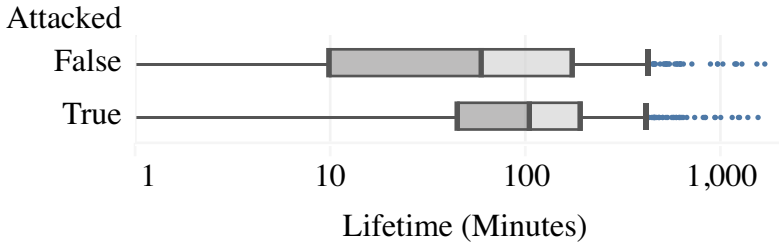

Figure 3.10: Lifetime of top-10 non-attacked trends (top) versus attacked trends (bottom). Attacked trends tend to stay longer (median: 105 minutes) in the trending list when they initially enter the trends list even when compared to other top 10 trends (median: 60 minutes).

3.5.2.5 Impact on Trends

Now that we have shown how successful the attacks are individually, we estimate the prevalence of this attack in terms of the percentage of daily trends that are manipulated. To measure the prevalence, we record how many unique target keywords we know to be artificially promoted by ephemeral astroturfing attacks per day and reached the trends list, and compare it to the total number of unique trends on the same day. From June 2015 to September 2019, we found 32,895 attacked trends, making up 6.6% of our top 50 trends data since June 2015. However, this is likely an underestimation. First, because not all trends' data are found in the 1% real-time sample. More importantly, as we observe in §3.5.2.2, attacks only aim for the top trends because only the top trends are visible and would make an impact. Therefore, using our real-time trend dataset, we

compute the percentage of the attacked trends to all trends positioning themselves in the top 5 and the top 10 trends. Figure 3.11 shows the percentage of top trends that are attacked for the trends in Turkey (upper) and the world trends (lower), positioning in the top 10 (bars) and the top 5 (lines). The daily average of attacked trends reaching the top 10 is 26.7%. This number goes as high as 47.5% for the top 5, reaching the highest on July 19, 2019, to 68.4%, 4 days before the June 23, 2019, Istanbul election rerun. Crucially, many of these keywords reached world trends. The daily average of attacked trends reaching the top 5 is 13.7% reaching the highest 31.6% while this number is 19.7% for the top 10 trends with a maximum value of 37.9%.

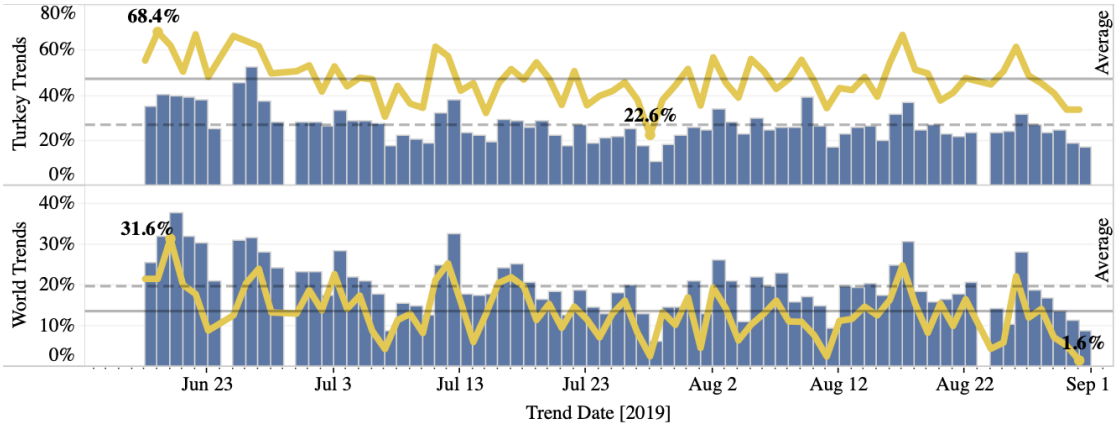

Figure 3.11: Percentage of the attacked trends reaching the top 10 (bars) and the top 5 (lines) trends in Turkey (top) and the world (bottom.) per day. The daily average of attacked trends positioning themselves in the top 10 trends in Turkey is 26.7% while this value goes high as 47.5% for the top 5. The highest value is 68.4% on 19 June 2020, four days before the Istanbul election rerun and the minimum value is 22.6%. The daily average of attacked trends positioning themselves in the top 10 global trends is 19.7% and 13.7% in the top 5, maximum 37.9%, and 31.6% respectively.

3.5.3 Tactics

3.5.3.1 Time of Day

We now turn to one of the tactics the attackers may be employing, sniffing the best time to execute attacks. For those trends which make to the top 10, the trends that are not associated with attacks generally enter the trend list in the morning while the attacked trends mostly enter the list at night as Figure 3.12 shows. The attackers may be choosing nighttime to maximize the binary success; they may be assessing the agenda of the day to decide on how many astrobots to employ, or whether to attack or not. It may be also because the organic trends tend to enter the trend list in the morning possibly due to news setting the agenda and creating competition. Alternatively, it may be due to maximizing the impact; the attackers may be considering the night hours as a better time to attack since people may be surfing on Twitter more at night and thus be more susceptible to attention hijacking.

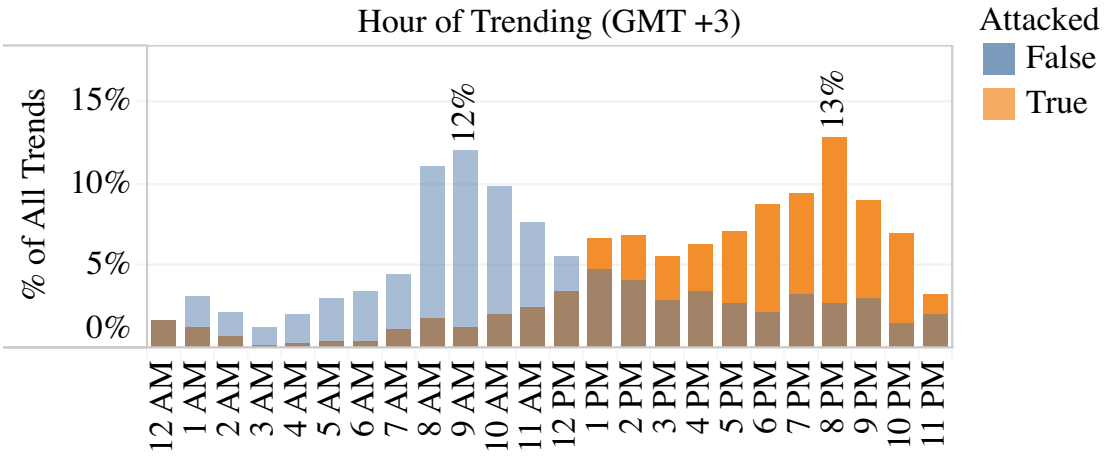

Figure 3.12: Percentage of keywords entering the trends list in a specific hour. The attacked trends enter the trends list mostly at night (Turkey time) while others enter in the morning.

3.5.3.2 Location Field

It is likely that attackers spoof locations to make the trend nationwide instead of in a single city. Additionally, the trending algorithm may be favoring trends with geotagged tweets or trends discussed in a wide range of locations. Similar behavior was reported in [76] in which pro-government Twitter users organize a disinformation campaign against Istanbul's mayor about a water shortage in Istanbul but the tweets are posted from 16 different cities. To show this, we collect the geotagged tweets in the retrospective data, 285,319 tweets in total. Of the 285,319 geotagged tweets in the retrospective dataset, 77.63% are associated with attacked trends even though their tweets make up 25.3% of all tweets. 95% of the geotagged tweets associated with attacked trends are deleted while this is only 14% for other trends. Fig. 3.13 shows the number of geotagged tweets and the percentage of deleted geotagged tweets to all geotagged tweets per trend.

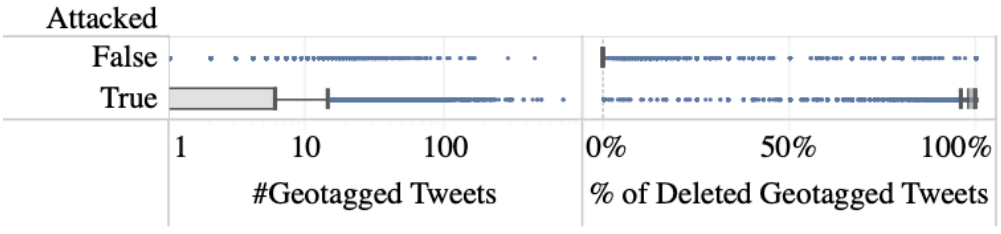

Figure 3.13: The number of geotagged tweets (left) and the percentage of geotagged tweets deleted to all geotagged tweets(right), per trend. Attacked trends have more geotagged tweets and the majority are deleted.

To verify these geotags are indeed fake, we tracked 5,000 users which we manually confirmed were astrobots, in real-time for one week. Out of the 3140 bots active at that

time, 384 had at least two distinct geolocated tweets. We then compute the total distance between all of the points (in chronological order) in a 5-day span for each account. The average distance covered in one week by astrobot accounts was 24,582 km: a round trip from Istanbul to the capital, Ankara, 70 times.

3.5.4 The Volume Of Trends

The Twitter API's *GET trends/place* endpoints both provide the trends and their *volumes* which signifies the number of tweets discussing the trend as computed by Twitter's internal tools. Though in reality the number of tweets posted to an attacked trend is higher than other trends, the *volume* of attacked trends is lower compared to other trends, as Fig. 3.14 shows. While the black-box nature of the volume field obscures the true reason, it may be that attacked trends were promoted by other bot-like accounts that Twitter discarded while computing the volume of tweets associated with trends.

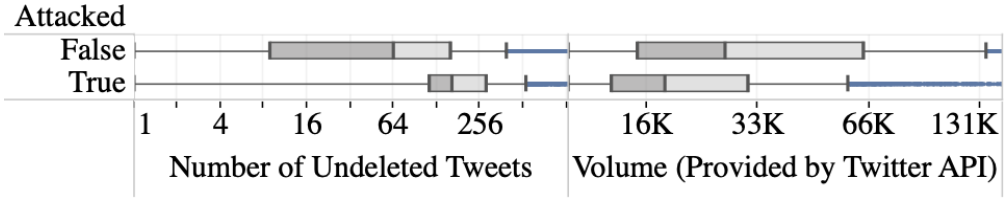

Figure 3.14: The number of undeleted tweets related to attacked trends and other trends vs the volume field provided by the Twitter API. While the former is higher for attacked trends (median is 166 vs 64 for other trends), the latter is higher for other trends (median is 27k versus 18k for attacked trend). This may mean that Twitter filters out the inorganic behavior associated with trends while computing the volume. The minimum volume is 10,000 likely because Twitter sets the volume to null when it is below 10k.

3.6 Account Analysis

In this section, we analyze the types of accounts the attackers employ. We sampled 1,031 astrobots that were employed in these attacks which were still active in March 2019. We inspected the profile and the recent tweets of the accounts and came up with categories based on the content and time of their tweets in an open-world setting. One author annotated all accounts and another annotated 100 to report inter-annotator agreement, which was K = 0.707 (substantial agreement.) The annotators established three categories that covered 947 accounts (92%): 1) inactive (zombie) accounts, which are either dormant for years or have no persistent tweets but actively post lexicon tweets and then delete them (n = 304), 2) retweeter accounts, whose timelines are made up of only retweets (n = 157), and 3) accounts that appear to be human due to their *sophisticated*, original, recent tweets (excluding retweets) and conversations with other users on the platform. We defined sophisticated as containing genuine sentences that convey meaning and have standard grammar (n = 486).

We suspect that most if not all of the users from the latter group are compromised accounts, which were also reported by Turkish media [17, 29, 242]. The most compelling evidence to support this is that the accounts' political orientations observed in undeleted tweets and deleted tweets are inconsistent. Pro-opposition hashtags such as *#HerŞeyÇokGüzelOlacak* (a candidate's slogan, *#EverythingWillBeGreat*) are the most prevalent and adopted by 104 users. However, the most prevalent hashtags among deleted tweets of this otherwise pro-opposition group of accounts are obvious spam advertising trend topic service and/or fake follower schemes. When we examine the hashtags in the deleted tweets, we find that they contradict the political views found in the other tweets: 43 tweeted the pro-government hashtag *#ErdoğanıÇokSeviyoruz* (*#WeLoveErdoğan*), and 19 tweeted with the anti-opposition hashtag *#CHPİPinPKK-lıadayları* (*#PKKMembersAmongCHP&İyiParty*) which claims that the opposition is aligned with terrorists.

We also contacted and interviewed 13 users whose accounts appear to be compromised, using a non-Twitter channel when we were able to locate the off-platform (Twitter = 8, Instagram = 3, Facebook = 2). We informed the user that their account was being used in a botnet and verified that their account was compromised, with the attacker taking partial control. The users either did not know they were in the botnet, or they were helpless and did not think it was a big enough problem to address since the attackers only use them to astroturf trends and quickly delete their tweets.

On June 12, 2020, Twitter announced that they suspended and published the data of 7,340 fake and compromised accounts that made up a centralized network [278]. The accompanying report by the Stanford Internet Observatory claimed that the accounts, among other tactics, employed astroturfing, aimed at pressuring the government to implement specific policy reforms." [132]. The report did not mention fake trends created by the attack we describe here and nor their prevalence. To show that part of these accounts removed on that occasion were indeed astrobots, we cross-referenced these accounts with those in our retrospective dataset. We found an overlap of 77 accounts which we manually identified as astrobots as they tweeted lexicon tweets. Of these, 27 had lexicon tweets that were published by Twitter and publicly accessible. We examined the non-deleted tweets of all 77 accounts to identify their purposes. Only 5 of these accounts appeared to be pro-government while 25 exhibited bot-like behavior since they were only employed to promote hashtags on policy reforms. Eight users were openly anti-government while the rest appeared to be apolitical. This further backs up our claim that some of the astrobots are non-pro government accounts that are compromised, as this is also how Twitter framed the accounts they suspended in this dataset. There are likely many more compromised accounts astroturfing trends in this dataset, but we cannot identify more without access to any deleted tweets, which Twitter did not share.

We combined this data with the deleted tweets found in our retrospective data and identified 77 astrobots tweeting lexicon tweets. Of these bots, 27 were identified through the data Twitter provided since the attackers did not delete these users' lexicon tweets.

We examined the non-deleted tweets of these 77 accounts to identify their purpose. Only 5 appear to be pro-government while 25 have bot-like behavior, as they were only employed to promote hashtags on policy reforms. Eight users were openly anti-government while the rest appear to be apolitical. Our findings are in line with Twitter's, which announced that they had suspended non-pro government users that were compromised. There are likely many more compromised accounts astroturfing trends in the dataset, but we were not able to identify more. Since Twitter did not share the deleted tweets, we needed to rely on the 1% sample for deletions.

3.7 Attack Ecosystem

So far, we have claimed that the goal of the attack is to reach the trends list. However, if we take a step back there's a more important question to ask: "*Why* do the attackers want to reach the trends list?" In this section, we analyze the trends themselves and the ecosystem that supposed the attack to uncover the motivations of attacks.

3.7.1 Topical Analysis of Trends

To understand what topics the attackers promote, we first perform a qualitative analysis on the topics of the attacked trends. We collected all 6,566 unique astroturfed keywords that trended in 2019 and labeled them according to which specific group (e.g., political party) that each trend promoted if any. We also annotated a supercategory for different types of groups. Two annotators labeled 3,211 keywords, one using network information if available (i.e. if two keywords are promoted by the same set of users) and the other using only manual examination. The manual examination consisted of reading the keyword itself or searching for it on Twitter and/or Google for the context. The annotator agreement was K = 0.78 (substantial agreement). The remaining 3,355 were annotated by only one annotator due to the absence of network information. The resulting supercategories with the group annotations in their descriptions are the following:

Illicit Advertisement (n = 2,131): Trend Spam, i.e. trend topic (TT) services, or the fake follower schemes (n = 259), betting websites (n = 1421), a local social media platform (n = 27), a logo design service (n = 20), illegal football streaming websites (n = 24) and others (n = 380). Advertisements often promote the same entity multiple times using slightly changed keywords. This may be because the attackers believe that the trending algorithm favors new keywords. We also observed that these trends are not usually tweeted about by real users. The account of the advertised entity was the first, and in some cases, the only, account to include the trending keyword in a tweet, i.e., attackers push a gambling website to trends, then the betting website's Twitter account uses the new trend and becomes visible in the "Popular Tweets" and "Latest Tweets" panels on Twitter.

Politics (n = 802): Political slogans or manipulations in support of or against a political party or candidate. Pro-AKP keywords in support of the Turkish government

(n = 348) and those that target the opposition (primarily CHP) negatively (n = 124) are the majority. There are also slogans in support of the main opposition party and its candidates (n = 118), other parties (n = 42), or targeting AKP (n = 20). The rest are keywords related to politics but not political parties.

Appeal (n = 1,219): Appeals to government suggesting some policy reforms, as in [132]. These state the demand of the client either in a camelcase form that makes up whole sentences (e.g., MrErdoganPlsGiveUsJobs) or is a very long hashtag (e.g., #JobsTo5000FoodEngineers). The demands are for jobs (e.g., food engineers, contracted personnel, teachers, etc.) (n = 730), for pardoning prisoners (n = 157), for military service by payment (n = 54), and on other profession-related issues. Some of the demands are heavily opposed by the government, which suggests that attackers do not always work in favor of the government.

Cult Slogan (n = 592): Trends that are about various topics but are all sourced from either the Adnan Oktar Cult (n = 474), Furkan's Cult (n = 105), or the Tevhid cult (n = 13), as the users campaigning using the corresponding trends explicitly state they are associated with the cult. All of the cults' leaders were arrested and some of the trends demanded the release of their leaders. Other trends include promoting the cults' views, e.g., spreading disinformation about the theory of evolution and the LGBT community. Turkish media has reported that Adnan Oktar and Furkan cults manipulate trends using bots [22, 157].

Boycott (n = 92): Appeals to people to boycott Metro and MediaMarkt (n = 45) or other companies (n = 47).

Miscellaneous (n = 1,730): Trends that are about any topic including social campaigns, TV shows, names of individuals, or random slogans. As they do not have interest groups, they may have been purchased by people who do not have the intention to campaign and may not be involved in multiple attacks. This corroborates that attacks are a business model with a wide range of clients, which is also reported by the Turkish news media [17, 77, 242].

3.7.2 Astrobot Network

As our attack model indicates, each trend is promoted by a set of *astrobots*. The same set of bots can promote any trend as long as the bots are still controlled by the attackers. Thus, a set of bots consistently attacking the same trends are assumed to be controlled by the same attacker. Then, the same set of bots promoting keywords related to conflicting groups (e.g., opposing political parties) would indicate that the attacks are not executed by the interested parties, but that the attacks are provided as a service to different clients. To explore this, we extract and characterize communities of astrobots by analyzing their topical and temporal activities. The latter provides insights into how Twitter may be defending the platform.

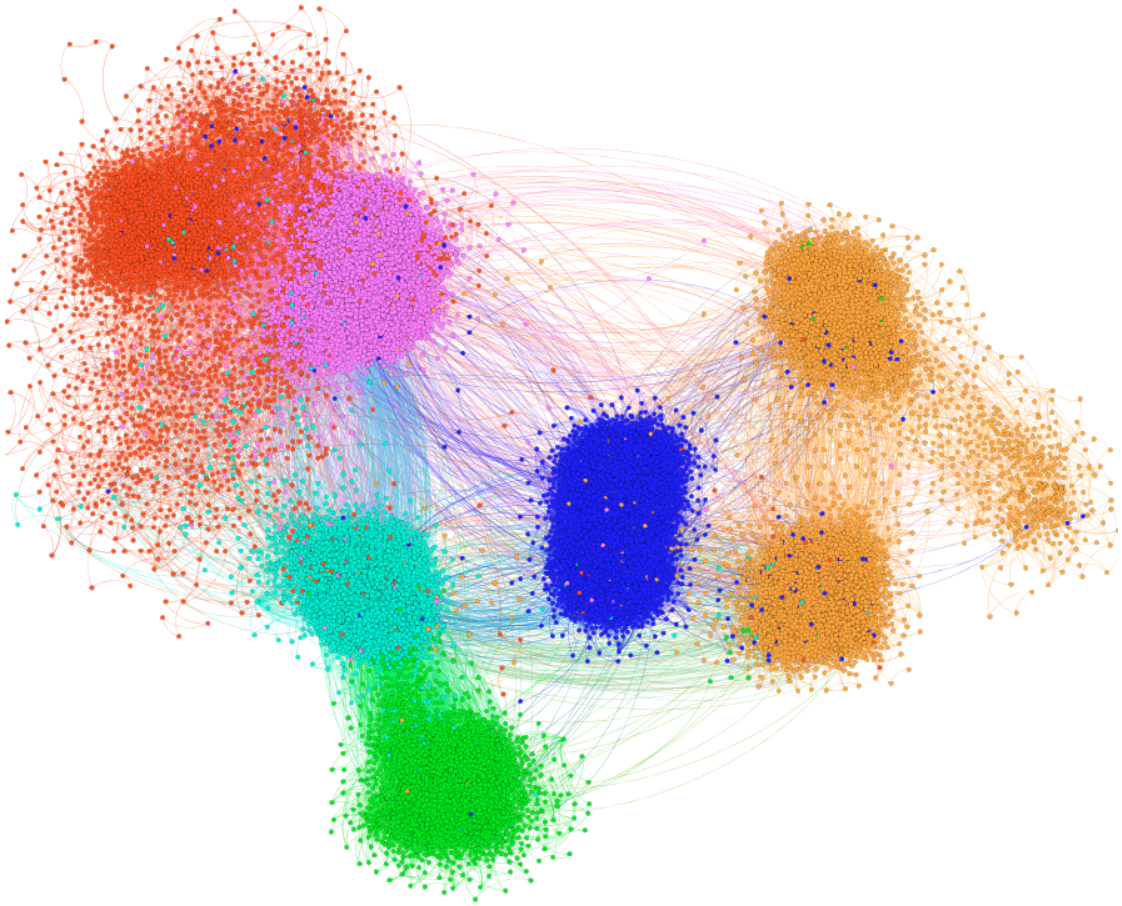

Figure 3.15: The astrobot network visualized in OpenOrd [190] layout using Gephi [32]. Colors indicate the communities obtained by the Louvain method [38]. The attackers lost control of the green and cyan communities by February 2019 while the remaining communities still participate in the attacks by September 2019. Spam trends that promote the fake follower service to compromise more users or promote the top trend service are mainly sourced from the blue community which has a central position in the network.

We build the astrobot graph network in which the nodes are accounts and the edges indicate that the accounts participated in the same attack (both posted a deleted lexicon tweet containing the trend). This network had 33,593 users active in 2019, 71.6% of which were still active as of July 2020. Surprisingly, the intersection of the set of users promoting the trend and not deleting the tweets (147,000 users) and the set of astrobot accounts was only 817, suggesting that the astrobots' only function is to push keywords to trends and stay idle otherwise. This is likely part of the stealthy nature of ephemeral astroturfing attacks; the attackers do not want any of their bots associated with the campaigns they promote, so they do not employ them for non-ephemeral astroturfing attack tweets. Instead, the clients outsource pushing trends to attackers and then execute any other activity themselves.

From the users active in 2019, we removed those who attacked only once to remove noise, leaving 21,187 users. We performed community detection using the Louvain

method [38]. The Louvain method greedily optimizes modularity, which ensures nodes that have strong connections are in the same community. Thus, astrobots consistently attacking the same trends will be in the same community. We found 6 communities, with modularity 0.711. We name these communities by the coloring scheme: green, cyan, blue, orange, pink, and red. Fig. 3.15 shows the resulting network. Table III shows the number of trends and users and the percentage of users that remain on the platform by July 2020 within each community, as well as any pattern(s) we found. We now describe the temporal and semantic patterns the communities follow in detail.

Table 3.2: Statistics of the communities. Persist denotes the percentage of users not suspended or deleted within the community as of July 2020. Summary refers to the pattern(s) that characterize(s) the communities.

Community	Users	Persist	Trends	Activity	Topic
Green	2,079	81%	291	1/19 - 2/19	Misc.
Cyan	3,701	78%	839	6/18 - 2/19	Appeal (Pardon)
Blue	4,845	70%	1,043	1/19 - 9/19	Ads (Spam)
Pink	4,719	74%	2,422	3/19 - 9/19	Ads (Betting)
Red	2,627	73%	941	3/19 - 9/19	Various
Orange	3,216	71%	913	4/19 - 9/19	Cult (Furkan)

Temporal Activity Studying temporal activity is key to learning how many networks of astrobots are active at a given point in time and how quickly Twitter addresses the coordinated activity. Fig. 3.16 shows the first and last times the astrobots were found to be participating in an attack which gives us a rough idea of the active times of their respective communities.

We found that the cyan community is the only community in which the majority of the accounts (79%) were actively participating in the attacks before 2019 while the rest became active in mid-2019. Exceptionally, the green community's users were active since January 2019. The green and cyan communities stopped attacking in February 2019, however, most of the accounts in these two communities remain on the platform despite not participating in recent attacks. All users that remain on the platform in the green community and half of such users in the cyan community last posted an undeleted tweet in February, as Fig. 3.17 shows. Precisely, 1,887 users became dormant on February 1, 2019. 23% of the users in the green community and 6.7% in the cyan community last tweeted a lexicon tweet in February, none of which were deleted, suggesting that the attackers could not or did not delete the lexicon tweets from their final attack using these communities. The other half of the users in the cyan community remained on Twitter as of July 2020 but did not participate in an attack after February. This suggests that the attackers may have lost control of the accounts either due to internal problems or because Twitter imposed a session reset or a location or IP ban on the accounts.

The fact that two of the communities were inactive by early 2019 and three new communities then became active indicates that the attackers replaced the cyan and green communities with new ones. Interestingly, while the majority of the creation

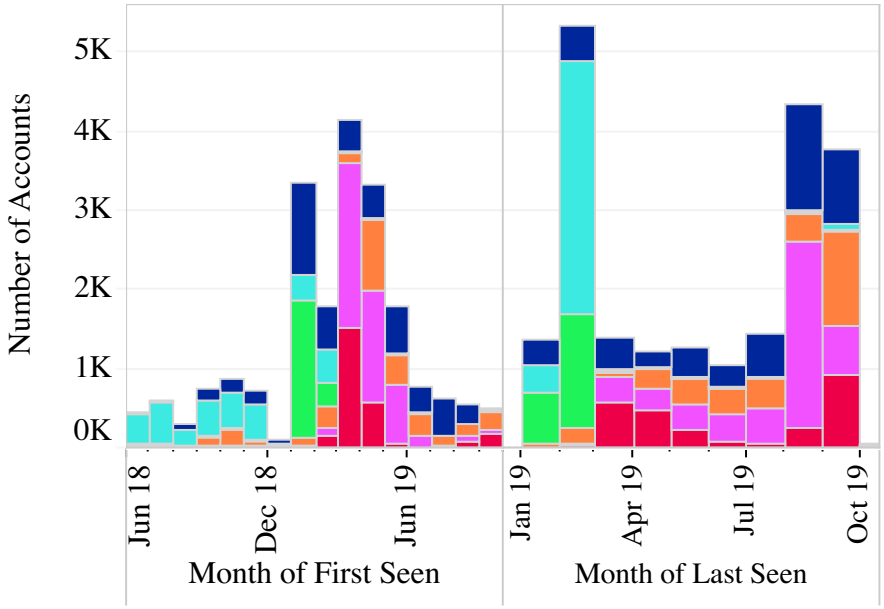

Figure 3.16: The time the accounts are first and last seen attacking. The users from the cyan community were active even before 2019 while the rest of the community became active in 2019. Accounts in the green and cyan communities appear to discontinue attacking in early 2019.

dates of the accounts in the other communities are from 2016, 62% of accounts in the green community were created between 2012 and 2014 even though they did not become active in any attacks until January 2019. Attackers may have bulk purchased and/or bulk compromised these accounts and were detected by Twitter quickly and taken down. The rest of the four communities were still participating in attacks as of September 2019. This indicates that there are four databases of astrobots owned by at most four different attackers.

Topical Activity We now analyze the interplay between attackers and clients by analyzing the topical categories of the trends and the astrobot communities promoting them. Fig. 3.18 shows the distribution using the topics from the previous subsection. Except for the green community, in which 60% of trends were labeled as miscellaneous, no community was dominated by a topic and/or group. Some topics were mostly or uniquely associated with one community, suggesting that groups promoting those topics only collude with one attacker, although the same community promotes other topics as well.

The majority of the bet related ads (80%) and Oktar's cult slogans (68%) were in the pink and red communities. Most (80%) spam ads promoting fake follower schemes and trend topic services were in the blue community. The fact that this community is central to the whole network suggests that the attackers controlling this group provided users and trends for other groups. Food engineers appealing for jobs were also almost uniquely associated with the blue community, while cult slogans related to Furkan were associated uniquely with the orange community. Political trends were dispersed

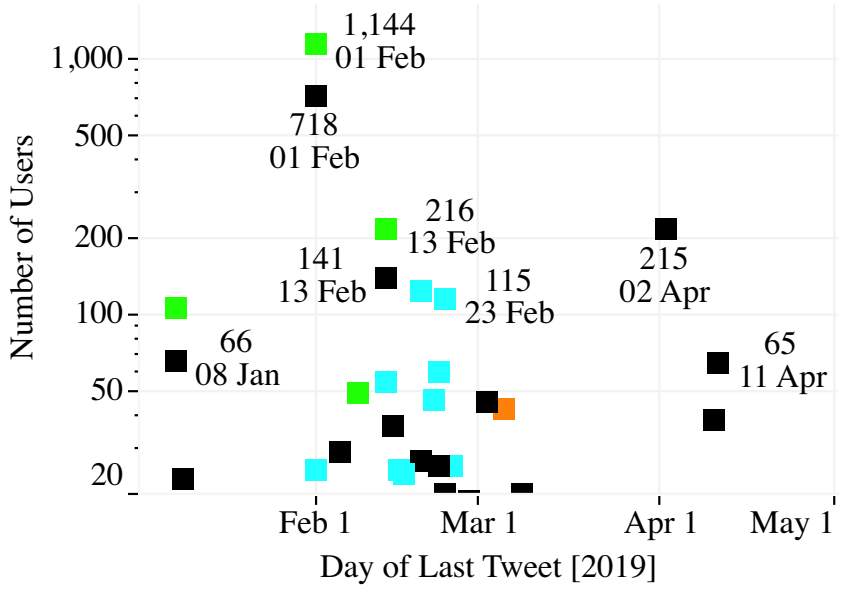

Figure 3.17: The date of last not deleted tweet of each account per community. Accounts shown in black are not assigned to a community. The huge spike in accounts that last tweeted on February 1 and never since may show that attackers lost control of these accounts, although the accounts are not suspended.

throughout the communities and political parties often shared the same community with their rivals. For instance, the blue community contains the highest number (80) of pro-CHP (the main opposition) trends but also has 80 trends in support of its fierce rival, AKP. Similarly, 40% of the pro-AKP trends are associated with the pink community but the pink community has also 15 pro-CHP trends. This further corroborates that the attackers are independent of the parties and provide fake trends as a service.

3.8 Implications

Ephemeral astroturfing attacks principally have an impact on users and the platforms that they attack in terms of (i) platform integrity, (ii) account security, and (iii) attack attribution. It also has a tremendous impact on data integrity in data science studies. We discuss further implications to security research and propose countermeasures.

Platform Integrity Systematic attacks on popularity mechanisms compromise the integrity of the mechanism and the platform. On Twitter, users expect to see legitimate popular content, so when they are shown content that is not popular, they no longer trust that what is shown on the Twitter trends list is actually popular. As with many systems, when the authenticity of a component is compromised, trust in the entire system diminishes, e.g., the price of bitcoin falls after prominent exit schemes. If Twitter trends fails to reliably display authentic trends, trust in trends and Twitter as a whole is diminished. Twitter recently took steps to preserve platform integrity such as suspending accounts involved in coordinated inauthentic behavior, however, they have not addressed

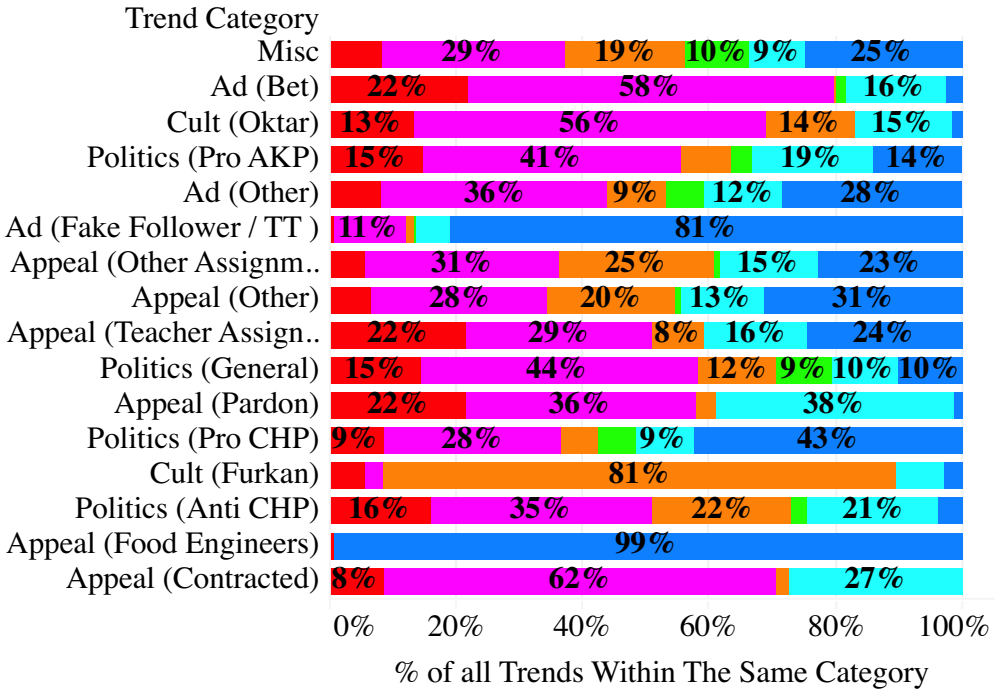

Figure 3.18: The trends according to topics (those with at least 100 trends) and the astrobot community the trends are promoted by. Some interest groups such as contract employees are merged into one.

ephemeral astroturfing attacks which are contributing to a loss of trust among the users affected by the attacks.

Account Security Ephemeral astroturfing reinforces the practice of selling accounts. Because astroturfing attacks attempt to mimic widespread popularity, they require a critical mass of accounts, they necessitate a black market for compromised and fake accounts. Ephemeral astroturfing is unique in that it allows for the use of active, compromised accounts and not only fake accounts. As long as ephemeral astroturfing remains effective, more compromised accounts will be needed to boost the target keywords. While it is challenging to disrupt these markets directly, e.g., via takedowns, they can be disrupted by removing the market demand, rendering fake and compromised accounts useless.

Attack Attribution Malicious online activities are often difficult to attribute to an actor, and astroturfing attacks are no exception. Organic campaigns that are launched by users can generally be attributed to a certain group, ideology, or event. However, in the case of ephemeral astroturfing, the actions of the adversaries are quickly hidden. This makes it possible for adversaries to conduct illicit activities including the promotion of scams and illicit businesses. Ephemerality makes it more difficult to attribute an attack to a specific group, while at the same time legitimizing the activity by making it seem as though the activity is the result of grassroots organizing.

Data Integrity Beyond astroturfing, data science studies often rely on the assumption that data is a static entity. This is especially the case in social media studies, where data is often collected ex post facto. Such an assumption should be taken very carefully, as social media posts and accounts can be deleted or suspended. If accounts or posts are deleted, then the dataset used for evaluation and analysis may be incomplete. In the case of ephemeral astroturfing, we find that a *critical* segment of the data may be deleted: the tweets that illegitimately created the popularity of a topic. Future analysis of a trend that does not consider deleted data may misinterpret how a topic became popular. For example, in September 2018 the trend *#SuriyelilerDefolsun* (#SyriansGetOut) was pushed to the trends list using ephemeral astroturfing attacks, as shown in Fig. 3.2. The hashtag attracted the primarily negative attention of the public after reaching trending. However, academic studies that use the hashtag as a case study [54, 209] or refer to it [53, 166] all attributed the hashtag to the public, completely unaware of the fact that the hashtag was trending due to bot activity, even going as far as to say that social media users launched the hashtag due to a fear of a new wave of mass migration [218].

Impacts on Society Ephemeral attacks expose users to illegal advertisements, hate speech targeting vulnerable populations, and political propaganda. For example, ÇiftlikBank was a Ponzi scheme that claimed to be a farming company aimed at growing quickly and aggressively maximizing profits. They employed ephemeral astroturfing attacks to promote themselves 29 times using slogans such as *ÇiftlikBank TesisAçılışı* (ÇiftlikBank Facility Opening) and *ÇiftlikBank BirYaşında* (ÇiftlikBank is one year old) which give the impression that it is a growing business. They did not trend organically until December 2017, and only then because they started to raise suspicions [41]. They attempted to counter this suspicion by using ephemeral astroturfing to push *#ÇiftlikBankaGüveniyoruz* (#WeTrustInÇiftlikBank) into trends. ÇiftlikBank's boss scammed $129 million before escaping in March 2018 [19, 140].

Taxi drivers in Istanbul used ephemeral astroturfing to protest Uber [110]. Some of the slogans aligning with their campaign were used to sow hate against the drivers, e.g., *#KorsanUberKapanacak* (#PirateUberMustShutdown), *#RüşvetçiUber* (#UberTakesBribes), and *UberAraçSahipleriAraştırılsın* (#UberDriversShouldBeInvestigated). Other hateful messages targeted specific individuals demanding their arrest, e.g., *#FetöcüKuytulTutuklansın* (#ArrestKuytulHeIsATerrorist). Alparslan Kuytul is the leader of Furkan Cult and has an anti-government stance. Others spread hate speech and disinformation targeting vulnerable populations; the LGBT community was targeted at least 24 times by these attacks in 2019 with trends such as *LgbtiPedofilidir* (LGBTisPedophilia) and *DinDevletDüşmanı SapıkLgbti*, (PervyLGBT is enemy of religion and state). Occasionally, counter campaigns were launched by the attack targets, also employing ephemeral astroturfing attacks, e.g., *#UberiSeviyoruz* (#WeLoveUber) and *#HalkUberiSeviyor* (#PeopleWantUber) were used to counter the taxi slogans. Additionally, people seemed to react to the prevalence of trends that appear to be sourced from Adnan Oktar Cult by astroturfing trends like *#AdnancılarMilletSizdenBıktı* (Adnan Oktar

Cult, people are sick of you,) and *#SizinA*kAdnancılar* (Expletives directed at the Adnan Oktar Cult using abbreviated profanity).

Politically motivated groups employed attacks for smear campaigns spreading disinformation and hate speech. During the 2019 local elections in Turkey, many pro-government groups astroturfed trends to target the opposition (e.g., *#CHPPKKnınİzinde*, which indicates that the opposition's party follow a terrorist organization) and particularly opposition candidate Ekrem İmamoğlu who eventually won the election and became mayor of Istanbul. Trends targeting the candidate involved slander asserting that he lied (e.g., *#EkrandakiYalancı*, which means "Liar on TV") and that he stole votes. The most popular astroturfed trend on this issue, *#ÇünküÇaldılar* ("Because They Stole"), was explicitly organized and pushed to the trends by pro-government groups [23] and joined by the rival candidate Binali Yıldırım [85]. Ekrem İmamoğlu condemned the campaign [25]. After, the Supreme Electoral Council decided to rerun the elections but did not state there was any ballot ringing involved; Binali Yıldırım later stated he expressed himself in a colloquial language and the campaign was "an operation to create a perception". [28].

Although many users are exposed to these trends, the extent of the impact is unclear as we cannot measure engagements with trends directly. Meanwhile, public engagement metrics such as the count of retweets and likes per tweet are open to manipulation. However, based on their appearance on other platforms, some astroturfed trends succeed in receiving the public's attention. For example, the mainstream media in Turkey framed many political slogans that trended due to ephemeral astroturfing as grassroots organizing, e.g., *#ÇünküÇaldılar* (#BecauseTheyStole) [24], *#HırsızEkrem* (#EkremIsAThief) [27], and *#SüresizNafakaZulümdir* (#IndefiniteAlimonyisTorture) [21]. Users also posted these slogans on Ekşi Sözlük, a local social media platform where users post entries for terms, because "the public discusses them." *#ÇünküÇaldılar* received 352 entries and *#AtatürkAtamOlamaz* received 145 entries. Perhaps one of the most impactful ephemeral astroturfing attacks was *#SuriyelilerDefolsun* (#SyriansGetOut), which was astroturfed on September 3, 2018, sparking widespread controversy. It was discussed extensively by the media [47, 52], academic works [53, 54, 166, 209, 218] and other social media websites such as Reddit [148], Ekşi Sözlük [188], and kizlarsoruyor.com [37].

For Security Research Although we focus on one case of ephemeral astroturfing, the methodology that we present in this study can be extended to other attacks on popularity mechanisms. All popularity mechanisms work through parsing content to determine what is popular at the moment, though for different mediums and with different definitions of popularity. Considering deleted activity or content as valid leaves open an attack vector, making ephemeral attacks possible. Our results shed light on the problem of astroturfing, framed as an attack on popularity metrics, and how prevalent a problem it can become when left unchecked and unaddressed.

3.9 Generalizability

Ephemeral astroturfing attacks can generalize to any platform where an algorithm determines trending or popular content and does not take deletions into account or have wide periods between consecutive checks for popularity. However, this attack has yet to be explored on platforms other than Twitter. Traditional forums rank threads based on the time of the last reply, thus, spammers post comments to old threads to make them more visible, a practice called bumping [57]. However, forums generally account for deletions and rank the bumped threads lower when the bumper deletes their reply. Reddit considers the current number of upvotes and downvotes at the time it computes the ranking of threads and is therefore likely resistant to ephemerality, i.e. coordinated upvotes proceeded by removing those upvotes [239]. Other possible vulnerable platforms include sites with reviews, like Amazon or the Google Play store, but so far no relevant public analysis of these platforms exists. This attack can also generalize to Twitter trends in any region.

3.10 Counter-Measures

Due to the use of active, compromised accounts, defenses against ephemeral astroturfing attacks are inherently challenging. These accounts, whose owners are victims of the scheme, cannot simply be banned. If the attacks were being executed via a malicious application, Twitter could suspend access to the app, as in [223], but in this case, tweets are posted from official Twitter apps (e.g., *Twitter for Android*). Otherwise, ephemeral astroturfing attacks fit an easily detectable pattern. We outline two main paths for defenses: detecting and inoculating. First, Twitter can implement a detection mechanism to prevent malicious tweets from being considered for Twitter trends, or even made visible at all. They can extend the detection method laid out in §3.4 to find the tweets and accounts involved. Once a trend is found to be manipulated, it can be removed from trends or even prevented from ever reaching them. The second option is to render the attack useless. The fact that these attacks are successful implies that the Twitter trending algorithm does not consider the deletion of tweets. A simple defense, then, is to account for deleted tweets when computing trending topics. For example, the trending algorithm can track the tweets that contain the keyword and heavily penalize the trend's rank for each deleted tweet.

In addition to direct countermeasures, platforms can also work to ensure that even if a popularity mechanism is manipulated via deletions, that users can be aware of potentially suspicious behavior. On Reddit, for example, when a comment is deleted there is public evidence left behind that indicates that a comment was deleted. On Twitter, this translates to an indicator that a tweet that contained the trending keyword was deleted. In this way, when users click on a trend, they are not only shown tweets, but also a series of deleted tweets, which indicate that something suspicious has occurred.

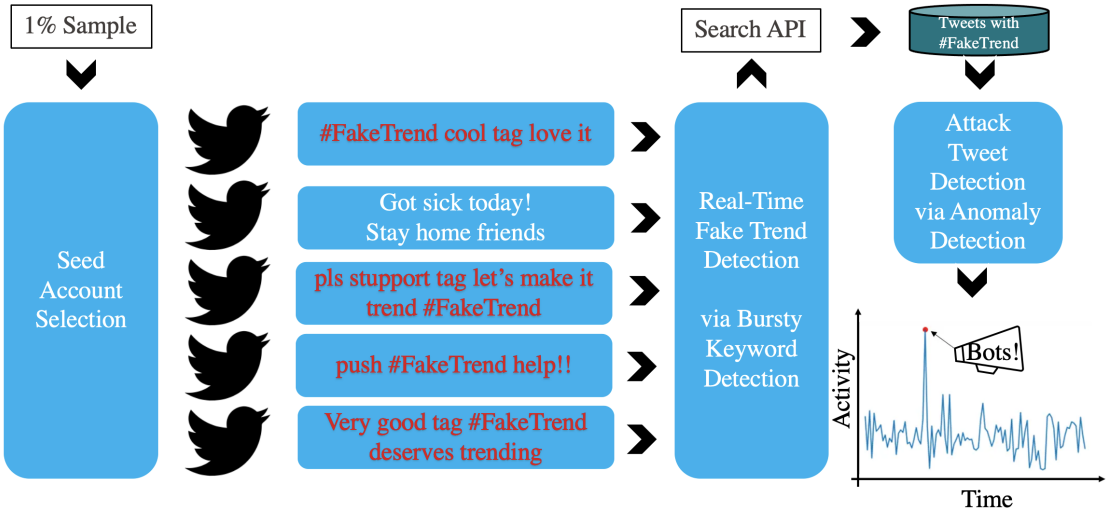

Figure 3.19: Overview of the framework. First, we choose a set of seed users as astrobot candidates. We then listen to their activity and detect fake trends they promote using bursty keyword detection. We then collect the data containing the fake trend using Search API. We finally detect the attack tweets from this data using anomaly detection. We then announce the malicious activity to the public, highlighting the bot activity.

3.11 Real-Time Detection of Attacks for the Public Good

We implemented our own countermeasure against the attack. It has two goals. The first is to collect fake trends and the bots from the full data instead of the 1% sample to increase recall of attack tweets/bots' activity data. The second goal is to announce the fake trends and number of the bots involved to raise user awareness. We do this *during* the attack, so the other users may have the opportunity to observe the tweets by bots before they get deleted. We deployed a Twitter bot to announce the fake trends. We now present our real-time detection framework.

3.11.1 Proposed Framework

We propose a fully automated framework. The system receives a set of seed users as input (§3.11.1.1). It then listens to their activity using Twitter's Streaming API, which provides the full user activity in real-time, including the tweet deletions with deletion times. The framework continuously searches for frequent n-grams promoted by multiple accounts, which may be trend candidates (§3.11.1.2). When it finds a suitable candidate, it collects all the tweets containing this candidate using Search API by setting the candidate keyword as input. Finally, it detects the bots' tweets used for the attack (§3.11.1.3). The framework is depicted in Fig. 3.19. We now explain every component in detail, formulate the problem, and provide the ground truth and result.

3.11.1.1 Seed User Selection

To run the framework, we need a set of known astrobots to use as seeds. The seed should contain an unbiased sample of astrobots to cover the full network of astrobots. We adopt a simple approach. We exploit the information that astrobots delete their tweets that promote a trend before that trend reaches trending, as shown earlier. We collected the last 5000 users who observed this behavior in 2022. We add all the users to the seed list without any classification to avoid classification bias. Thus, the sample is likely to contain false positives, but we assume this will not affect the framework's performance. Since this sample is from 1% of all tweets, we assume it is an unbiased sample. We automatically update the seeds every day with the same procedure.

3.11.1.2 Real-Time Trend Detection

In this component, we collect the tweets astrobots post in real-time. We then detect the fake trends based on the collected activity. The detection has two goals. The first is to collect the data containing the data as early as possible before they are removed. We prefer a detection with high recall in this case. The false positives cause extraneous requests to Twitter's search API, but they do not make the system inoperable due to rate limits as they are not too frequent. The second is to announce the fake trend to the public in real-time while bot tweets are still present. Both precision and recall are important in this case as we want to inform the public about fake trends and not mislead them. Thus, we prioritize models with high F1 scores. Since our goal is to detect the fake trends before the bots remove their activity, we cannot exploit the deletion activity of the bots and only rely on the correlated activity of the bots.

We now formulate the problem and our objective and provide our process in creating ground truth and the results.

Problem Statement Let $T = t_0, t_1 ... t_n$ be the stream of the astrobots tweets. Let FT be the target keyword the astrobots promote in the tweets $T_{ft} = t_s, t_{s+1} ... t_{s+k} \subset T$ and $k < n$. Detect FT at the minimum m such that $s < m < k$.

Since the attacks observe bursty behavior, we constrain the problem by introducing the timeslot T_s so that all or the part of the tweets in T_{ft} should be posted in T_s. We adopt a data mining approach and treat each tweet as a transaction. The objective is then to find w transactions that contain FT with support s such that $s <= w$.

Ground Truth We listened to the tweets seed users in real-time between 17 May and 31 May. We collected the accounts' tweets and deletion notices thrown by Twitter API, which indicated the id and the deletion time of the deleted tweets. We removed the undeleted tweets and the tweets that do not contain trends. We then examined the remaining tweets containing the trends in that period and marked the attacks. That is, we look for a series of uninterrupted lexicon tweets that promote a trend. To simplify the annotation, we upper-bounded the number of tweets to inspect to 10. We found 96 fake trends mentioned at least once by the bots in that period, which we used to

compute the recall. We compute the precision by considering the non-trend frequent patterns extracted by our method as negatives. We found that our methods also extract frequent patterns that are targets of unsuccessful attacks. We discard those keywords in the evaluation process. We use the trends between 17 May and 25 May as the validation set and the trends between 26 May and 31 May as the test set.

Experiments and Results We use the FP Growth algorithm [137] to extract frequent hashtags and n-grams. We perform a grid search on the validation set to find the optimal w and s. We found that lower w and s increases recall while decreasing precision. We report the scores on the validation set and the test set. We report two models: the first model, named *announce* is to announce the fake trends to the public and is optimized by the F1 score on the validation set. The second model, named *collect* is to collect the fake trends and is optimized by the recall on the validation set. We also introduce models with an additional constraint: the target keywords must be either hashtags or 2-grams that start with uppercase (e.g., WeWant JobsNow). We observe that all the keywords obey this rule with one exception. This may be an artifact of Twitter's choice of trends. Filtering out the 1-grams and the 2-grams that start with lower case increases the model's precision with no decrease in recall. We decided to deploy this model until we observe this rule is broken frequently. Table 3.3 shows the results.

Table 3.3: The best parameters and the performance of the classifier. * refers to the models with the constraint.

Model	w	s	Prec-V	Rec-V	F1-V	Prec	Rec	F1
Collect	3	3	0.63	0.98	0.76	0.78	0.97	0.87
Announce	20	16	0.98	0.91	0.94	0.97	0.80	0.88
Collect*	3	3	0.82	0.98	0.89	0.93	0.97	0.95
Announce*	5	5	0.98	0.93	0.95	1.00	0.85	0.92

3.11.1.3 Attack Tweet Detection

In this component, we detect attack tweets and the astrobots posting them after collecting their data by search API after the real-time trend detection. We do this after filtering out the non-deleted tweets for higher precision. Our goal is to highlight the size of attacks while announcing them and collecting the bots who participated in the attacks. We now formulate the problem.

Problem Statement Given a fake trend FT, let $T = t_0, t_1...t_n$ be the tweets containing FT. Let $A = a_0, a_1...a_n \subset T$ be the tweets that are part of the attack. Determine A.

Ground Truth We collected the data of 72 trends using Search API after detecting that they were fake in real-time. We look for a series of uninterrupted lexicon tweets that promote a trend and are later deleted. We annotated 8446 tweets as attacks and 8296 tweets as negatives in total.

Experiments and Results We use unsupervised anomaly detection methods to detect anomalous volumes of tweets. The volumes are computed as the number of tweets in a time slot. We use the dates and the minutes the tweet posted as the timeslots. If there is an anomalous volume of tweets in a time slot, we classify every tweet in that time slot as attack tweets. Our caveat is that this approach ignores accounts that coincidentally post a tweet containing the fake trend during an attack. However, we found only one instance (i.e., tweet) of such behavior during the annotation process, which is negligible.

We use the data of 60 trends as the train and validation set and 12 trends as the test set. We use the validation set to find the best parameters for a given model. We evaluate our models based on several baselines, such as classifying lexicon tweets as attack tweets. We now describe the methods we use.

Lexicon: In §3.4 we introduced a text classifier so that we can detect trends targeted by astrobots. Although the classifier achieves this goal with high performance, it yields a low recall on individual attack tweets. For the attack detection task, we use it as a baseline and improve over it.

Lexicon Correlation (Lex-Corr): Our second baseline exploits the fact that the attack tweets are correlated with each other. Therefore, a tweet that comes before or after an attack tweets within *sec* seconds would also be an attack tweet. We first classify the attack tweets using the lexicon classifier. We then populate this set by classifying every tweet posted within *sec* seconds of an attack tweet as an attack tweet. We repeat this until there is no new attack tweet.

Peak Detection (Peak D.): We detect peaks in the volume of tweets by comparing each volume to their neighbors using scipy [292]. The height of each peak should be at least h tweets.

Local Outlier Factor (LOF): We preprocess the data using standard scaling. We detect anomalous volumes using Local Outlier Factor [42] implemented by sklearn [215]. We use the outlier factor (o.f.) as the parameter to be tuned using the validation set.

Isolation Forest (Iso F.): We use the same settings as of LOF, but instead use Isolation Forest [181].

Table 3.4: The Results of Attack Tweet Detection Methods with Best Parameters

Method	Parameter	Prec-V	Rec-V	F1-V	Prec	Rec	F1
Lexicon	-	0.975	0.733	0.837	**1.000**	0.482	0.652
Lex-Corr	sec=1	0.975	0.847	0.906	0.997	0.618	0.763
Peak D.	h = 40	**0.979**	0.959	0.969	0.993	0.861	0.922
LOF	o.f. = 0.05	0.956	0.838	0.893	0.968	0.766	0.855
Iso F.	**o.f. = 0.005**	0.970	**0.977**	**0.974**	0.978	**0.981**	**0.979**

We find that the Isolation Forest algorithm provides inarguably the best results in recall and F1 score. It's also more robust than the others due to better scores on the test data. We additionally remove the retweets from the data as retweets are not lexicon tweets in our case. This slightly increases the precision from 0.97 to 0.978 in exchange for generalizability. We deploy this final model.

3.11.2 Deployment

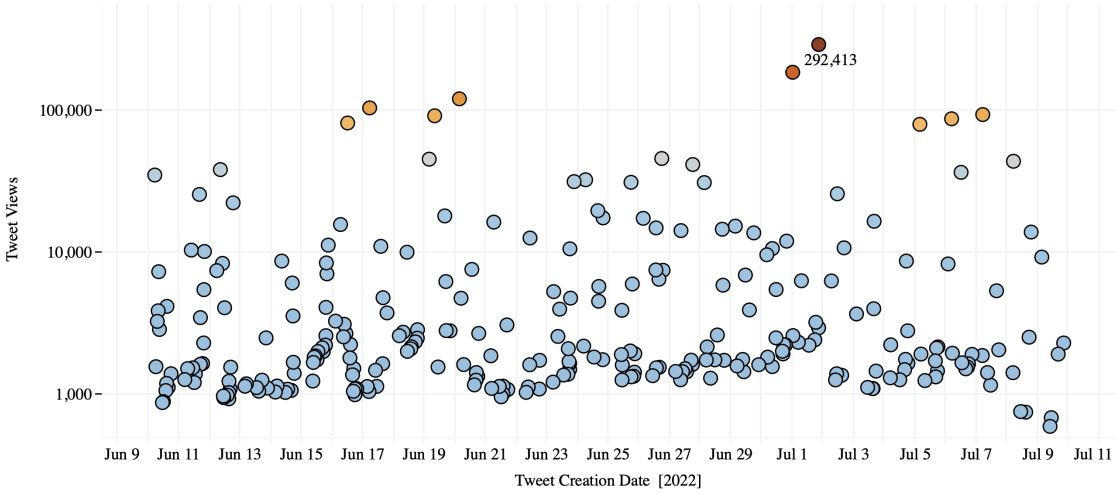

Figure 3.20: The number of tweet views per tweet announcing a fake trend. Colors closer to red highlights the tweets with high views.

We anonymously deployed this system on 23 May 2022 as a Twitter bot. By 9 July 2022, we detected 426 attacks. Our bot received public attention and gained 3000 followers without any promotion from our side. This is mostly because people could see its announcement when they clicked on the fake trends. As Twitter API v2 allows access to the number of tweet views (impressions) for the authenticating user, we were able to collect how many people we could reach. Fig. 3.20 shows the tweet views between 10 June and 9 July. Our tweets were viewed at least by 600 users. The median tweet view we have is 5500. 46 of our announcements were viewed more than 10,000. The maximum views we acquired were 290,000. It was when we announced a fake trend that reported a political sabotage attempt. The results show that our bot is successful in announcing fake trends to the public to some extent. It also implies that a large audience may view posts promoting fake trends.

3.12 Limitations

Study of ephemeral astroturfing is limited to the platforms in which the content promoted by the popularity mechanism and the deletions are made available. While working with Twitter data, we are limited to only 1% of tweets, as provided by Internet Archive's Twitter Stream Grab. Larger samples are not publicly available. This sample may not include attack tweets for every trend, so we may not be able to detect all attacked trends. Thus, we can only report lower bounds. Furthermore, trends with more data available in 1% are more likely to be detected, which makes the study biased towards attacks on a larger scale. Additionally, bots with more data available in 1% are more likely to be detected, which also biases the results towards more active bots. We acknowledge those biases and assume they will not make a huge impact on the final results. We are also

limited to local and global trends and are not able to analyze tailored (personalized) trends. The trending algorithm is black-box, and it is not reasonable to reverse engineer it using only a 1% sample. Thus, we study the attack based on the behavior we observe in the data and only develop a classifier to detect one specific, ongoing attack instance.

3.13 Ethical Implications

To detect and analyze attacks retrospectively, we used the public data provided by Twitter and the Internet Archive both of which have been analyzed extensively by previous work. We also collected tweets posted by astrobots during the real-time detection. We used their deletion notices to detect attack tweets in real-time. To protect user privacy, we use an algorithm that does not rely on the content of the deleted tweets we collected and delete those tweets after receiving the notice.

We acknowledge the ethical concerns with the honeypot account. To mitigate this, we signed up a newly created account which is normally filtered by Twitter's spam filters, and minimized the amount of time that the account was active.

3.14 Summary

In this chapter, we have defined and described a new attack on social media platforms that employ compromised accounts. We presented a case study in which adversaries performed these attacks to manipulate Twitter trends in a specific region. We have proposed a method to detect promoted keywords and used it to analyze the successes of the attacks by various metrics. We also characterize the accounts used in the attacks on their ecosystem. We discussed the implications and counter-measures. We also proposed and implemented our counter-measure which is announcing the attacks to the public in real-time. We believe our study will enhance analyses and prevention of ephemeral astroturfing attacks in other contexts.

This research was conducted using the Internet Archive's Twitter Stream Grab and trends data, so all data is public. and the study is reproducible. In addition, the IDs of the tweets and users annotated in this study as well as the annotated attacks are made available[2].

[2]https://github.com/tugrulz/EphemeralAstroturfing

Chapter 4

Implications of Compromised Accounts on Bot Research: Retweet Bots

Characterizing Retweet Bots: The Case of Black-Market Accounts

ICWSM 2022

This chapter studies and discusses the *implications* of compromised accounts used as bots. Malicious Twitter bots amplify harmful narratives to disrupt the public discourse on social media through automation. Past studies claim that bots are getting increasingly human-like and difficult to distinguish from legitimate accounts. Here, we focus on retweet bots that artificially inflate content to study such claims. We propose a new dataset of bots that have been uncovered by purchasing retweets from the black market. We characterize retweet bots for the first time. We found evidence that those accounts are mass-compromised. We also analyze their differences from human-controlled accounts. From our findings on the nature and life-cycle of retweet bots, we point out several inconsistencies between the retweet bots used in this work and bots studied in prior works. Our findings challenge some of the fundamental assumptions related to bots and in particular how to detect them.

4.1 Introduction

An extensive amount of research has focused on detecting automated accounts, *bots*, on Twitter. While some studies directly address bots' functions, such as spamming [313], pushing slogans on top of Twitter trends [104], and inflating follower counts [68], others use the general term *social bot* to mean *accounts that mimic humans* without addressing how the accounts are automated. By only considering generic social bots and their detection, we cannot learn about the actual nature of the bots themselves.

Retweet bots are the bots whose primary functions is retweeting others in an automated manner, often as a result of some commercial activity. While there are studies that detect coordinated groups of accounts, of which retweet bots may be a part, there

is no work that tries to understand what a single retweet bot looks like, how it behaves, and how it differs from a human account. If a user is a dedicated fan of a football club and, thus, uses their account only to retweet that club's Twitter posts, what makes them different from a retweet bot?

Understanding the nature of retweet bots is challenging in the absence of reliable ground truth on retweet bots. The standard method for labeling bots, annotating retweet bots by hand, is inexact. As illustrated by the football club fan, determining whether an account is controlled by an algorithm or a human is not straightforward. Using human annotation is error-prone and can lead to results that are biased by the annotator's assumptions about what a bot is.

We address this by studying retweet bots whose services were directly purchased by prior work. Using this dataset of reliable retweet bots, we also examine both the assumptions and findings made by prior works and find some that do not hold up against the bots in this dataset. This allows us to observe inconsistencies in how bots behave. Particularly, we found that these bots were *compromised* and we did not observe that the adversary put effort to pass them as humans after they take over the accounts. From this finding, we also study and discuss the implications of compromised accounts on bot research.

We focus on four research questions:

1. Where do these retweet bots come from? Were they created for this purpose or compromised? (§4.4)
2. What is the lifetime of these retweet bots? (§4.5)
3. How do the retweeters in our dataset act differently from human users? (§4.6)
4. Are there any differences between the bots examined in this work and those found in prior studies? (§4.7)

In answering these questions, we 1) present the first study focusing on retweet bots exclusively; 2) characterize retweet bots, providing evidence that some are mass-created and controlled by one center entity while some are compromised and used aggressively; 3) challenge fundamental assumptions about the nature of bot accounts, such as account age and over-activity; and 4) we discuss challenges in bot detection with respect to retweet bots that are compromised.

4.2 Related Work

The literature on the detection and analysis of bots principally defines and annotates bots either by their *nature* or their *primary function*. The popular term "social bot" in reference to "bots mimicking human behavior" is an example of the former [40, 114], which are usually reliant on human annotation, which we know to be unreliable [73]. The latter, e.g., spammers [60, 143, 313], fake followers [68], and astroturfing bots [104],

are usually less reliant on human annotation since the function of an account is more straightforward to define and detect based on specific behavior. The bots we focus on in this work are distinguished by their primary *function*: retweeting other accounts. We will refer to these accounts as *retweet bots*.

In order to forgo error-prone machine or human-based detection, some studies [68] opt for the more reliable method of directly purchasing a bot service. Past work has shown that this method can be used to study bots that inflate follower counts and to analyze the authors of such posts [87, 89]. We follow this more reliable method of bot collection and analyze the *accounts* controlled by a vendor (or vendors) who sells retweets.

Prior work on retweet bots is not devoid of research on accounts, however, these studies focus on coordinated groups of accounts [182, 288], including those who retweet others [135, 191], and not on individual accounts. By studying individual accounts we learn how the accounts became retweeters, how long they were active, and how they were different from genuine accounts. Such studies are also constrained by a single topic (e.g., finance [71]) or by an assumption (e.g., that bots always act at roughly the same time [55] or that bots' timelines are similar [70]).

To the best of our knowledge, there are only two prior works that provide a per-account analysis of retweet bots. Unlike the data presented in this chapter, both rely on human annotation. Dutta et al. [90] took a much more restrictive definition of *retweeter*, effectively only studying trend spammers and not a broader population of accounts. They also leverage human annotation which partially relies on whether or not a large number of tweets/retweets were posted within a short time period, something we find is not the case with the retweet bots in our dataset. Giatsogloue et al. [126] studied both retweeted posts and retweeters and proposed a retweeter classification method. The overlapping observations between our work and this are that 1) retweeters have a high follower friend ratio, in contrast to popular belief that they do not, and 2) they retweet with similar time delays. We expand on these observations and explore further.

4.3 Dataset Overview

We use two distinct types of data for this analysis: 1) data from retweet bots and 2) data from human (genuine) accounts which we use as a control group.

4.3.1 Retweet Bots

The retweet bot dataset is made up of bots that retweet others' tweets. The dataset was introduced in a paper by Golbeck [129]. In this work, the author created a fake and "uninteresting" Twitter account with no followers and posted "uninteresting" tweets in order not to attract genuine retweeters to the tweets. They then contacted vendors selling "retweet services" and purchased 100 retweets for each post. The goal of this study was to detect whether a post had been promoted by retweet fraud or genuinely

received the retweets. We build on this work by exploring the accounts that participated in the retweeting activity.

To build this dataset, we started with the 18 "uninteresting" tweets included in the Golbeck study and collected all of the accounts that retweeted any of these tweets. Due to the 3.5 year gap between the original study and this study, most accounts were either suspended or inactive, leaving only 862 non-suspended accounts. Although these accounts were not suspended and therefore Twitter has not flagged them as bots, because the activities (retweets) of these bots were directly purchased, we can be certain that these 862 accounts were in fact acting as retweet bots during the time of Golbeck's study.

To extend this dataset, Golbeck extended the search for accounts laterally. That is, for each bot A that retweeted one of the "uninteresting" tweets, find all of the other posts, e.g., t, that A retweeted. Because A only retweets posts they are paid to retweet, we know that the author of t paid for A and other bots to retweet it. Therefore, we can reasonably assume that *some* of the other accounts retweeting t are also retweet bots and they are recorded as such. This yielded a dataset of 6,112 accounts, 5,332 of which were suspended. We discarded the 780 non-suspended accounts from this study since there is some uncertainty in this collection method and, at least so far, Twitter does not suspect that these accounts are fake, so they may be genuine. We keep the suspended accounts since Twitter has more or less corroborated these results. We do not take for granted that these are all retweet bots. We explore later whether the suspended accounts are genuine or were suspended for some other reason, but find that this is very unlikely the case. Our final dataset consists of 6,199 accounts. Most of the analysis in this chapter is focused on the 862 accounts whose actions were directly purchased and were not suspended (so we have the full data from each account).

Using the Twitter API, we collected the timeline (most recent 3,200 tweets) from all 862 non-suspended accounts in October 2020. We refer to this dataset as the *timeline* dataset. This dataset consists of 1,212,030 retweets and 125,974 tweets. Some of the bots in the timeline dataset have few retweets, e.g., 48 bots have less than 100 retweets. One likely explanation is that there were more retweets, but they have since been deleted. We include all of the accounts in the dataset and the analysis regardless of retweet count.

Collecting the data from the 5,332 suspended accounts was more challenging. Internet archive's Twitter Stream Grab provides 1% of all tweets since 2011 [1] and has been used extensively by past research [98, 101, 261]. By mining this dataset, we collected roughly 1% of all tweets from these accounts and their profile information. We call this dataset the *archive* dataset. This dataset consists of 301,932 retweets and 29,899 tweets. Fig. 4.1 shows the histogram of the number of tweets and retweets per user for each dataset.

The main difference between the archive accounts and timeline accounts is that the former are suspended while the latter are not. This leads to differences in characteristics of the accounts, which we will discuss in the next section.

The datasets are made available for reproducibility[1].

[1] https://github.com/tugrulz/RetweetBots

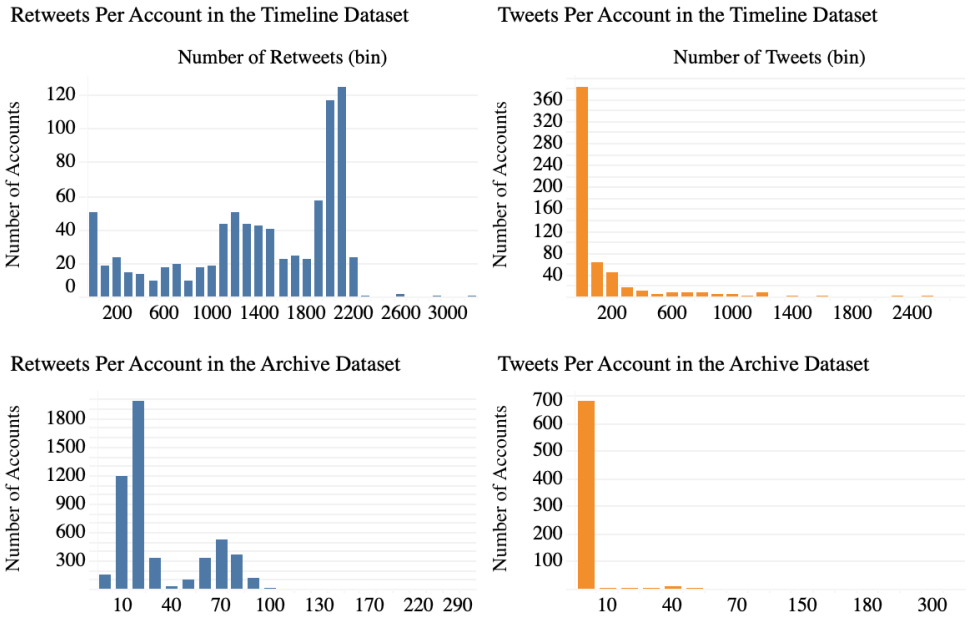

Figure 4.1: Number of retweets (left) and the tweets (right) per account in the timeline dataset (top) and the archive dataset (bottom). As the timeline dataset is collected using Twitter API, it consists way more tweets and retweets than the archive dataset. However, there are more accounts in the archive dataset.

4.3.1.1 Control Groups

Understanding the differences between retweet bots and human accounts requires a control group of human (genuine) users. For this, we rely on multiple datasets of accounts that have been annotated as human-controlled in previous studies [68, 69, 70, 71, 127, 191, 289, 312][2]. We collected the timeline (most recent 3,200 tweets) of 27,622 users in 2021. The datasets' statistics are summarized in Fig. 4.2.

In some cases, our labels, which are reliable because their activity was directly purchased, do not agree with labels in other datasets. In one such dataset of human-controlled accounts [71], which was labeled via classification, 664 and 37 bots from the timeline and archive datasets respectively are labeled as bots. However, this dataset also labels 133 bots from the timeline dataset and 1 from the archive dataset as humans. We considered these users as bots and excluded them from the humans dataset. We do not use any datasets of other types of bot accounts in the control groups because no prior dataset provided a differentiation between retweet bots and non-retweet bots.

4.4 RQ1: Nature of Retweet Bots

We first focus on meta-data analysis of the dates in which the accounts were created. We then focus on the content of the accounts and their tweets to find indicators that an

[2]https://botometer.osome.iu.edu/bot-repository/datasets.html

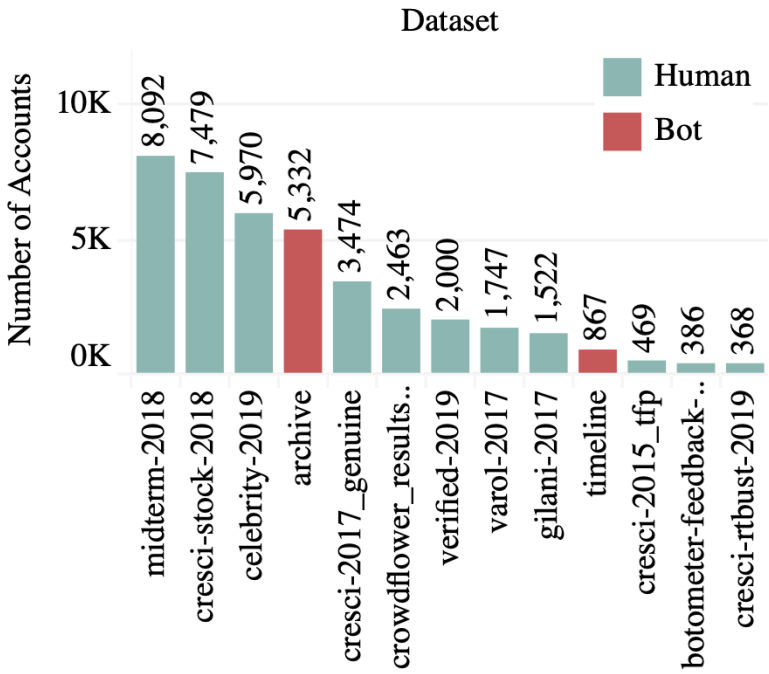

Figure 4.2: Dataset statistics. Each bar represents one dataset that we used in this work. The *bot* bars are datasets that we built for this chapter based on work by Golbeck [129], and the *human* bars are genuine accounts from previous studies.

account has been compromised. We find that the accounts in the archive dataset were illicitly created for this purpose, but that accounts in the timeline dataset were more likely compromised normal accounts, hence they are not suspended like the accounts in the archive dataset.

4.4.1 Evidence of Mass Creation

Many studies analyzing social media manipulation focus on accounts created around the same date and new accounts as signs of deliberate manipulation. This is based on the assumption that accounts that are created at the same time are likely to be controlled by one entity, and as such are fake (vs. either genuine or compromised).

We followed this assumption and computed the number of accounts created per day. A histogram is shown in Fig. 4.3 to illustrate this. *We found several periods in which the accounts in the archive dataset, but not in the timeline dataset, were bulk created.* The most significant period was between the 18$^{\text{th}}$ and 21$^{\text{st}}$ of October 2013 in which 3,750 accounts (70.3% of accounts in this dataset) were created.

4.4.2 Evidence of Mass Compromisation

The second potential origin of the accounts in this dataset is that they began as normal accounts and then were later compromised and used as retweet bots. While the accounts in the archive dataset were created at once, *the accounts in the timeline dataset appear to be compromised.*

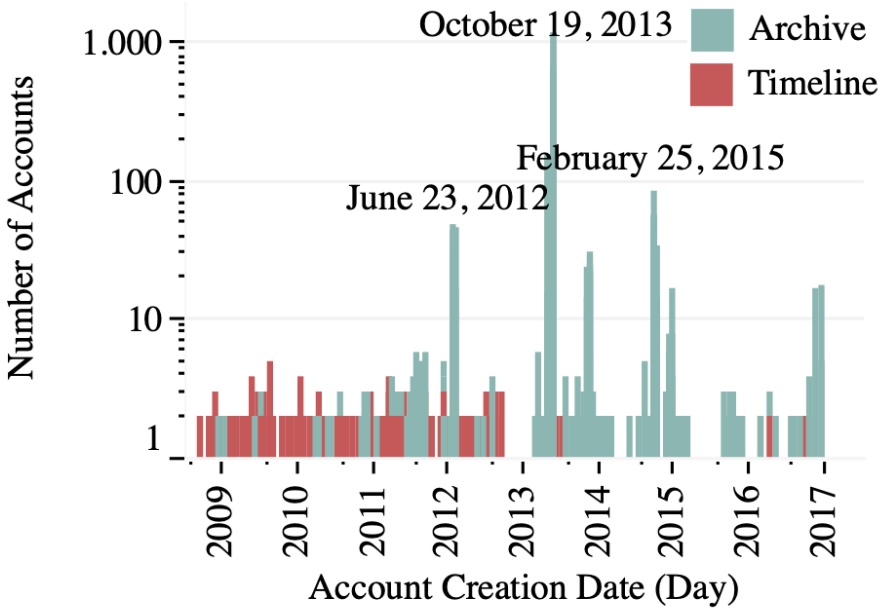

Figure 4.3: Creation dates of accounts in the timeline and archive datasets. Most of the accounts in the archive dataset but not the timeline dataset were created in bulk.

We find evidence of this by analyzing the *content* of the tweets. That is, we compared the tweets in the timeline dataset authored during the suspected time of compromise (March-August 2017, see Fig. 4.6) to those authored before this period. There were 4,708 tweets from 184 users during the retweeting period, while there were 31,847 tweets by 322 accounts before March 2017.

To highlight the differences between the tweets in these two different periods, we use wordshift graphs [120]. Wordshift graphs compare two corpora and rank the words by their contributions to the differences in these two corpora. The contribution is computed using Shannon entropy. We randomly sample 10 tweets per account in order to give equal weights to all accounts and create two corpora, one for before the period of retweets and one during. Fig. 4.4 shows the top words for each. We manually inspected the tweets containing these words.

We find the most evidence from the words that were prevalent during the retweet period, but not before, in the timeline dataset (the top left of Fig. 4.4). Similar to prior work [318], we found that some users were posting tweets that directly stated that their accounts had been "hacked" and are now recovered. We see this in the prevalence of the words *hacked* and *account*, as users recover their accounts over this period. The substring *hack* was present in 134 tweets, and we manually verified that 42 of them (by 33 users) stated that the user's Twitter account was hacked. Some users complained about the retweets from their account, e.g., *My account got #hacked and I've tried to dlt all the retweets, but it still says I have over 2,000 tweets. How do I get rid of them? #help.* Some users also announced that they were leaving their account because it was compromised and urged their followers to follow their new account. One user even changed their name to *hacked* and changed their description to their new handle.

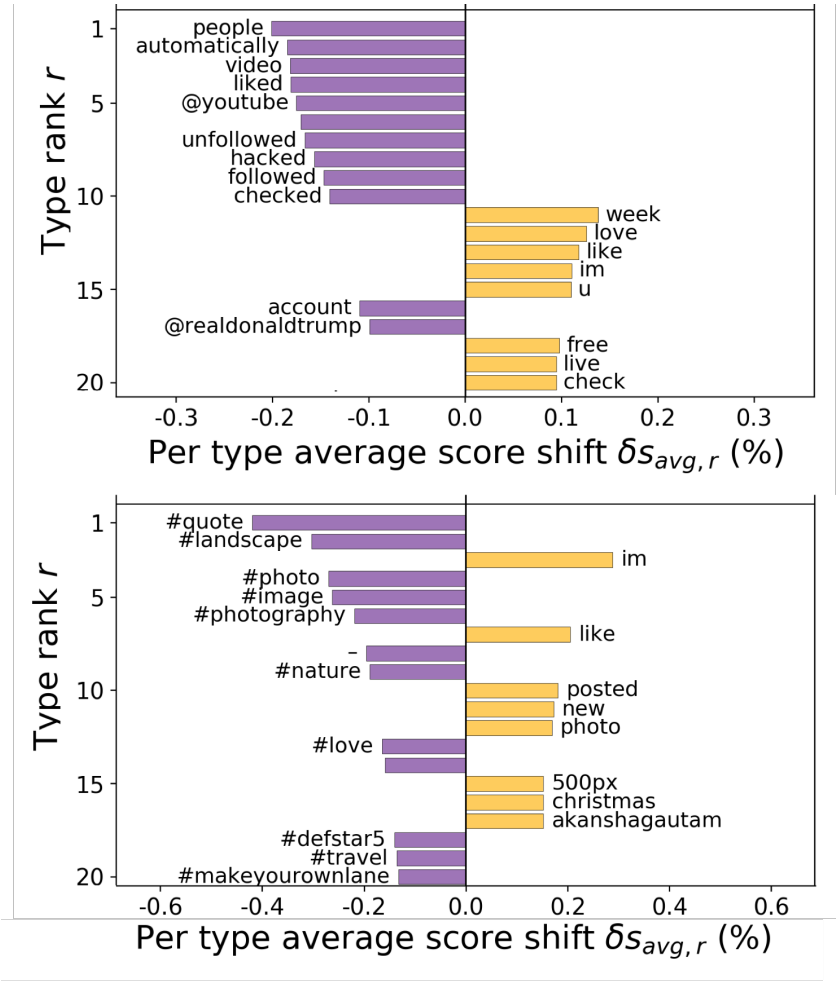

Figure 4.4: The wordshift graphs of accounts in the timeline dataset (upper) and archive dataset (lower). The words on the *left* represent the tweets during the period of retweet activity and those on the *right* represent tweets from before.

Alongside *hacked* and *account*, we also find *people*, *automatically*, *followed*, and *unfollowed*, all words posted by a common spam app that reports how many new accounts followed/unfollowed the account owner, i.e. *x people followed/unfollowed me//automatically checked by...* The words *video*, *liked*, and *@youtube* come from a Twitter app that posts the users' activity on Youtube to their Twitter feed. It is not clear if the users actually signed up for this service or if it is due to the account being compromised. The popular words *love*, *like*, *im*, etc., distinguish the two periods. This is likely because the tweets authored by the users during the period of retweets are mainly automated messages from a script so they do not contain such otherwise popular verbs.

We do the same analysis on the suspended users in the archive dataset. The retweet period for these accounts was longer and the edges less defined. We analyze the period in which at least half of these accounts were actively retweeting: April 2015 to October 2017. There were 10,196 tweets by 484 users before the period and 7,054 tweets by 138 users within the retweet period. Inspecting the wordshift graph, we find no patterns as

we did in the timeline dataset. There are no tweets by the spam app, only two users posted Youtube activity, and we did not identify a single user complaining about getting hacked. We did find other patterns: 17 users shared quotes from famous people, and the hashtagged words (e.g., *#landscape*) were used to share a blogpost containing photos accompanied by a quote (these appear to be part of a promotion). Cresci et al. [69] report a similar behavior, finding that novel social bots share quotes to appear genuine. Note that these hashtags are the most popular hashtags on Instagram[3] which may be the reason why these accounts target them as genuine accounts often use them as well. In conclusion, we believe that those users are fake rather than being compromised and their non-retweets are also used for promotions.

4.4.3 Are Any of the Accounts Genuine?

Prior work [89] has found that there are genuine users who sign up to blackmarket schemes to retweet others in exchange for retweets of their own content. These accounts become part of an illegitimate scheme called collusive retweeting. *We find no evidence of this happening in the accounts analyzed for this chapter.* Inspecting the *favorite counts* and *retweets counts* of the accounts' tweets, we see that most accounts do not get attention from other accounts, as would be the case if such a scheme were present in these datasets. Most of the accounts in both the timeline dataset (58%) and the archive dataset (98.2%) received no retweets. Even fewer accumulated between 1-10 retweets: 25.8% in the timeline dataset and 0.5% in the archive dataset. Finally, just 24 accounts in the timeline dataset and 36 in the archive dataset received at least 100 retweets in total. We see this mirrored for favorites counts. Our caveat is that the users who participated in collusive retweeting may have removed their retweets before data collection, therefore hiding this activity.

4.5 RQ2: Lifetime of a Retweet Bot

Now that we have an understanding of where these accounts originated, we try to understand their lifespans. We investigate the time and the volume of the activity of the accounts. *We observe that the accounts in our dataset were overactive in a short time period and were otherwise idle, even though they were not suspended.*

We found that 816 (94%) of the accounts in the timeline dataset were active between March-August 2017, peaking in August with 758 accounts as seen in Fig. 4.5. Note that we were not able to collect the activity of 263 accounts before March 2017 due to API limitations. The number of active accounts gradually falls to 379 on August 25th and then suddenly to 17 on August 26th. It then never exceeds 40. We make the same observation for the volume of retweets. As seen in Fig. 4.6, most of the retweets were posted between March 2017 and August 2017, peaking on July 30, 2017, with 14,069 retweets. The plausible explanation for this is that some malicious actor took possession of these accounts in this period and used them aggressively to retweet others. Twitter

[3]https://influencermarketinghub.com/most-popular-instagram-hashtags

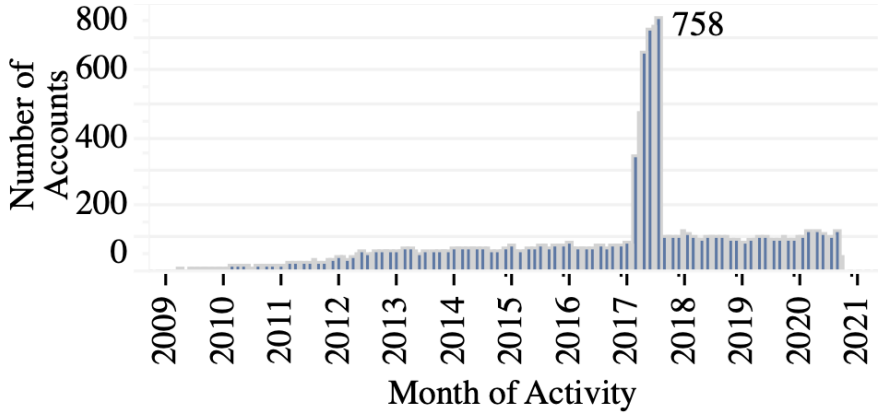

Figure 4.5: The number of accounts active in the timeline dataset per month. Most of the accounts were active between March 2017 and August 2017.

then may have imposed a "softban" on these accounts without actually suspending them, e.g., by asking for phone verification.

We made a similar observation with the suspended accounts from the archive dataset. As seen in Fig. 4.7 the number of active accounts peak in Summer 2017, drops abruptly from 2,544 in September 2017 to 842 in October 2017 and then later to 81 in February 2018. However, the accounts were active for a longer time period than those in the timeline dataset, i.e., there were already 2,170 active accounts in April 2015. Unlike the timeline dataset accounts, these were suspended by Twitter. What is unclear is whether they went inactive due to the suspension or whether they became inactive and were later suspended.

These results suggest that the retweeters we analyzed stay active for 6 months to 2.5 years. Although this may not generalize to all retweets bots, it gives an indication of how long a retweet bot may be active.

4.6 RQ3: Retweeters vs. Humans

We next analyze how retweet bots differ from humans in terms of the time and volume of their activities. Table 4.1 summarizes the results. Specifically, we focus on the following patterns: 1) volume of activity (§4.6.1), 2) percentage of retweets (§4.6.2), 3) time between consecutive retweets of the same user (§4.6.3), 4) time between consecutive retweets of the same post (§4.6.4), 5) time between the retweeted post and the retweets by the retweeter accounts (§4.6.5) 6) diversity of retweeted users (§4.6.6). The percentage of retweets and time between consecutive retweets of the same post clearly demonstrate the retweet bot activity while the other activities have patterns that challenge some of the fundamental assumptions in the bot literature. Each pattern is described by a different measurement. We compute the mean of that measure for each group (humans and bots) and report the difference in terms of that measure. We use Welch's t-test to test the statistical significance of the difference. Welch's t-test is best suited for this task

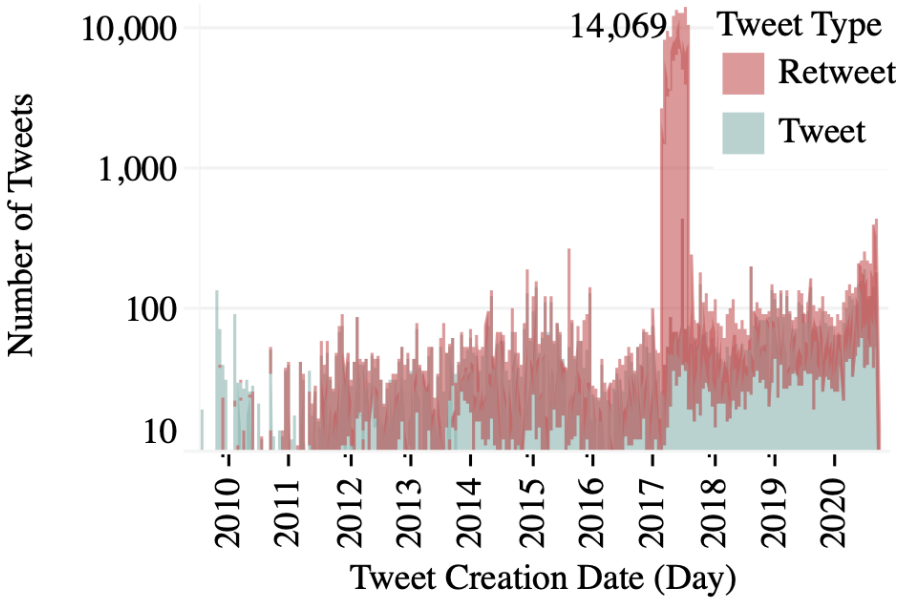

Figure 4.6: Number of tweets and retweets per day by the accounts in the timeline dataset. The accounts retweeted aggressively between March 2017 and August 2017, despite the low number of original tweets.

as it does not assume equal variance between groups. For patterns in §4.6.1 and §4.6.5, we compute additional measures. Two of the measures report the percentage of accounts with a binary property within each group. We report the difference in percentages and apply a chi-squared test to test the statistical significance.

Section	Measurement	Human	Bot	Diff
§4.6.1	Mean status count	6.1k	24.5k	18.4*
§4.6.1	% of accounts with >50 tweets/day	42%	16%	26%**
§4.6.1	Mean of max # daily tweets	60.6	40.2	20.4*
§4.6.1	Mean # daily tweets	9.7	12.2	2.5*
§4.6.2	% of RTs	34%	91%	57%*
§4.6.3	Mean of per-user median time between RTs	253	69	184*
§4.6.4	Mean of per-post mean time between RTs	1,469	3.4	1,465.5*
§4.6.5	% of RTs within 1 min.	4%	1.25%	2.75%**
§4.6.5	Mean retweet delay	313	246	67*
§4.6.6	Mean Diversity	0.50	0.52	0.2*

*p <0.0001, Welch's t-test, **p <0.0001, chi-squared test

Table 4.1: Summary of the quantitative differences between retweeter bots and humans. All were statistically significant.

4.6.1 Volume of Activity

We begin by analyzing the overall activity of retweet bots and human accounts. Fig. 4.8 shows the statuses count (the total number of tweets and retweets) for each account in each dataset. Humans were more active overall, and their statuses count follows a power

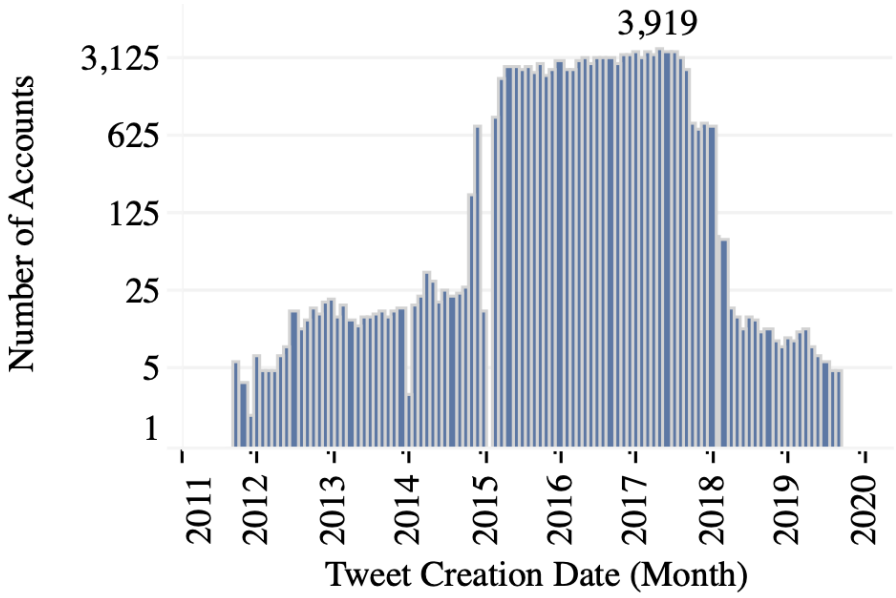

Figure 4.7: Number of accounts active per month. Most accounts were active between April 2015 and September 2017.

law. Meanwhile, the statuses counts of the bots in the timeline dataset are unimodal, and those in the archive dataset are bimodal. The latter is likely due to the presence of multiple vendors. The difference between the mean statuses count for bots (6,153) and humans (24,560) is 18,407.

We also find that the human accounts in our dataset were more likely to be overactive than bots when considering daily activity. The average number of daily tweets for bots and humans is similar, 12.2 and 9.7 respectively, however, bots tended to retweet more consistently, as we see in the variance of the number of daily tweets, 21 and 370 respectively. We see this inconsistency as well when we look at the maximum of statuses posted in a single day, i.e., the number of statuses each account posted on their most active day. Human-controlled accounts posted an average of 60.2 statuses on their most active day. This figure is 40.2 for bots.

As seen in Fig. 4.9, 70% of retweet bots had a maximum between 25-40, which is either due to a threshold set by the vendor or is constrained by the number of tasks the account vendor receives every day. Of the human accounts, 42% tweeted at least 50 times in a single day (a threshold used by Howard et al. [146] to indicate an overactive bot), while only 16% of retweet bots were this active.

4.6.2 Percentage of Retweets

One major difference we expect to see with respect to these bots in particular is their retweet activity. We compute the *retweet percentage* for each account, i.e., the share of retweets in all statuses the account posted. The overall average retweet percentage for bots is 91.3% and for humans is 34.5%. Fig. 4.10 demonstrates how stark this difference

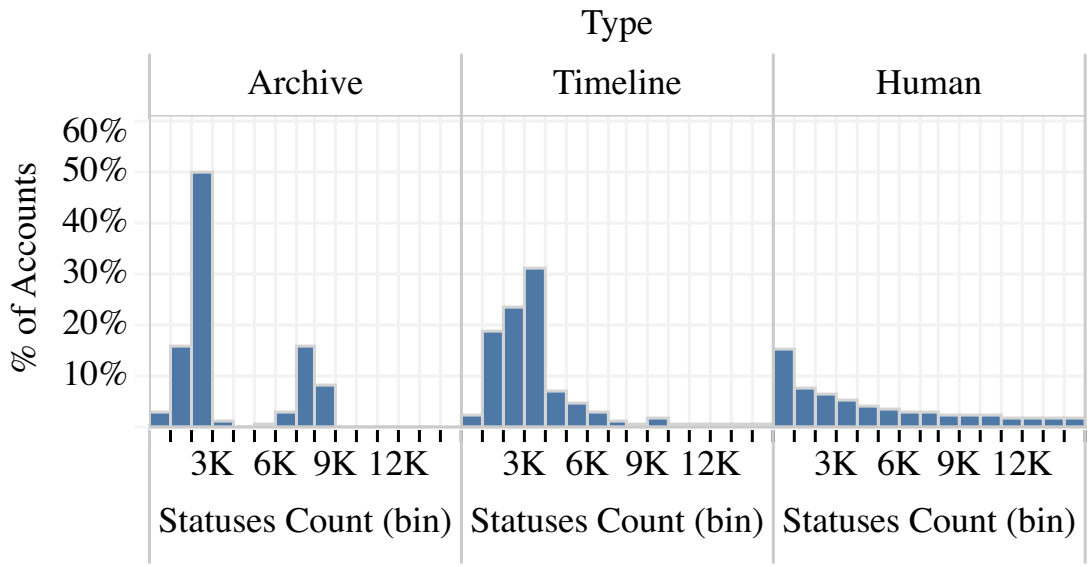

Figure 4.8: Histogram of status counts per day per account. While the humans' status counts distribution follows a power law, the bots' are unimodal/bimodal, concentrated between 1,000 and 3,000, with the accounts in the archive dataset having another focal point between 7,000 and 9,000.

is: 86% of retweet bots vs 9% of humans have a retweet percentage of at least 80% and 95% of retweet bots vs 28% of humans have a retweet percentage of at least 50%.

Upon further inspection of the human accounts with a high retweet percentage, we found that a majority come from just two datasets: 43% of the human accounts with a retweet ratio greater than 50% come from cresci-stock-2018 and 22% from midterm-2018. Additionally, cresci-rtbust-2019 contains a high percentage of such accounts. The accounts in these datasets were collected using topics that are vulnerable to manipulation (i.e. elections and financial campaigns). Midterm-2018 and cresci-rtbust-2019 were annotated by hand and cresci-stock-2018 by classification.

4.6.3 Time Between Consecutive Retweets of User

We further find that bots are not overactive in shorter periods, as it may be expected that they retweet in bulk and then stay idle. Fig. 4.11 shows the median time between consecutive retweets of each account. Bots waited roughly 60 minutes between consecutive retweets while humans tended to retweet more rapidly. Note that this is regardless of their retweet count; the bots with a very high retweet count also have the same statistic. The vendor controlling the accounts may be setting sleep times or may be limited by when the requests are made. The average of the median time differences between consecutive retweets is 69 minutes for bots and 253 minutes for humans.

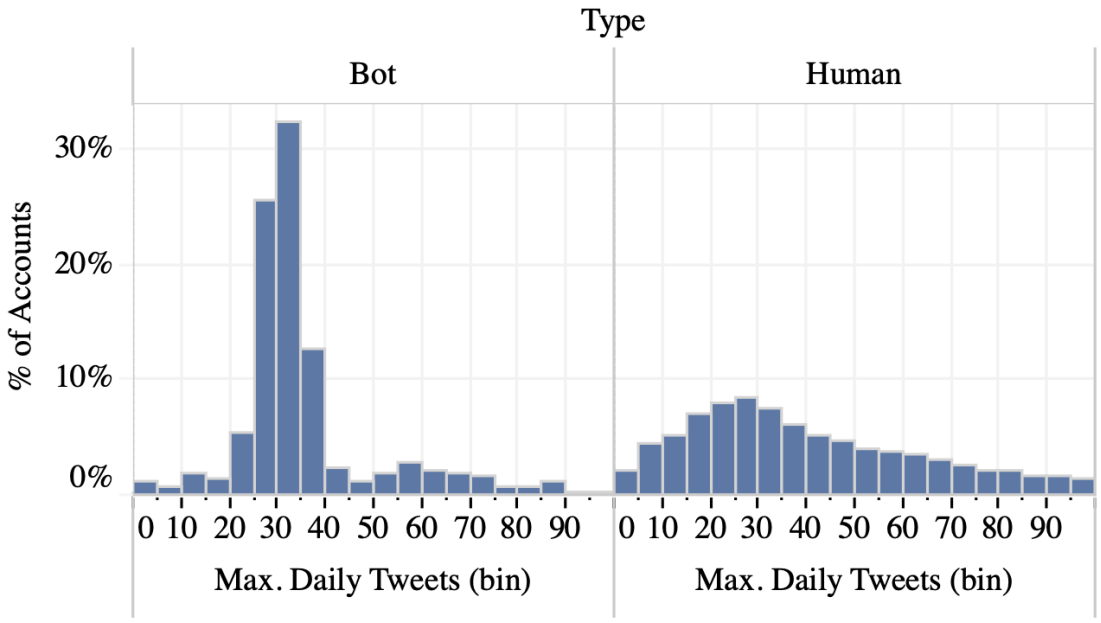

Figure 4.9: Maximum number of daily tweets per account. This is concentrated between 25-40 for accounts in the timeline dataset. Meanwhile, humans are more likely to be overactive and reach more than 50 tweets per day.

4.6.4 Time Between Consecutive Retweets per Post

Since it is the job of a retweet bot to promote posts, one might expect that while they tend to have a long delay between their own retweets, they retweet a new tweet quickly after it is posted. Fig. 4.12 shows the time between each retweet of tweets that were paid to be promoted by the bots. This time difference is roughly within five seconds, and no outlier exceeds 50 seconds.

To compare this finding with organically retweeted posts, we collected all 17,456 tweets that were retweeted by human users in our control group and had retweet counts similar to those of the bot-promoted posts (between 40 and 70). We compute the mean time between each retweet. As Fig. 4.13 shows, these figures are low for both groups, however, humans have more outliers due to late retweets, so the interval of the mean time between retweets is higher. The average time between retweets is 3.4 seconds for posts promoted only by bots and 1,469 seconds for posts by humans. Retweeters retweet at roughly the same time, while humans tend to retweet with delays.

4.6.5 Retweet Delay

Humans are able to retweet a post as quickly as they see it (e.g., after receiving a notification), but this may not be the case with retweet bots whose services are sold on the blackmarket. Fig. 4.14 shows the retweet delay, i.e., the time between the tweeting being posted and the time of the retweet, for retweet by humans and bots. Humans were much more likely to retweet a tweet within the minute it was posted, while retweet bots were more likely to wait. The accounts most likely to retweet within a minute were

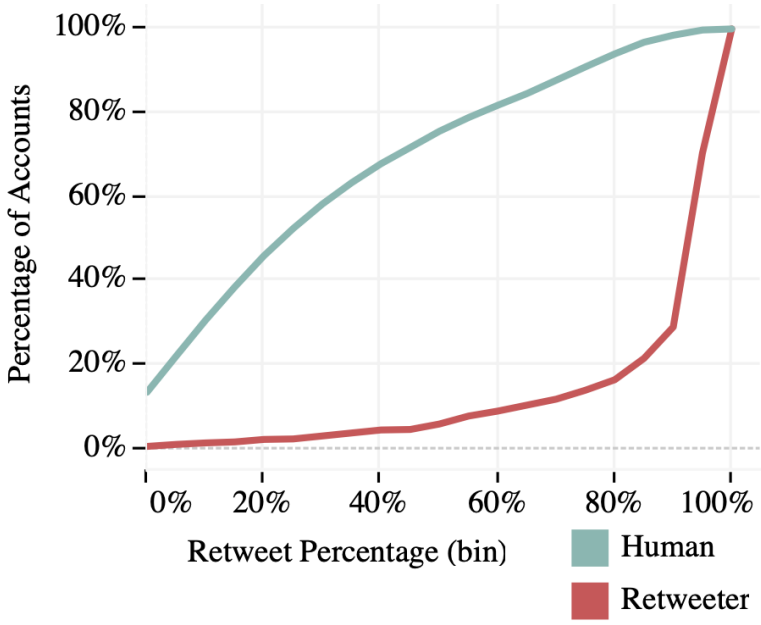

Figure 4.10: Cumulative frequency distribution of bots (in the timeline dataset) and human accounts according to their percentage of retweets. On average, retweeter accounts have a higher percentage of retweets.

celebrity accounts (5.4% of their retweets are within a minute) and verified accounts (6.4%). These only account for 2.3% of accounts in the archive dataset and 1% of accounts in the timeline dataset. Overall, 4% of humans' retweets and 1.25% of bots' retweets were posted within a minute.

While the median time difference between the original post and the retweet per account is more or less uniformly distributed (but gradually declining) for humans, it is concentrated between 8-18 minutes for suspended accounts in the archive dataset and 60-80 minutes for accounts in the timeline dataset as seen in Fig. 4.15. Overall, bots retweet faster, with an average retweet delay of 246 minutes vs 313 minutes for humans.

These results may suggest that while humans who are active on Twitter can retweet a tweet within a minute, there is always a delay for vendor-purchased retweet bots. This is either because the vendor places a deliberate delay or because it takes at least a minute for the customer and/or the vendor to send a retweet command to retweet bots. The fact that the median time difference of retweets exceeds 8 minutes suggests that some customers do retrospective commands, i.e., they first wait for their posts to get genuine attention and only purchase retweets if they fail to.

4.6.6 Diversity in Retweeted Users

While humans follow a set of people and, thus, are likely to retweet the same set of users, we expect retweeters to retweet from a diverse set of whichever accounts recently paid them. We compute diversity by diving the number of unique users an account retweeted by the number of retweets, shown in Fig. 4.16. We observe a normal distribution for all human datasets except for the cresci-stock-2018, which claims to be made up of human

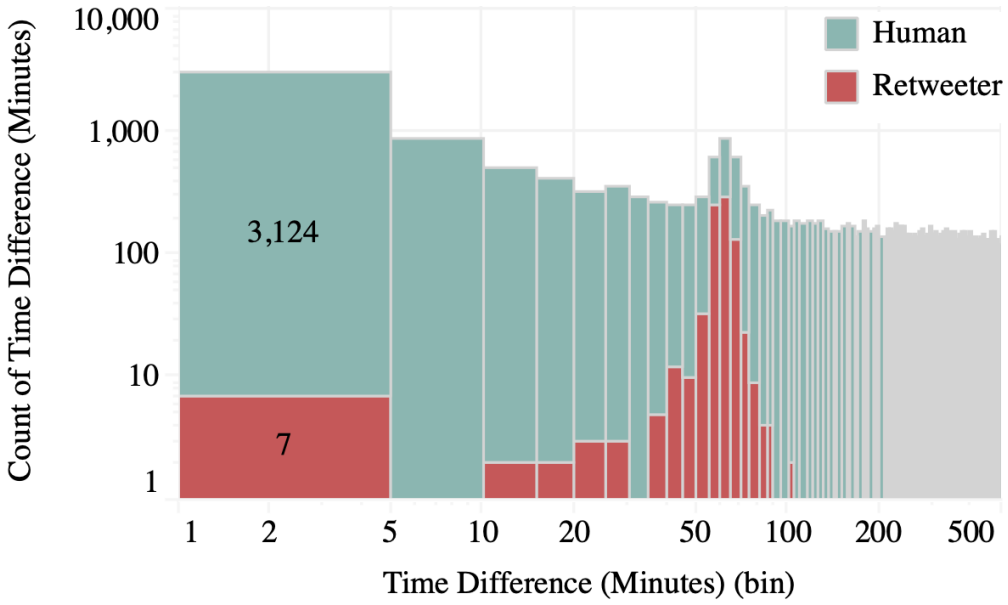

Figure 4.11: Median time difference between consecutive retweets per user. Retweet bots are more likely to stay idle between retweets.

accounts with many retweets and has a distribution similar to the retweet bots (we have excluded the incorrectly included bots).

4.7 RQ4: Differences to Prior Studies

Many bot detection methods utilize machine learning classification based on labeled datasets of bots. Often these labels come from human annotators or rely on other signals such as accounts that were banned by Twitter. Using machine learning classification for such a task has several drawbacks, but one, in particular, is that they rely on features found in training data that are inherently unstable. In this chapter, we utilize datasets collected without the use of machine learning and instead rely on observing directly purchased behaviors and lateral propagation. In this section, we explore some deviations from the narratives and assumptions found in prior work that impact these features and hint at the unreliability of such classification methods. In some cases, these comparisons are difficult to make because many studies neglect to perform any analysis of the features used and instead rely on the classifier to determine the feature importance. That is, as long as the classifier performs well, the direction in which the feature was predictive (or even if it was predictive) is not considered.

4.7.1 Delayed Activity

One common feature found in prior bot detection studies [35, 68, 289, 307, 312] is account age. The assumption is often that new accounts are more suspicious than older, more established ones. However, in both datasets, we found that the high activity period of the accounts occurred years after they were created. Most of the accounts in the archive

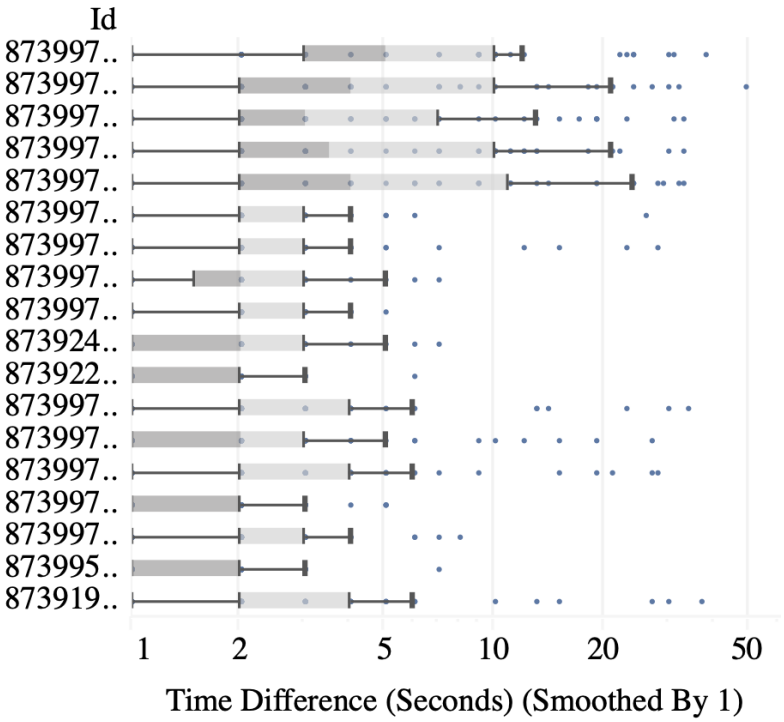

Figure 4.12: Time difference between consecutive retweets per post. The time differences do not exceed 50 seconds. The time difference is smoothed by 1.

dataset were created in 2013 and became active in April 2015. Meanwhile, accounts in the timeline dataset were mostly created between 2009-2013 but only became active in March 2017. Given the length of these delays, it is likely that the account owners are strategically not using freshly made accounts for malicious activity to avoid suspicion and thus detection by Twitter. This finding does not imply that creation date is not a signal for fake account detection, only that there are accounts that do not have this signal.

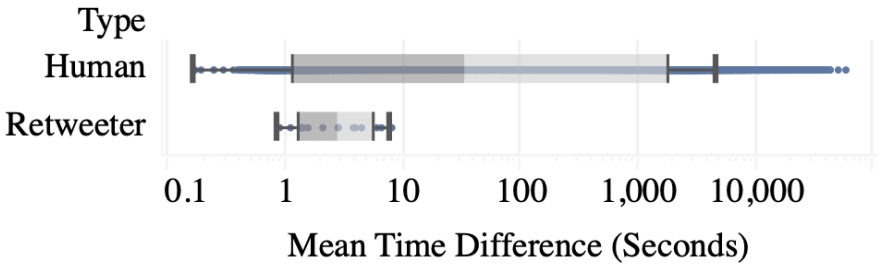

Figure 4.13: The mean time difference between consecutive retweets per post. It is lower for posts promoted by retweets when compared to humans.

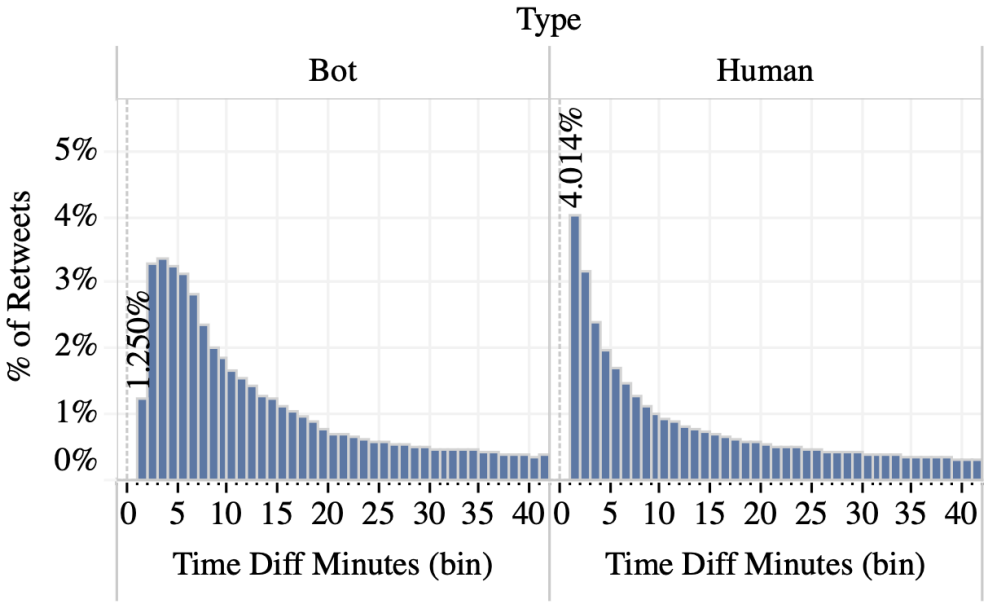

Figure 4.14: Time difference between the retweeted posts and retweets. Bots are more likely to have a small delay than humans who react more quickly.

4.7.2 Volume of Activity

Some studies use the volume of activity as an indicator of fake accounts. Howard and Kollayi [146], classified overactive accounts (at least 50 tweets/day) as bots. Similarly, Dutta et al. [90] annotated accounts as retweeters if they retweeted many tweets in a short time. In answering RQ2, we extensively analyzed the volume of activity of the accounts in both datasets. We found that the maximum daily activity for retweeters was between 25-40, much lower than the threshold of 50, and even lower than humans. Our analysis corroborates prior work [121] claiming this assumption is faulty.

4.7.3 Retweet Delay

In answering RQ3, we learned that the bots in our dataset were on average slower or as slow as humans in retweeting a new post. Prior work [191] introduced normal and suspicious retweeting patterns. They define a normal pattern as one with a delay based on the assumption that the fact that people see their timeline in reverse chronological order introduces delays in retweets. That is, a normal distribution centered at 100 minutes. They define a suspicious pattern in which users retweet in a matter of seconds. Our analysis contradicts this analysis, suggesting that humans are (even more) likely to retweet in seconds and their retweeting activities do not have long delays.

4.7.4 Diversity

Some studies [70, 191] find that the retweet bots they analyze, even if they take precautions to appear genuine, exclusively retweet from one or a small subset of accounts. In

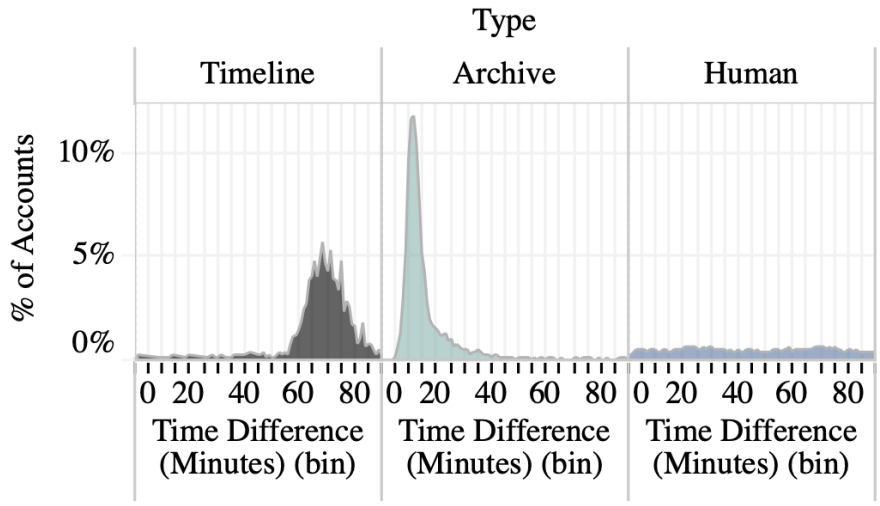

Figure 4.15: Median time difference between the original post and the retweet per user. It is concentrated between 8-18 minutes for accounts in the archive dataset and 60-80 minutes for accounts in the timeline dataset.

answering RQ3, we found that, on the contrary, they maintain a diverse set of accounts to retweet from.

4.7.5 Friends and Followers

Some prior work assumes that bots have a low follower to friend (i.e. users the bot follow) ratio because 1) they are only used to inflate follower counts of others [68], 2) they follow many accounts in order to receive "followbacks" [58], and/or 3) they fail to gain followers because they are illegitimate and thus have uninteresting profiles [255]. We found that 57% (3,175) of accounts in the archive dataset and 24.4% (212) of accounts in the timeline dataset had more followers than friends. Fig. 4.17 shows the proportion of followers to friends for both datasets. Note that one potential source of bias in this type of ratio is accounts with very few friends and followers (e.g., an account with one friend and two followers). However, 56.8% (3,031) of accounts have more than 100 followers. Meanwhile, we could not find a single instance of a retweet bot following another bot.

4.7.6 Temporality

As we found while answering RQ2, the bots used for analysis in this chapter are not bots at every moment of their existence. Compromised accounts began as normal accounts, became bots, and then either became normal accounts again or became inactive. Fake accounts began as inactive accounts, became bots, and then resumed their inactivity. Most bot detection works do not consider this temporality and try to detect bot activity based on the entire timeline of an account's life. For example, we found that retweet bots which were previously classified as social bots by prior work [71], were in fact a mixture of human and inactive accounts. The accounts were compromised for only a short period, then they became inactive, possibly due to a soft ban by Twitter. It is not

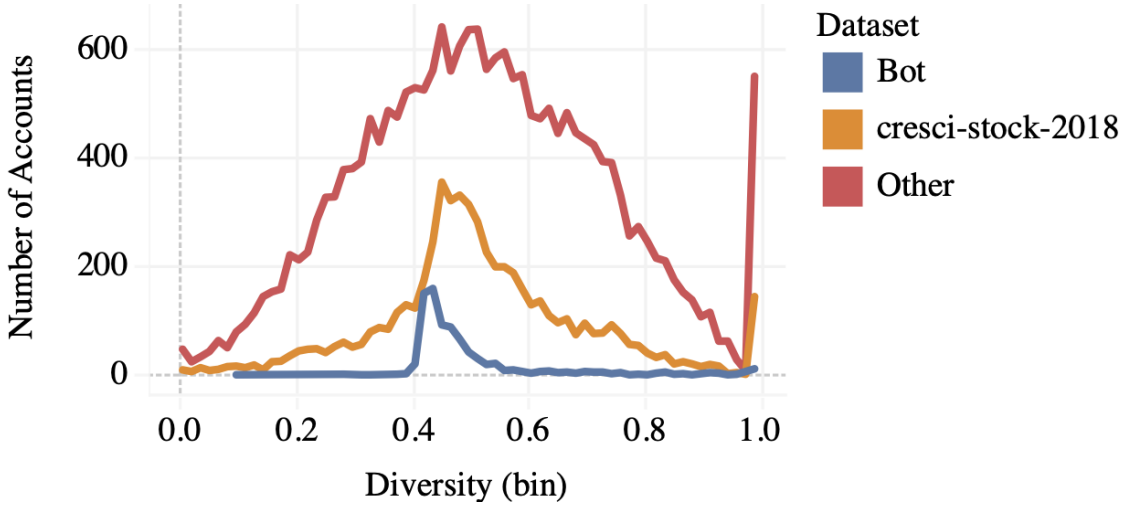

Figure 4.16: Number of accounts by the diversity of retweeted accounts; computed by dividing the number of unique users retweeted by the number of retweets. It follows a normal distribution for human accounts but is concentrated at 0.4-0.5 for accounts in the timeline and cresci-stock-2018 datasets.

surprising that these accounts were classified as bots, they have an unnatural spike in the number of retweets just after compromisation, nor is it surprising that a bot detection system trained on these accounts would classify inactive accounts as bots, which was the case with Botometer [103, 121]. The issue comes from the underlying assumption that a bot account, by its nature, is always automated, which neglects the fact that they can be compromised and act as a bot for only a particular period of time.

4.8 Implications of The Compromised Accounts on Bot Research

This section discusses the implications of our findings on bot research.

4.8.1 The Nature of Bots

Many studies focus on quantitative methods to detect or analyze the prevalence of bots, while few focus on explaining what the detected accounts truly are. No study investigates whether the detected bots actually emulate humans or how they do it. This has direct implications: the retweet bots in our study appeared to be *emulating* humans because they *were* humans. They had full and genuine profiles because they belonged to humans once. Bot detection studies could address this using qualitative methods to determine what the accounts were truly created for to explain their behavior. Learning their behavior through qualitative analysis may also advance bot detection systems.

4.8.2 Temporality

We saw when studying temporality that we must consider *when* an account is a bot and not only *that* it is. This goes before and beyond feature engineering and into the

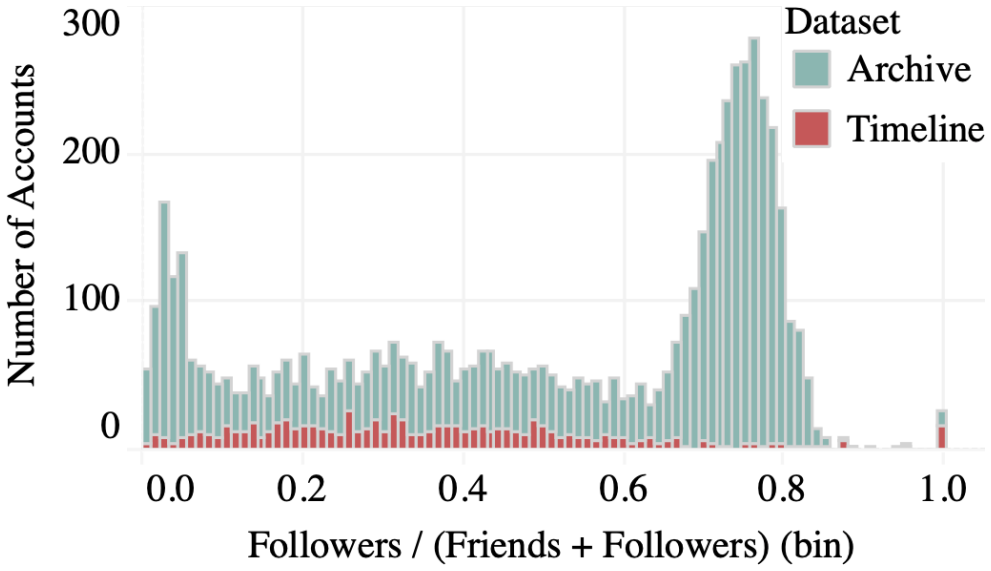

Figure 4.17: Follower ratios of accounts in the archive and timeline datasets. Surprisingly, accounts in the archive dataset had many more followers than friends.

actual problem setup and preprocessing of the training data. Studies should consider an account not as a static entity but as something that evolves and changes throughout its life. Studies proposing new bot datasets should specify what period of activity the annotators inspected during annotation. For the data used in this chapter, it is not advantageous to label retweeters in the timeline dataset as "bots" outside of March-August 2017, before they were compromised or after they left the botnet. Additionally, the compromised accounts used as astroturfing bots in the previous chapter should be classified as bots only when they are used as such. We already observed the implication of neglecting the temporality, as consistently classifying inactive accounts as bots is a reproducible error in Botometer [103].

4.8.3 Anomalous Behavior

Compromised retweet bots do not observe some of the anomalous behaviors as assumed by the previous studies. However, we found that they observe other kinds of anomalous behavior. For instance, even though they do not pass the 50 tweets per day threshold employed by a previous study, they retweet at an abnormal rate compared to their past activity, as we saw in Fig. 4.6. They do not consistently retweet the same people: on the contrary, they retweet a diverse set of users, which may be clients paying for their service. They do not immediately retweet the clients, but they retweet together with other retweet bots within a minute. The studies proposing methods to detect compromised accounts exploit such anomalous activity and the shift in the account behavior [93, 161, 162]. Meanwhile, to the best of our knowledge, bot detection studies do not consider such features while annotating the bots or detecting them. We advise bot researchers to consider such features in future work.

4.8.4 Monitoring Black-Market Activity Via Compromised Accounts

The adversaries used the compromised accounts in our dataset to create fake engagements for clients. Those clients may purchase engagements to manipulate social media, e.g., to do political astroturfing. To scale up the size of their campaigns, they may also be employing other bots, e.g., buying fake followers to inflate their follower counts. Thus, the bot research can identify clients and the other bots they employ by monitoring the black-market activity of the compromised accounts.

Compromised accounts may also assist in investigating criminal activity. This is because clients may purchase engagements for illegitimate activities, such as illicit advertisements that may not attract genuine engagements. Monitoring the accounts may reveal their malicious promotions. Although Twitter may be regulating cyber-criminal activity on its platform effectively, the adversaries may redirect victims to external websites through compromised accounts where they can promote their illegitimate business without scrutiny. Thus, the activity of compromised accounts may also be a reliable source of open-source intelligence on cyber-criminal activity. We leave the analysis of such illegitimate activity to future work.

4.9 Generalizability

While we only focus on Twitter, our results may give insights into platform manipulations on other platforms. In particular, we see that black markets sell fake engagements by compromising legitimate accounts. This can easily translate to fake Instagram likes and follows or Reddit upvotes. Indeed, people on the internet sometimes complain about the random accounts their account automatically follow [46] which may signify such illegitimate activity.

Retweets are visible on the timeline, so they are easily noticeable by the owner of the compromised accounts. Meanwhile, some engagements are less visible and harder to track by the account owner. For instance, Instagram users would not see a post they like on their news feed unless they follow the owner of the post. They must go through the "Your Activity" page and inspect their past likes. Thus, an adversary can stealthily use an Instagram account to send fake likes and avoid detection by the account owner. Platforms may have to detect such fake likes in users' behalves to protect their security.

4.10 Counter-Measures

Our findings have implications for bot research, particularly on the integrity of bot detection systems proposed by such research. This section presents recommendations that researchers can consider while implementing counter-measures to account for the implications we discussed.

4.10.1 Feature Engineering and Overfitting

Bot detection methods must consider the inherent assumptions made in feature engineering. The primary culprit here is that what is intuitive is not always true. For example, it may be intuitive that bot accounts follow more accounts than follow them back. However, we found no evidence of this in our analysis, so a classifier trained on data biased to this assumption may not identify the bots studied in this work.

Even in the instances when feature engineering is more quantitative than qualitative, the data itself may not be representative of bots more generally, so any conclusions drawn from it about which features are important for bots at large may be inaccurate. Of course, this is always a problem for machine learning classification: when a new class or variant of an existing class is not considered in training, the classifier cannot be expected to classify it correctly. Hence, when the results of a bot classifier are used to label a dataset for analysis, these results must be considered as reflecting *the types of bots that were used for training the classifier* and not bots more generally.

To help remedy this, studies could consider focusing on bots according to their functions instead of using broad umbrella concepts like "social bot" which may bias the labeling and classification. In this way, we can be more accurate and precise when talking about bots and their behaviors. This can also result in more reliable account labels, as discussing bots in terms of their functions can allow for the direct purchasing of bots' activity instead of relying on humans.

4.10.2 Beyond Classification

Our analysis shows that there is a sharp distinction between retweet bots and humans with respect to retweet percentages, so a simple rule-based classifier based on retweet percentage would already suffice to classify retweet bots *like the ones from this study*. However, if applied to live Twitter accounts, it would likely result in the misclassification of thousands of real users who use Twitter to amplify their favorite users, a problem already pointed out by Rauchfleisch and Kaiser [227]. This is because while we know that the retweeters analyzed in this study are bots based on the fact that their activity was directly purchased, we do not know which overactive retweeter accounts are solely controlled by humans. A next step to remedy this would be to find ways to collect a dataset that represents such human-controlled accounts.

The style of data collection used in this study is often difficult to obtain, requires institutional backing to purchase such accounts, and results in smaller datasets than classification. In cases where we use human-labeled data or classifier outputs, either by necessity or convenience, we must consider the assumptions that went into building the datasets and the classifier and both understand and acknowledge the implications of these assumptions on our results.

4.10.3 Data Considerations

Collecting data for bot studies in a way that is ethical, inexpensive, well-labeled, and large-scale is challenging. There is often a trade-off between the certainty of labels, ease of collection (and therefore scale of the dataset), and ethical data collection. The dataset used in this study sacrifices size and collection ease in favor of label certainty. In doing so, we found that there were some assumptions and findings from prior work that did not hold in our dataset. Future research should consider these trade-offs carefully in light of such findings.

4.11 Limitations

The dataset we propose in this chapter contains labels that are more reliable than seen in prior work. However, we still cannot overcome all of the limitations of data collected from the Internet. First, all data in this work came from particular black market websites. As such, we can only learn about this specific set of bots, possibly controlled by a few central points or controlled in the same manner. Other markets, or even other users on these particular markets, may have different strategies for their retweet bots which could yield different results. This is the first work to analyze retweet bots collected in this manner, so no comparisons to prior work can be made beyond those in the previous section to datasets collected using different methodologies. Still, these comparisons help us understand how different data collected in different situations can lead to different results, as we showed in answering RQ4.

Second, our dataset is still limited by the Twitter API and which data are available on the Internet Archive. These datasets do complement each other in that they contain different snapshots of Twitter, but they are still just small portions of the broader dataset of tweets. The Twitter API allows us to capture all of the *not removed* content of *non-suspended or private* accounts, leaving a sizable blind spot in terms of accounts that Twitter has suspended. This limits us in analyzing retweet bots that have purged their tweets. However, as Fig. 4.1 and Fig. 4.5 suggest, the majority of users have more than 1,000 retweets and were active when the retweet bots were the most active, indicating that this limitation does not harm our analysis. The API limitation of the statuses/user_timeline endpoints also prevented us from collecting a specific user's tweets beyond the last 3,200, which affected 263 users. This is a minor issue since we were still able to collect all tweets from the majority of users and a large amount (3,200) from the remaining 263 users. The Internet Archive dataset only contains 1% of tweets. This limits our analysis while answering RQ3, where 100% of tweets is necessary to learn the volume of daily activity of users (A), the percentage of retweets (B), the temporal analysis (C, D), and the diversity analysis (F).

4.12 Ethical Implications

Collecting high-quality data to study bots is challenging, but this work demonstrates the necessity of correctly labeling accounts as bots or humans. As such, we must be able to label some bots based on a third channel of information (e.g., purchasing). This introduces some ethical issues that do not arise when hand labeling or using classification, as in most cases, this involves participation in the black market.

To weigh the ethics of using data collected in a way that could be unethical, we refer to recent work by Ienca and Vayena [149] on responsibly using hacked data for research. While our work is not a straightforward application of their arguments, it is nonetheless a useful tool for understanding the ethics of using third-party datasets that may have been collected either unethically or in breach of the Terms of Service of the platform. This work argues that one must weigh the public value, optimization of resources, uniqueness, and cross-domain consistency of the leaked data (in our case, the archive dataset) against consent issues, possible secondary harms, breach of privacy, and lower quality data of an otherwise non-leaked dataset (a hand-labeled dataset collected using only the Twitter API). In this chapter, we opted not to actively participate in the black market but instead use data that others had collected from black market sources. In this way, we mitigate the harms of further participation in the market. This also mitigates another ethical dilemma: hand-labeling is by its nature error-prone and, thus, inevitably leads to human accounts being labeled as bots, as we found in answering *RQ4*. This can cause harm to these users who are treated as bot accounts. Finally, these data are also very unique, and without them, we are unable to claim a reliable ground truth.

In terms of data collection, throughout this work, we did not collect any data from Twitter except by use of the Twitter API, and with this data, we respect the rules and regulations written in the Twitter Terms of Service. We do utilize a dataset that was released by The Internet Archive, which may have been collected in breach of the Twitter terms of service because it contains deleted and suspended content. Referring again to the work of Ienca and Vayena, we weigh the utility and uniqueness of using deleted content collected by a third party, some of which is content that Twitter removed due to inauthentic bot activity and some of which was deleted by the bots to hide their activity [104], against the privacy of users who have deleted their content since the data was collected by The Internet Archive. Indeed, bot research has immense value for the public good as bots spread fake news and promote harmful narratives.

4.13 Summary

In this chapter, we presented a dataset of retweet bots directly purchased from vendors on the black market and analyzed them to learn where they originate, how long they are active, and how they differ from human accounts. This unique dataset gives us new insights into the world of bots. Particularly, we found that they were compromised, which gave them human traits. In studying the behavior and lifespan of retweet bots,

we also found several inconsistencies between our results and those obtained using bots studied in prior works. These results challenge some of the basic understandings about bot behavior and operation and highlight the need for studies that use reliable datasets and make decisions about bots based on them.

Chapter 5

Detecting Compromised Accounts in the Wild: Misleading Repurposing

Misleading Repurposing on Twitter

Under Review by ICWSM 2023

This chapter proposes a methodology to *detect* and extensively analyze a social media manipulation technique that is previously known but was not studied: *misleading repurposing*. On social media, an adversary can compromise and/or purchase a legitimate account and change its identity of it via, among other things, changes to the profile attributes to use the account for a new purpose while retaining its followers. Such repurposing of an account, which we name as *misleading repurposing*, is difficult to detect as it's not always possible to detect past profile attributes of the victim retrospectively. We propose a methodology to flag repurposed accounts that uses supervised learning on data mined from the Internet Archive's Twitter Stream Grab. We found over 100,000 accounts that may have been repurposed. We also characterize repurposed accounts and found that they are more likely to be repurposed after a period of inactivity and deleting old tweets. We also provide evidence that adversaries target accounts with high follower counts to repurpose and some make them have high follower counts by participating in follow-back schemes. Finally, we present a tool to root out accounts that became popular and repurposed later.

5.1 Introduction

"As Gregor Samsa woke one morning from uneasy dreams, he found himself transformed into some kind of monstrous vermin." - Franz Kafka [158]

Social media platforms allow users to change their profile information in order to keep up with real-world or online identity changes. For example, a user may change their real-world name and want their online identity to reflect that change, they may

want to make their profile more anonymous, or they may make a career change and want to change their description field to reflect it.

Not all attribute changes are genuine, however. Sporadically, journalists, bloggers, activists, and Twitter users report instances of accounts changing identities overnight to such an extreme degree that the former identity is lost completely, and the "new" account is used for a different purpose. For example, the account of an attractive woman with thousands of followers switching to an account promoting a political party, or the account of a user that reported to be based in the UK suddenly changing their name and location and taking on the identity of a patriotic American citizen. In instances such as these, accounts keep their followers, but transform all of their characteristics at once, including their name, screen name, description, location, website, and even the style or language of the tweets. We refer to this type of drastic shift in identity and/or characteristics as *repurposing*.

Such drastic changes, in which the entire identity of the account is changed suddenly, are usually the result of malicious activity. Consider an adversary who aims to execute some malicious activity that requires many followers, e.g., spam propagation, illicit advertisements, propaganda, political manipulation, etc. In this case, it's advantageous to use an account that has already accumulated followers and built trust with the public and the platform. Thus, the adversary may *compromise* or *purchase* a compromised account and repurpose it to use it for their malicious goals instead.

Compromised accounts are not the only type of trustworthy accounts that are later repurposed. Some accounts are only created to be repurposed later. That is, an adversary first creates a fake account not for its final purpose but for the sole purpose of gaining popularity and visibility via gaining followers. This is because a new fake account posting only spam or political content will get little attention from genuine users, so a more attractive and human-appearing account is first created to gain followers. Once this first goal is achieved, and the account has risen in popularity, the account owner changes the account to achieve the intended goal. Often the final goal of the malicious user is not to use the account but to *sell* the now popular account to another user who then changes it to fit their own purpose.

While repurposing can occur on any platform, exact policies for changing profile information vary, meaning that some platforms are more susceptible to account repurposing. Facebook limits accounts to be only personal profiles that correspond to a real person and disallows owning multiple accounts, so anonymity and name changes are uncommon and regulated. Users can create a *page* to post anonymously to some extent, but pages' features are limited, e.g., they cannot befriend or interact with personal profiles. On the other hand, Twitter allows for personal profiles, anonymous accounts using pseudonyms, and hobby accounts (e.g., parody, commentary, and fan accounts). These different types of accounts are functionally the same; unlike Facebook, there is no distinction between personal profiles and pages, so a personal profile can easily be transformed into a hobby account. While Twitter prohibits any account transfers or

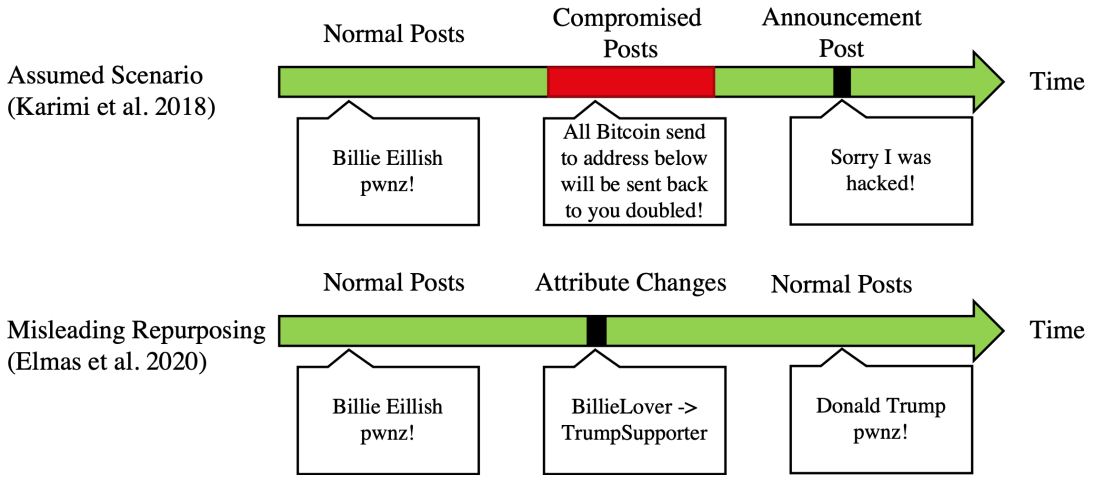

Figure 5.1: The scenario is assumed by previous compromised account detection methods (above) and the scenario proposed by our work (below). The first scenario assumes the account will observe and retain anomalous post when it was compromised and will self-state that it was compromised. In our scenario, the compromised account does not show any anomalous behavior or does not self-state that it was compromised.

sales, no rule explicitly regulates the repurposing of one's own account. Facebook, however, prohibits name changes that are "misleading" or "substantially change the Page's subject matter" [221]. It further notifies the followers of pages and groups when a name change occurs, unlike Twitter.

Compromised and/or sold accounts are challenging to detect in the wild. This is because the malicious activity that gives away the illegitimate nature of these accounts is ephemeral: they change their names and descriptions overnight. Previous detection methods in compromised account detection [93, 161, 162] may not be effective in finding repurposed accounts due to such ephemerality. They also assume that the compromised accounts observe anomalous behavior and share spam, which may not hold in the case of misleading repurposing. Fig. 5.1 shows the two scenarios to highlight their differences. By studying misleading repurposing, we also propose a detection method to root out such compromised accounts that may go unnoticed by the previous methods.

The contributions of this chapter are as follows:

1. We introduce the concept of misleading repurposing and suggest a definition. (§5.3)
2. We present the first large-scale study of misleading repurposing using a massive retrospective dataset. (§5.4)
3. We establish a hand-labeled ground-truth dataset of repurposed accounts using datasets published by Twitter. (§5.5)
4. We provide an analysis of repurposed accounts and find that they were more likely to build towards and/or have higher follower counts. We also found that some accounts were repurposed after staying dormant for a while and deleted their old tweets. (§5.6)

5. We propose a classifier to flag repurposed accounts in the wild, which may be compromised and/or sold. (§5.7)

6. We propose a tool to study and visualize repurposed accounts in the wild. (§5.11)

The structure of this chapter is as follows: We first create a concrete definition of misleading repurposing. Then we propose a framework to find repurposed accounts. We build a dataset of repurposed accounts. We characterize the repurposed accounts. We present our classifier to detect repurposed accounts in the wild. We lastly present a tool to study and visualize repurposed accounts in the wild.

5.2 Related Work

5.2.1 Twitter Attribute Changes

Previous work analyzed how users change their profile attributes, primarily to uncover under which circumstances these changes are made [205, 245, 301]. Jain et al. [152] found that users change their attributes to maintain multiple accounts, change user identifiability, and for username squatting. Regarding screen names in particular, Mariconti et al. [189] found that adversaries hijack the screen names of popular users who recently changed their screen name in order to gain visibility, often with malicious intent. Zannettou et al. [322] found that 9% of 2,700 accounts operated by Russian trolls changed their screen name. None of these works reported that accounts change of screen name as a signal to repurposing.

5.2.2 Accounts Changing Ownership

Accounts can change ownership either from being compromised or on mutual agreement between the previous and current account owners, often as a result of commerce. Compromised accounts are well studied in the literature, including their detection [93, 161, 162], what the compromised accounts are used for [286] and user reactions to their accounts' compromisation [244, 318]. Thomas et al. [265] studied how Twitter accounts were bought and sold on illicit forums. Our work builds onto these works as studying repurposings roots out compromised and/or sold accounts, which may go unnoticed by the previous detection methods.

5.2.3 Platform Manipulation

Past research primarily focused on accounts with automated behaviors, e.g., spammers [34, 143], fake followers [68], impersonating bots [128], dormant bots [259], retweet bots [100], and astroturfing bots [105]. Most of the research on non-automated sockpuppets relies on datasets published by Twitter. Our work tries to break out of this pattern. Timely detection of fake accounts through classification of repurposing behavior could lead to early detection of accounts that participate in social media manipulation.

5.2.3.1 Style Change Detection

We employ style change detection methods in this thesis after a review of the current methods. Traditional methods employ handcrafted features. These include frequency usage of special characters, numbers, uppercase letters, functional words, POS tags, long words, hashtags, mentions, URLs etc [3, 36, 333] and readability features such as Flesch reading ease [331]. More advanced methods use state-of-the-art sentence representation models such as BERT [83]. In fact, the best performing models of the PAN 2020 [317] and PAN 2021 style change detection tasks [316] all employ Google's BERT language model to creature features out of the representation of the texts [151, 257, 328]. Style change detection can be used for broader problems such as authorship attribution [9, 174, 208] and authorship verification [44]. Our problem differs from these problems as we do not need to attribute the account to a particular author, but only need to verify that the account is now being used for a new purpose.

5.2.4 User Analysis Tools

In this work, we also propose a tool that allows retrospective analysis of users. Several web applications provide a summary (e.g. number of tweets, followers, the hashtags and the topics the user uses and their interactions) using the recent tweets of a given user such as foller.me, accountanalysis.app, twitonomy.com, followerwonk, or of the authenticating user such as Twitter's own analytics tool[1]. However, they use up-to-date statistics which is extracted using only the most recent tweets due to API limitations, contrary to our approach which uses a retrospective dataset. Additionally, our tool is the first to provide follower growth, change of attributes, and deletions to the best of our knowledge.

5.3 Defining Misleading Repurposing

5.3.1 Definition

To study *misleading repurposing* extensively, we first must come up with a definition of this behavior. Repurposing is "to adapt for use in a different purpose". This definition covers the anomalous cases in which we are interested such as an account used as a personal account is adapted to use as a personal account of another person. However, it also introduces many false positives. For instance, Joe Biden becomes the president of the United States and he adapts his personal Twitter account to use for a different purpose: to share announcements about the current president of the United States. Such behavior does not intend to cause harm and is not the focus of this study; the same person and/or account may have multiple purposes and/or (slightly) change purposes over time.

We are interested in repurposing behavior that *misleads external observers about the past of the account*. Therefore we define misleading repurposing as a substantial change to the account attributes so that the initial identity (i.e. the entity it represents) or purpose (i.e. what it is created for) of the account cannot be inferred from the new state

[1] https://analytics.twitter.com

of the account. Thus, the account *misleads others about who they were and what they were doing*. This definition is similar to Facebook's definition of inauthentic behavior which is stated as "misleading people or Facebook about the identity, purpose or origin of the entity that they represent." [111]. It also covers compromised accounts that are sold and changed overnight for a new purpose.

Meanwhile, legitimate changes introduced to the accounts (e.g., marriage, anonymization, or becoming president) might make it difficult to infer the initial identity or purpose of the account as well. To exclude such cases, we also require that the new state of the account should have a recognizable identity or a purpose that is irrelevant to the initial identity or purpose. For personal accounts, these would mean the account has gone through an identity change and now represent a new person. For non-personal accounts such as organizational accounts or hobby accounts centered around a subject, we adopt Facebook's definition and say "substantial change to the account's subject". Thus, we define misleading repurposing as a "change of the account's identity and/or a substantial change of the subject matter with the intention to mislead the public about the past of the account." This also covers personal accounts which are repurposed to be non-personal accounts and vice-versa.

Finally, misleading repurposing is not necessarily malicious, e.g., legitimate hobby page can be repurposed to be used to promote another hobby page that is also legitimate. Our goal in this chapter is to uncover misleading repurposing, independent of whether the intent or the account was malicious. Uncovering misleading repurposing is a starting point in revealing malicious accounts that disrupt the public dialogue through social media manipulation.

5.3.2 Cases

Misleading repurposing can be employed for any reason, however, the most interesting scenarios for studying account misuse are those in which the objective is public manipulation. We identified two such scenarios: 1) coordinated manipulation and 2) fake influentials. We detail media reports and white papers reporting the real-life instances of the use cases we describe.

5.3.2.1 Coordinated Manipulation:

Adversaries use multiple accounts in a coordinated way to influence the public. The accounts may have little impact on their own but can be effective when deployed to sway a specific discussion.

Adversaries do not need to create accounts anew to achieve a new goal; they can repurpose the same accounts and adapt them to the new goal. For example, an instance of a number of fake accounts attacking far-right French politicians in a coordinated manner. All of the accounts used the same email address and stolen photos. Later, the accounts were repurposed: they deleted all of their tweets and claimed to be "some sort of artificial neural network company or laboratory filled with fake content" [305].

Similarly, 12 "troll" accounts that were suspended after working to influence the Brexit debate initially reported their locations to be in Germany and had bios in German [183].

Furthermore, adversaries do not need to create accounts themselves; they can buy accounts that were mass created by a third party or with stolen credentials for as cheap as five cents per account [265] and then repurpose them. For instance, Uren et al. [282] uncovered a set of low-impact, coordinating accounts that were repurposed after being either bought or compromised, likely by actors linked to the Chinese government. These accounts were repurposed to promote government propaganda in Hong Kong. The new account owners made only a minimal effort to hide the fact that the accounts were repurposed so the authors were able to trace the previous owner of the account through unchanged profile attributes, past tweets which were left undeleted, and a bot run by the previous owner which was still reposting tweets after repurposing. Some were even tweeting in a different language before the repurposing.

5.3.2.2 Fake Influentials:

Adversaries sometimes employ accounts with a high number of followers in order to influence many people at once or to be seen as influential. For instance, a Russian troll account named "TEN_GOP" presenting itself as the "Unofficial Twitter of Tennessee Republicans" reached over 100,000 followers and was retweeted by many celebrities and politicians [297]. Misleading repurposing makes it easy to create such fake influentials.

For instance, Sözeri [252] reported an account by the name of "Oy ve Hilesi" (English: *Vote and Fraud*) which was seemingly created only to attack a pro-opposition NGO called "Oy ve Ötesi" (English: *Vote and Beyond*) during 2015 Turkish elections. The account was created in 2014. He found that just prior to being called "Oy ve Hilesi," the account had the identity of "a sexy girl" and was tweeting romantic quotes as part of a scheme to artificially gain followers. Once it reached 40,000 followers, the account was sold on a webmaster forum for 200 Turkish Lira (∼$70). The new owner deleted the old and irrelevant tweets, changed the name, the description and the profile picture and, thus, shifted from a fake personal profile of a woman to an anonymous account used to attack the opposition. The account did not immediately unfollow the 40,000 accounts that it was following at the time of the repurposing but waited until much later to do so. This may be because the account's followers were following it only for a follow in return and would unfollow when being unfollowed. As of May 2022, the account retains 34,000 of its followers.

Similarly, Grossman et al. [133] reported sockpuppets that were posing as Qataris living in Saudi Arabia. The authors emphasized the high number of followers the accounts have and claimed that these accounts "increase their audiences with follow-back spam behavior," and then are repurposed to mimic public figures after changing their screen name and deleting all the tweets associated with the previous identity of the accounts.

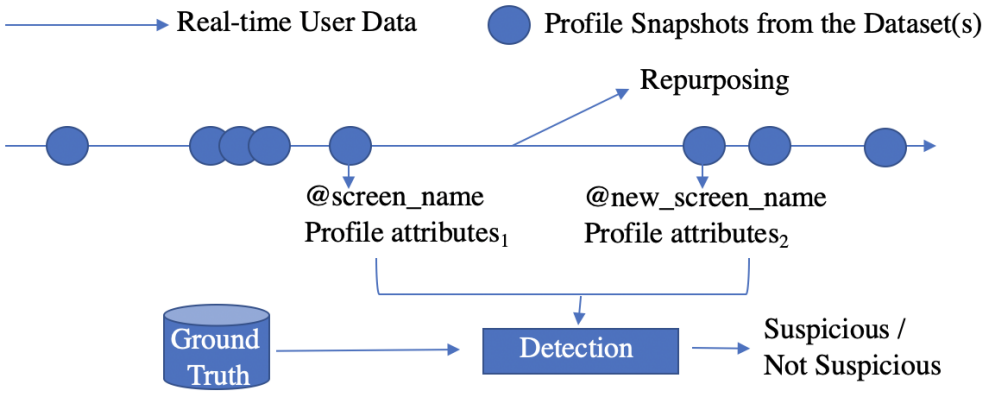

Figure 5.2: Summary of our methodology. We use snapshots of an account and detect if a repurposing might have taken place based on the ground truth we built.

5.4 Building a Dataset of Repurposed Accounts

In order to study repurposing more broadly, we must first find more instances of repurposed accounts. As this is a rare event, we take a machine learning approach to simulate the function and mass-label accounts as repurposed. We first hand-label a set of accounts as repurposed or not, then train a classifier to find more instances on the wild.

Our methodology consists of collecting a historical dataset that contains past profile snapshots to reconstruct the users' Twitter history, hand-labeling this data to establish ground truth, and building a classifier to detect suspect repurposed accounts in the wild. The process is summarized in Figure 5.2.

5.4.1 Base Dataset (Archive)

In order to detect if an account has been repurposed, we must, at the very least, have a snapshot of the account before the repurposing and a snapshot after to witness the change. To this end, we use a dataset of public Twitter data that is archived by the Internet Archive's Twitter Stream [1]. This dataset contains a 1% sample of all tweets, including retweets and quotes, and includes a profile snapshot of the user who posted the tweet and, if applicable, the retweeted/quoted user. As the sample includes retweets, popular users with many retweets will appear more often than accounts with fewer of these interactions. Thus, the dataset is biased towards active and popular accounts, which is an advantage in this study, as such accounts have the greatest potential to have an impact. We name this dataset *archive*.

Even though it contains only 1% of tweets, this dataset is massive. At the time of analysis (October 2020), the dataset dated from September 2011 to June 2020 and contained 446 million user ids. We create an abbreviated version of this dataset by only considering accounts that changed their screen name since we found this signal in every example of repurposing we studied. In general, this is a rare action on Twitter since a screen name is a unique identifier and the way that profiles are searched and shared (i.e.

twitter.com/justinbieber). Additionally, Wesslen et al. [301] found that this attribute is stable for most users. We found that only 13.3% (59 million) of users in the archive dataset changed their screen name, confirming this finding.

5.4.2 Ground Truth Datasets

To establish ground truth, we hand label a set of accounts. Our preliminary analysis shows that randomly sampling Twitter users who changed their screen names to find positives is not an efficient strategy as repurposing is extremely rare. Additionally, negative cases that are randomly sampled are very trivial to classify: they are often slight changes to screen names with little or no changes to names and description fields. Therefore, we follow a multi-step approach: we first hand-labeled a collection of accounts from a set that we know contains repurposed accounts and build a simple classifier on this data with sufficient precision and recall. We then deploy this classifier *in the wild* (i.e., on normal Twitter accounts) and detect positives. We then manually annotate these positives and create a second ground-truth dataset.

5.4.2.1 Civic-Integrity Ground Truth Set (Integrity)

We use the datasets published by Twitter that involve state-sponsored accounts that undermine elections integrity for the first step. Some were already reported to be repurposed by previous work. By October 2020, there were 35 datasets focused on 16 countries [276]. The datasets do not include past profile attributes. Thus, for each user id in these datasets, we extracted the historical data from the archive dataset. Of the 83,481 unique user ids, 38,426 were found in the archive. We found 17,220 screen name changes involving 8,370 accounts. We name this dataset *integrity*.

5.4.2.2 In The Wild Popular Users Ground Truth Set (Popular)

We have a ground truth set of accounts with many positive cases. However, this set is biased. We observed that malicious accounts often change their name and description field drastically for the purpose of misleading repurposing. This makes the probability of misleading repurposing given the drastic change in profile attributes close to one, making two events seem the same. Additionally, the base rate of the positive cases in the integrity dataset is very high compared to a random sample. As a result, in our preliminary experiments, we observe that our classifier reporting good scores on the integrity dataset performed poorly in the wild and yield many false positives. Thus, we used the accounts detected by our initial classifier as a ground truth set for the second step. We then took an active learning approach and deployed new and more complex classifiers to improve our initial classifier. We call this new ground truth set the *popular* dataset.

We describe our annotation process and the annotation scheme we suggest in detail in the next section.

5.5 Annotation

We use human annotation to build a ground truth for repurposed accounts. We refrain from crowdsourcing this task because 1) the integrity dataset is only fully available to researchers given access by Twitter; 2) the archive dataset, although public, contains sensitive information, e.g., former names of real users; and 3) expert annotation is more reliable than crowdsourcing for complex tasks [235, 250].

5.5.1 Procedure

We treat each instance of a screen name change as a separate data point. For each change from screen name s_i to s_j, we select the last available snapshot of the profile with s_i and the first snapshot with s_j. The same user may have multiple screen name changes and, thus, may be represented multiple times. We presented the annotators with the following semantically interpretable attributes: name, screen name, description, location, home page url, profile settings language, most common tweet source and tweet language. We asked the annotators the following question and allowed for the responses: *Yes(+)*, *No(-)*, and *unsure*.

> *Did the account change in a way that makes it seem that the account is now owned by a different person/organization, or has the account rebranded itself substantially?*

Two authors/domain experts, A_1 and A_2, independently coded all cases. We report the annotator agreement using Cohen's kappa. Due to the subjective nature of the problem, we observed many cases in which we needed to code and determine a common answer. The relatively low agreement of a non-expert further emphasized this need. Since this problem is unexplored, there is no coding scheme available. Thus, A_1 and A_2 developed a coding scheme and made decisions for differently coded cases when needed. First, each annotator independently annotated the data using only the initial annotation question. Then they compared the annotations and computed the annotator agreement. They then discussed and coded the cases in which they disagreed or were both *unsure*. Finally, a decision was made for each case. Below, we present the cases and the decisions.

5.5.2 Annotated Data

Integrity Dataset We first selected English and French profiles for validation by multiple annotators. A_1 and A_2 independently annotated 200 cases. A_1 labeled accounts in this sample and passed 100 *positive* and 100 *negative* cases to A_2 for annotation. The inter-annotator agreement between the authors was $\kappa = 0.8$ (substantial agreement). The agreement on negatives was higher (93%) than positives (87%), i.e., it is easy to discard *negative* cases which are likely more prevalent "in the wild". The annotators discussed 20 cases in which they did not agree and came up with a verdict for each.

To expand this labeled dataset, A_1 labeled an additional 1,476 profiles in English, French, and Turkish (Turkish accounts were made available after the initial annotation was complete). This resulted in 512 *positive*, 910 *negative*, and 254 *unsure*.

Popular Dataset For the popular dataset, we used a stratified sampling approach and sampled 400 accounts from the list of users with the most followers before the repurposing and 600 accounts from a random sample of users who had more than 5,000 followers. Half of those accounts tweeted in English while the other half tweeted in Turkish.

For the popular dataset, the annotation was done simultaneously: A_1 and A_2 independently annotated the samples. This resulted in $\kappa = 0.66$ including *unsure* cases (i.e. one decided *negative* while the other decided *unsure* was considered disagreement) and $\kappa = 0.81$ when cases decided as *unsure* cases were discarded from the data. We observe that the disagreements were mostly due to overestimating the prevalence of repurposings as all accounts in this dataset substantially change their name and descriptions.

A_1 and A_2 then discussed the cases in which they did not agree and either came to a consensus or assigned a label of "Disagree" in the case of disagreement or *unsure* if both annotators were *unsure* of the case. This annotation resulted in 562 *positive* cases, 127 *negative* cases, 278 *unsure* cases, and 33 disagreed cases.

To expand this labeled dataset, A_1 additionally annotated 1,500 accounts with the same sampling strategy. This annotation was done more conservatively and the goal was to increase *negative* examples, since in-the-wild *negative* cases with dramatic name changes are rare. Only the cases where the annotator was highly confident were annotated as *positive*; *negative* and *unsure* cases were not checked for a second time. This yield an additional 421 *positive*, 248 *negative*, and 831 *unsure*.

5.5.3 Annotated Cases

Our coded cases and decisions for each are presented in Table 5.1. We explain each in detail.

5.5.4 Positive Cases

Different Identity: Misleading repurposing is evident when an account purports a completely new person or hobby page or an organization when compared to its old version. The account has a new name, has a new website, and moved to a new location. It is easy to infer the purpose of the old snapshot and the new snapshot from the description field and they are dramatically different. Accounts representing personal profiles of real people and official profiles of organizations generally observe these phenomena, which make it easier to annotate as such accounts are used for the social media presence of their owners. Hobby pages also comply with this criteria as they make their purpose clear.

Misleading "Purposing": Misleading *re*purposing is not evident because the purpose of the previous snapshot is not clearly established (e.g., it is blank or basic), but the

Table 5.1: Summary of Cases in Our Annotation Framework

Case	Example	Verdict
Different Identity	Lawyer John Doe becomes Dr. Mohammad Lee	Positive
Misleading "Purposing"	Empty profile of "sdfdsfsd" becomes politics enthusiast John Doe from USA	Positive
Commercial Activity	Account named "FOR SALE" becomes John Doe	Positive
Same Person	Jane Doe marries and becomes Jane Brown	Negative
Slight Change In Subject	Philosophy Quotes becomes Inspiring Quotes	Negative
Purpose Overloading	Jenna The Traveler becomes politics enthusiast Jenna Abrams	Negative
No Purpose / Unclear	White Horse becomes Black Rose but still shares quotes	Unsure
Organization Rebranding	Windows Phone becomes Lumia	Unsure
Lazy Compromisation	An American starts to tweet against HK Protests in Mandarin	Unsure
Lazy Repurposing	Patriotic Somalian changes name, but keeps description	Unsure
Person↔Org. Unclear	John Doe becomes "Lonely Boy's Pen"	Unsure
Change pseudoynms	Excalibur17 becomes Rebellion47 but still plays DOTA	Unsure

new snapshot has a purpose. Because the previous snapshot appears to be created and later given a purpose, the purpose of the new snapshot is still inconsistent with the old snapshot, so we regard this case as *positive*. We observe this case only among 30 accounts from the Russia 08/2018 dataset (named "ira" by Twitter) which used a (likely malicious) source to post tweets called "masss post$x$" ($x \in \{2, 3, 4, 5\}$). The account names and screen names were random strings. They originally reported their locations as cities in England but after repurposing, all changed their locations to "USA" or a specific US state. Similarly, all changed their names to names that are more prevalent in the US, (exceptionally, one adopted the name of a local news outlet). The accounts were created on various dates in 2014. We annotated those accounts as *positive* because the location fields were inconsistent and there is strong evidence that these accounts were bulk created and/or purchased to be repurposed later. This case appeared to be rare among the popular accounts.

Commercial Activity: An account that is sold and later repurposed to clearly purport a new identity is the ideal *positive* case. However, some accounts do not clearly establish their purpose in their description field but explicitly state that they are for sale instead. We consider these cases as *positive* if the new snapshot appears to purport a new identity, which likely signifies that the account was sold and now being used for a new purpose.

5.5.5 Negative Cases

Same Person: A user changed their profile attributes but it is evident that they are the same person. This is often apparent because keywords are shared between snapshots or the writing style of the account does not change. We observe that many teenage

pop artist fans change their attributes frequently to express their admiration in different ways. We annotate these cases as *negative* if we can confidently infer that the account purports the same person and purpose. Otherwise, we annotated them as *unsure*.

Slight Change in Subject Matter: Although the profile attributes have changed, the subject matter did not change substantially and it appears to be the same or almost the same. An exception is the organizations that rebrand themselves, which is explained in the next section in detail.

Purpose Overloading: The account appears to change its purpose but it still purports the same person or page. We observe this case among accounts suspended in October 2018 originating from Russia. We identified 43 accounts that appear to be personal profiles which had initially blank or politically neutral description fields but then later adopted description fields that explicate their political stance (e.g., pro-Israeli, patriot, conservative, #Blacklivesmatter) alongside a corresponding change in demographic attributes (e.g., Christian, Black). All 43 accounts changed their screen names, but they appear to keep the same identity (e.g., Jenna the Traveler became Jenna Abrams) and so we did not consider them to have been repurposed per our definition. Instead of repurposing, the accounts appear to add a purpose. Their current purpose of showing off their personality does not change and coexists with the new purpose. These accounts might have accumulated followings from politically neutral personal accounts, then repurposed them by overloading them with a political agenda without deleting their old tweets and/or repurposing. Such accounts can be repurposed over and over for any topic within the same context, i.e., a regular American citizen account can be repurposed by overloading it with pro-republican content while still tweeting regular non-political tweets. Then the pro-republican tweets can be deleted and the account can be partially repurposed. This is a stealthy and potentially malicious strategy, but is not a case of misleading repurposing.

5.5.6 Unsure Cases

No Purpose or Purpose Unclear: It is difficult to understand the purpose of the account. For the English-tweeting accounts, these were primarily accounts in which cultural context that the annotator did not have was needed to make a decision (e.g., accounts from Nigeria, a country that the annotators were not familiar with). For the Turkish-tweeting accounts, this was often due to a lack of profile attributes that state the purpose of the accounts, e.g., we observe many accounts use pseudonyms instead of personal profiles.

Organization Rebranding: Some organizations were sold, rebranded, and/or changed name. Examples include musical.ly which became TikTok, Windows Phone which became Lumia, and Facebook which became Meta[2]. We consider those repurposings as *unsure* since it is not clear if they are repurposed nor if it is misleading. Other than those obvious examples, it is difficult to distinguish a rebranding and a new company

[2]Facebook renamed its old account to Meta but also kept a separate private account @Facebook to reserve its name.

without thorough research. Thus, if the purpose of the previous and the next snapshot of the organization is the same or similar even though they appear to be a different organizations, we annotate such organizations as *unsure*.

Lazy Compromisation: An account that is compromised and later repurposed to clearly purport a new identity is the same case as the ideal *positive* case. However, some put minimal effort to hide the compromisation: they slightly change the accounts' profiles or do not change them at all. Meanwhile, they tweet for malicious purposes so they change the accounts' purpose without changing the attributes. We observe this case among the accounts hijacked by Chinese users as reported by Uren at al. ([282]). If it is evident that the accounts were compromised due to a lack of changes in profile attributes, we consider this case as *unsure* as it is not misleading although the accounts are repurposed.

Lazy Repurposing Adversaries change the names of the personal accounts but not their descriptions, making it difficult to judge if those accounts purport the same people. We observed this behavior among some accounts originating from the U.A.E. which claim that they are patriotic citizens from Somalia.

Person ↔ Organization Unclear: An account that has the same purpose, but it is repurposed to be a page or an organization when it was a personal profile. It is not clear if this should be considered misleading repurposing because it could be the same person being professional or adopting a pseudonym for their hobby. We annotate such cases as *unsure*. Exceptionally, if a profile appears to be a user with a hobby turning their personal page into a hobby page, we consider this case *negative*. E.g., John Doe stating they he shares photography adopts the name DoePhotography.

Change of Pseudonyms: It is not clear if a person/page is repurposed to be a new person/page even though the name and description changed dramatically. These people/pages did not change their domain. We observe the former among esports gamers as they sometimes switch pseudonyms and teams, but they play and stream the same game. We observe the latter among meme pages as their names and description fields are also memes but there is no other indication of the specific purpose of the account. We annotate those cases as *unsure*.

5.6 Characterization

We next describe some of the characteristics of the accounts that have undergone misleading repurposing. Specifically, we discover that misleading repurposed accounts often 1) have more followers than other accounts that change their screen names, 2) utilize follow-back schemes to grow their follower counts, 3) delete tweets related to their former purpose, and 4) have a period of dormancy before the repurposing.

5.6.1 Follower Count

Accounts with a high number of followers are more likely to be misleading repurposed when they change their screen name than other accounts that undergo a screen name

change. This is likely because screen name changes are an anomaly for influential accounts since they lose any incoming links (i.e. twitter.com/jack does not redirect to Jack's new handle if he changed his screen name). Additionally, they may be more likely to be a target of account swap due to compromisation or commercial activity. This may also be an artifact of the data collection: users with a high number of followers are more active so it's more likely that we capture their screen name changes and, thus, their repurposing. Similarly, this does not entail that misleading repurposing is more prevalent among accounts with high follower counts. Figure 5.3 illustrates this difference.

In the integrity dataset, the mean followers count before the screen name change was 14,007 for repurposed accounts and 6,579 for non-repurposed accounts. The difference is 7,427 and statistically significant according to Welch's t-test. In the popular dataset, the mean followers count before the screen name change is 327,431 for repurposed accounts and 139,237 for non-repurposed accounts. The difference is 188,194 and statistically significant according to Welch's t-test ($p < 0.0001$ in both cases).

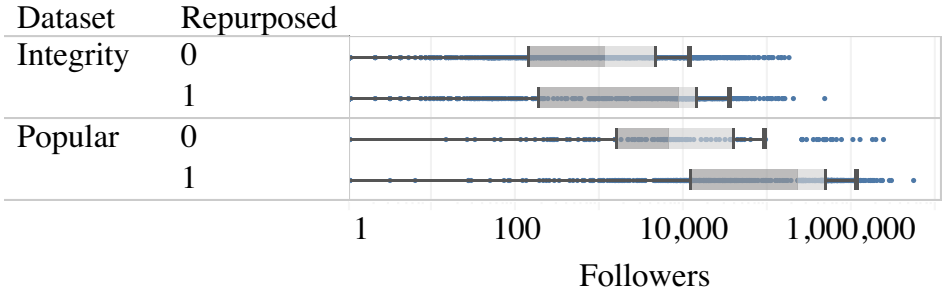

Figure 5.3: A box plot showing the number of followers for repurposed vs not repurposed accounts. High follower counts are more likely to indicate repurposing.

5.6.2 Follow-Back

Some, if not all, repurposed accounts appear to actively grow their accounts by joining follow-back schemes. They indicate that they follow back once another account follows them by using dedicated hashtags. Out of 1,595 repurposed accounts, 81 accounts used #FF ("Follow Friday", which is the most used hashtag in the dataset), 50 accounts used #Follow, 44 accounts used #IFollowBack, and 36 accounts used #TeamFollowBack. Meanwhile, out of 1,385 non-repurposed accounts, only 9 accounts used #FF, 7 used #TeamFollowBack, and 5 used #Follow. Our caveat is that these numbers are based on 1% of the tweets and there may be many more users using those hashtags.

5.6.3 Deletions

Repurposed accounts often delete the tweets which are irrelevant to the new purpose of the account. We observe this behavior by comparing the number of tweets before and after the account changed its screen name. Fig. 5.4 shows that the repurposed accounts are more likely to lose up to 96% of their tweets. Precisely, 519 of the 1,595 repurposings

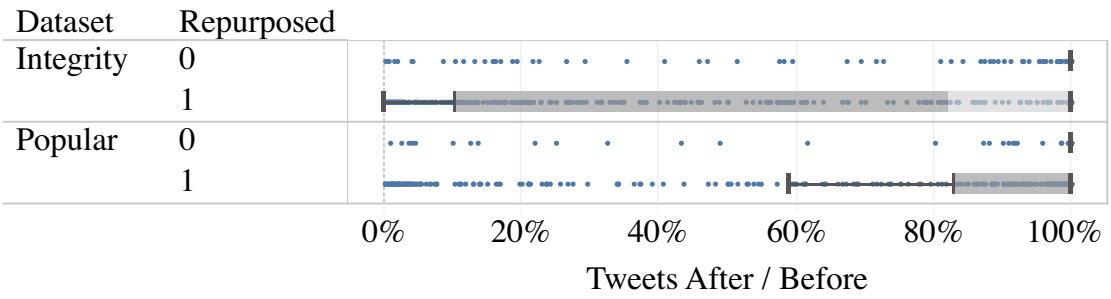

Figure 5.4: Box plot of the ratios of the number of tweets before and after an account changed its screen name for repurposed vs. not repurposed accounts. Accounts that gained tweets (i.e., created more tweets than deleted) are not included because they are not relevant here and have a ratio > 1. Repurposed accounts are more likely to delete their tweets fully or partially when they change screen names.

(32%) resulted in removing at least one tweet versus 75 of the 1,385 non-repurposings (5%). The difference is statistically significant according to the chi-squared test with $p < 0.0001$

5.6.4 Dormancy

We observe that repurposed accounts in the integrity datasets are more likely to be dormant for a long period prior to repurposing. This may be because the owners of the accounts no longer use them and eventually sell them. Alternatively, the accounts get compromised but since the original owner does not use them, they do not claim them and let them be repurposed by another malicious user. We did not observe this behavior among popular accounts.

5.7 Detection

Table 5.2: Results on the integrity dataset (-I) and the popular dataset (-P). Best performances in bold. We use F1 as the primary evaluation metric for the integrity dataset and AUC as the primary evaluation metric for the popular dataset due to distinct base rates. We report the other scores for completeness.

Model	F1-I	AUC-P	Prec-I	Rec-I	AUC-I	Prec-P	Rec-P	F1-P	TPR-P	FPR-P
EDT (Baseline)	92.3%	-	94.4%	90.3%	92.6%	81.5%	-	-	-	-
EDT-DSIM	92.8%	**88.4%**	95.5%	90.3%	97.3%	92.1%	92.7%	92.4%	92.7%	37.4%
EDT-STY	93.1%	73.2%	91.7%	94.6%	97.5%	88.9%	87.3%	88.1%	87.3%	51.4%
EDT (Retrained)	94.0%	78.5%	95.6%	92.5%	97.8%	88.7%	92.3%	90.5%	92.3%	55.1%
EDT-MD	94.0%	79.4%	95.6%	92.5%	98.4%	90.9%	89.3%	90.1%	89.3%	42.1%
EDT-DSIM-MD	94.0%	**87.6%**	95.6%	92.5%	98.6%	91.5%	92.1%	91.8%	92.1%	40.2%
EDT-DSIM-MD-STY	94.5%	84.8%	96.6%	92.5%	98.1%	92.3%	90.5%	91.4%	90.5%	35.5%
EDT-MD-STY	94.6%	76.4%	95.6%	93.5%	98.2%	91.0%	82.1%	86.3%	82.1%	38.3%
EDT-DSIM-STY	**95.1%**	83.1%	96.7%	93.5%	97.5%	91.2%	92.3%	91.7%	92.3%	42.1%

We next provide a classifier to detect misleading repurposing in the wild. The goal

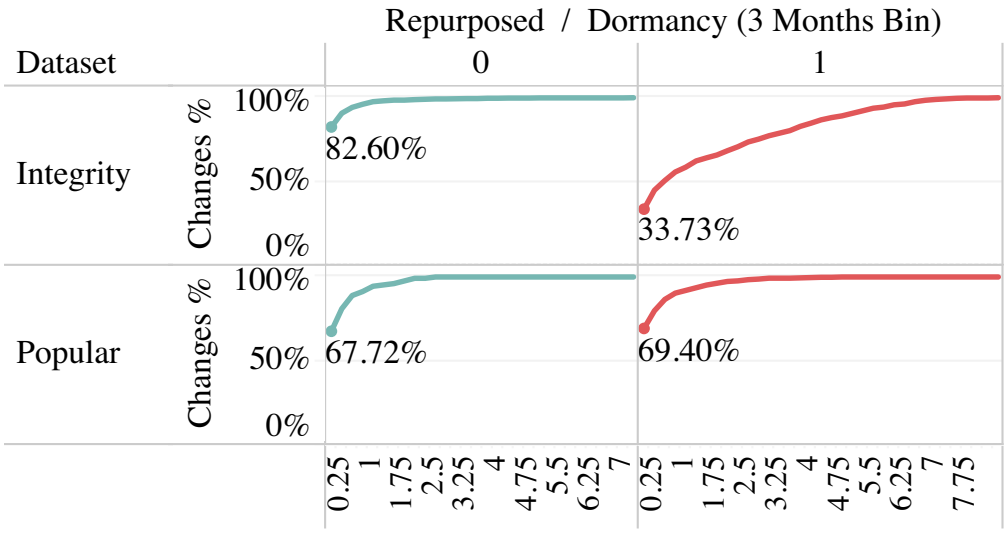

Figure 5.5: Cumulative distribution (CDF) of the percentage of accounts staying dormant. Bins are 3 months/quarter years. Misleading repurposed accounts in the integrity dataset are more likely to repurpose after staying dormant for a while than other accounts with screen name changes. We did not observe this behavior among popular accounts.

of this detection method is not to develop the framework that should be used by Twitter or other social media companies to detect repurposed accounts, as they have access to a richer set of signals and data. Instead, we provide a framework for researchers who do not have such privileged access to flag accounts that are potentially repurposed by only using publicly available data. Due to the subjective nature of the problem, we advise that the detection should always be accompanied by expert verification.

We tackle the following classification problem:

Problem Statement Let A be the account with the screen name scn_{t_i} at time t_i. Let PA_{t_i} be the profile attributes of the account A at time t_i. Let T_{scn_i} be the tweets the account posted under the screen name scn_i. Determine if the account A has gone through a misleading repurposing when it changed its screen name scn_{t_1} to scn_{t_2} such that $t_1 < t_2$ using PA_1, PA_2, $T_{scn_{t_1}}$, and $T_{scn_{t_2}}$.

We experiment with four classification strategies based on different feature categories: change of name and description, name/name similarity, profile metadata, and style change. We use a combination for the final classifier.

5.7.1 Change of Name & Description (EDT)

We observe that changing the name and description thoroughly at the same time with the screen name is a behavior that is indicative of misleading repurposing. Thus, we create features to capture this signal based on edit distances. We compute the Levenshtein distance between the string fields and use it as a feature. The formula is as follows:

$$NLD_{\text{attr}} = \frac{\text{lev}(\text{attr}_{U_{\text{prev}}}, \text{attr}_{U_{\text{next}}})}{max(len(\text{attr}_{U_{\text{prev}}}), len(\text{attr}_{U_{\text{prev}}}))} \tag{5.1}$$

where attr is the string attribute, *len* is its length, U_{prev} is the previous snapshot, U_{next} is the next snapshot.

We adopt an online learning approach and train a simple classifier using this feature to sample accounts from the popular dataset. Precisely, we train a decision tree classifier of depth two, which classifies screen name changing instances with $NLD_{name} > 0.721$ and $NLD_{description} > 0.742$.

We chose this classifier because it initially achieved sufficient precision and recall. Thus, we use this classifier as **the baseline** and build other classifiers to improve on it.

5.7.2 Name/Description Similarity (DSIM)

After training the initial baseline classifier, deploying it in the wild, and finding more positives and negatives, we made an observation: non-malicious users who thoroughly change their name and description field leave some artifact that is relevant to the past of the account in order not to mislead their audience, e.g., old screen names, email addresses. We compute features to exploit this behavior. We compute the longest common sequence between the two snapshots' names, screen names, and description fields combined to account for the longest common substring. We computed the raw number and Jaccard coefficient of common tokens between two texts to identify common entities. Finally, we compute the similarity between those two texts using sentence-transformers. We use the model "bert-base-multilingual-uncased" [83] since our data consists of different languages.

5.7.3 Profile Metadata (MD)

We employ additional textual features such as the home page of the profile, the self-stated location, and the profile image. We check if these attributes changed and also compute the edit distance and their NLD (except for the profile image). We also introduce non-textual (numeric) profile attributes: friends count, followers count, statuses count, and favorites count. For each profile attribute in each snapshot, S_i and S_j, we use the raw numbers, a_i and a_j; the difference, $a_i - a_j$; and the ratio of the difference and the maximum to capture the magnitude of the change, $(a_i - a_j)/max(a_i, a_j)$. We also introduce dormancy which is the time passed between two snapshots

5.7.4 Style Change Detection (STY)

If misleading repurposing occurs the style of the tweets may change because the ownership of the account may have changed. We create features based on this assumption using state-of-the-art style change detection techniques [151, 328] and the model "bert-base-multilingual-uncased" [83]. We concatenate the tweets before and after the change of the screen name and treat them as separate *paragraphs*. Iyer et al. [151] predicts the style change between two consecutive *paragraphs* by averaging the sentence vectors of two paragraphs. We produce the sentence vectors by exactly following their method: we split each paragraph into sentences and generate embeddings for each sentence. This

results in a tensor 12 x l x 768 dimensions, where 12 is the number of layers, 768 is the hidden size and l is the length of the sentence (maximum 512 tokens). We first sum the embeddings of the last 4 layers, producing tensors of size l x 768. We then sum this tensor over the first axis to produce a vector of size 768. We generate these vectors for every sentence in each document representing the tweets posted before and after the screen name change and sum them. We then take the average of the two vectors.

5.7.5 Classification

We train each classifier on the data annotated by only A_1. It consists of 512 positives and 910 negatives from integrity data and 421 positives and 248 negatives from popular data. We have 933 positives and 1,158 negatives in total.

We experiment with several supervised machine learning algorithms: SVMs, Logistic Regression, Naive Bayes, Decision Trees, Random Forest, and Neural Nets using sklearn [216]. We experimented with different parameters using grid search. While choosing the best model, we use the integrity dataset as the validation dataset and report the model for each classification strategy that performs the best on this dataset. We use the F1-score to evaluate the performance as the dataset is balanced. We found that Random Forest yielded the best scores consistently, so we only present the results of the Random Forest classification. As we noted before, misleading repurposing is very prevalent in the integrity dataset and even very simple classifiers perform well. Therefore, we test our classifier using the popular dataset. The goal is then to decrease the False Positive Rate among Popular accounts (FPR-P) while still sustaining a high True Positive Rate (TPR-P). Thus, we use Area Under the ROC curve (AUC-P) to evaluate our classifiers' performance. This metric is more reliable than Precision, Recall, and F1-Score when the dataset is imbalanced and positives are more prevalent as it takes FPR into account [33, 169].

5.7.6 Results

All results are presented in Table 5.2. We observe that the classifier EDT that is based only on the change of name and description field performs well on the integrity dataset but not on the popular dataset. This is likely because screen name-changing behavior generally entails misleading repurposing in the former dataset while not necessarily in the latter. We find that extra features based on the name and description field greatly improve this simple classifier because in most cases where the name and description change do not entail misleading repurposing, those fields are either semantically similar or have some traces referring to old snapshots of those fields. Profile metadata features that represent the characteristics of the accounts contribute to the performance of the integrity dataset but only slightly boost the popular dataset.

The style change classifier performs poorly on its own, suffering from a very low recall (57% on integrity accounts and 35.5% on popular accounts). Most of the true positives in the integrity dataset come from the accounts originating from China. As such, it

improves the combinations of other classifiers when used on the integrity dataset but is not as useful when used on the popular dataset. The popular sample only contains 1% of tweets from each user, so it is quite possible that with more data on each account this classifier would perform better. Style change may be more effective in the presence of more data. We leave a comprehensive style change analysis on social media to future work.

The performance of the classifier (EDT-DSIM) is only slightly higher on English accounts compared to Turkish accounts ($AUC = 88.1$ vs $AUC = 87.8$). It performs better on the accounts with the most followers compared to random accounts ($AUC = 89.8$ vs $AUC = 88.0$). This may be because repurposing is more evident in accounts with the most followers as they are more likely to put more indicators in their description fields.

5.7.6.1 Miss-classifications

We manually examined the false negatives and false positives introduced by the BASE-DSIM classifier.

False negatives occur in two cases. First, the account leaves the description field empty, leaving an insufficient amount of information for the classifier. As the popular dataset is collected using the baseline classifier, we only observe this among the integrity accounts. The annotators could annotate those accounts as repurposed due to changes in their names signifying a new person/organization. One limitation of our approach is that it relies on non-empty description fields. However, because accounts with blank description fields are rare, as we discuss below, this limitation is not critical. Second, the classifier captures common slogans and phrases as similar such as "Follow us" and "Updates about x". Such similarities may indicate that the owner of the account is the same or the new owner keeps the style but does not entail the absence of misleading repurposing. A special case is that the purpose of the account changes and it is misleading, but the owner appears to be the same and has the same specialization, so it continues to use buzzwords like Deep Learning. This is generally the case when the personal account becomes an organization account.

False positives occur mainly in two cases. The classifier fails to capture the similarity an annotator can see or there is enough information for an annotator but not for a classifier. An example of the former case is a religious page: the page adopts different names and quotes different religious texts but its purpose is to share religious quotes without a specific agenda (i.e. promoting a specific religious narrative) in both cases. Since there are no repeating texts and semantic similarity is fairly low compared to other examples, the classifier classified this example as positive. Examples for the latter case generally consist of the cases where description fields are entirely or almost empty in one of the snapshots but an annotator can judge that the purpose of the account is the same from the name alone.

5.7.6.2 Estimation

We deployed our classifier on the 1.57 million popular accounts that were active in the first half of 2020. We estimate that 180,689 misleading repurposings by 106,548 accounts. Given that the precision of the classifier is 92.1%, the expected number of screen name changes that are falsely classified as repurposing is 14,275. By May 2022, 22,063 (20.7%) were suspended and 7,800 (7.3%) were deleted. The suspension rate may be low because repurposing is not explicitly against Twitter rules.

5.8 Implications

Repurposing has security implications for Twitter and its users. First, in practice, repurposed accounts are new accounts that steal the followers of the accounts that they repurpose. These followers lend credibility to the account in the eyes of users and Twitter itself. With respect to the former, an account with many followers and an old account creation date are likely to appear more credible to users than a new account with only a few followers [201]. With respect to the latter, Twitter imposes a quality filter to filter out low-quality content and improve the user experience. Although this filter is a black-box, we can presume that the filter considers account age and the number of engagements, because these features have proven useful for bot and spammer detection [34, 143]. Repurposing evades both by stealing the history of an existing account. Twitter also works to detect and suspend accounts that participate in coordinated activities. A malicious user looking to engage many accounts in a coordinated manner may opt to purchase accounts from different sources and repurpose them together since this decreases the likelihood that the set of accounts was used in a coordinated manner in the past and thus, might avoid early detection. Consequently, repurposing *incentivizes* attackers to compromise accounts by creating a market demand for accounts that can be repurposed.

While inflated follower counts may give some credibility to fake accounts, the highest credibility impact comes from the tweets that are retweeted by someone the user knows, making it very important to build a genuine follower base [201]. Repurposing makes it easier for a malicious account to obtain engagements from influential people and/or organizations through a previously non-malicious account. For example, a "parody account" posted comical tweets about president Erdogan in 2014 and attracted 2,500 followers, mostly those with a pro-opposition stance. It was then repurposed twice the account of two different new, short-lived political parties in Turkey. The account was repurposed a third time in 2015 after amassing 7,500 followers, claiming to be an economist living in Canada with degrees from a prestigious university and a profile photo stolen from the web. By 2020, it was followed by many famous journalists and economists and had obtained 62,000 followers, and its tweets criticizing the government were made to the Turkish media [18].

5.9 Generalizability

In this chapter, we focused on Twitter to root out and propose a detection for misleading repurposing. However, we believe that misleading repurposing is a vulnerability for any social media platform that has to be accounted for. While Facebook keeps a log of old names and notifies its followers of the name changes, not all social media platforms have implemented counter-measures against misleading repurposing. For instance, hackers hijacked popular Youtube channels, such as the one by Chilean urban-music artist Aisack, changed their names and pictures to make them look like official Tesla channels, and impersonated Elon Musk, asking for bitcoin to send back the double amount [266]. To the best of our knowledge, the only counter-measure Youtube implements against misleading repurposing is to remove the check-marks from the verified channels that changed their names to prevent impersonation [142]. Channels can change their name with their subscribers intact, and the subscribers do not get a notification of the name change. Thus, misleading repurposing can generalize to Youtube as well as social media platforms with similar policies.

5.10 Counter-Measures

Misleading repurposing is a vulnerability of social media platforms that must be accounted for. The benefits of mitigation efforts are clear: the repurposed accounts are actively causing harm to online/offline communities, and banning such accounts will prevent further harm. However, because there are legitimate reasons to change profile attributes, mitigation is inherently difficult to address, and each solution has its issues.

Firstly, platforms can prevent misleading repurposing by disallowing changes of names or handles that serve as an identifier (e.g., screen names on Twitter). Since repurposing targets popular accounts, they can implement this counter-measure only for such accounts. However, this could impact genuine users who change their names. To remedy this issue, the platforms can implement a process in which users apply to change their names with justifications, and human moderators approve their requests.

The platforms can also notify an account's followers when a name change occurs. However, this disproportionately impacts users who change their names in real life, e.g., married/divorced women and trans users whose friends are notified of the change when they may prefer not to notify.

Lastly, platforms can build methods to detect misleading repurposing by using our framework and the methodology as a starting point and taking action on an individual basis. Researchers can also use our study to root out and investigate repurposed accounts in the wild. We further assist them through a tool we describe in detail in the next section.

5.11 WayPop Machine: A Wayback Machine to Investigate Repurposed Accounts

We propose our own counter-measure against misleading repurposing by enhancing its analysis and verification by researchers. Repurposed accounts often become popular due to the account's previous purpose and then switch their identity, e.g. a meme page accumulates followers. Later, adversaries purchase them to repurpose them to a malicious account, e.g., a political troll. External researchers looking at the up-to-date version of such accounts cannot understand why they are popular, as Twitter does not provide historical data (e.g., follower count of an account at a specific date). To aid them, we present a web application that features the follower growth of users in the past, their viral tweets, their deleted tweets, and any change to their profiles that may signify misleading repurposing. Our application stores, manages, and visualizes data from archive.org's Twitter Stream Grab. To the best of our knowledge, this is the first study to focus on why an account is popular.

5.11.1 Challenges

Our tool addresses two challenges: collecting and effectively using retrospective data.

Analyzing growth strategies of a possibly repurposed account requires users' 1) old tweets, 2) deleted tweets 3) historic profile attributes such as name and follower counts. Regarding the first, the current Twitter API only supports collecting the last 3,200 tweets of any given user, so it is impossible to collect the oldest tweets of an overactive user via the API. For the second, a user may have deleted the tweets relevant to understanding their follower growth. Finally, for the third, the Twitter API does not provide the historical profile attributes, so one cannot know the user's name and the number of followers a user had before and after a given tweet if they did not collect those attributes at the corresponding time.

Second, Twitter is vast, so a comprehensive dataset of user data would be massive and require lengthy processing times to analyze. Additionally, no methodologies or tools exist for analyzing the users' historical data at scale.

Our tool tackles these problems using the Internet Archive's Twitter Stream Grab. The dataset was collected in the past: it contains the old tweets of a given user, and the historic profile attributes at the corresponding time. It is enormous, 4.67 terabytes, when compressed. Our tool effectively stores and visualizes such historical data and provides visualizations explaining the follower growth over time and attribute changes.

5.11.1.1 Examples

Before introducing the specifics of the tool, we first provide brief examples of how our tool can be used to understand how an account became popular or repurposed. Fig. 5.6 shows how our tool analyzes a popular user with 5M followers. We manually identified three points of unusual growth. Further inspection shows that all correspond to "giveaways" in exchange for likes, retweets, and follows. Fig. 5.7 shows the follower growth of another

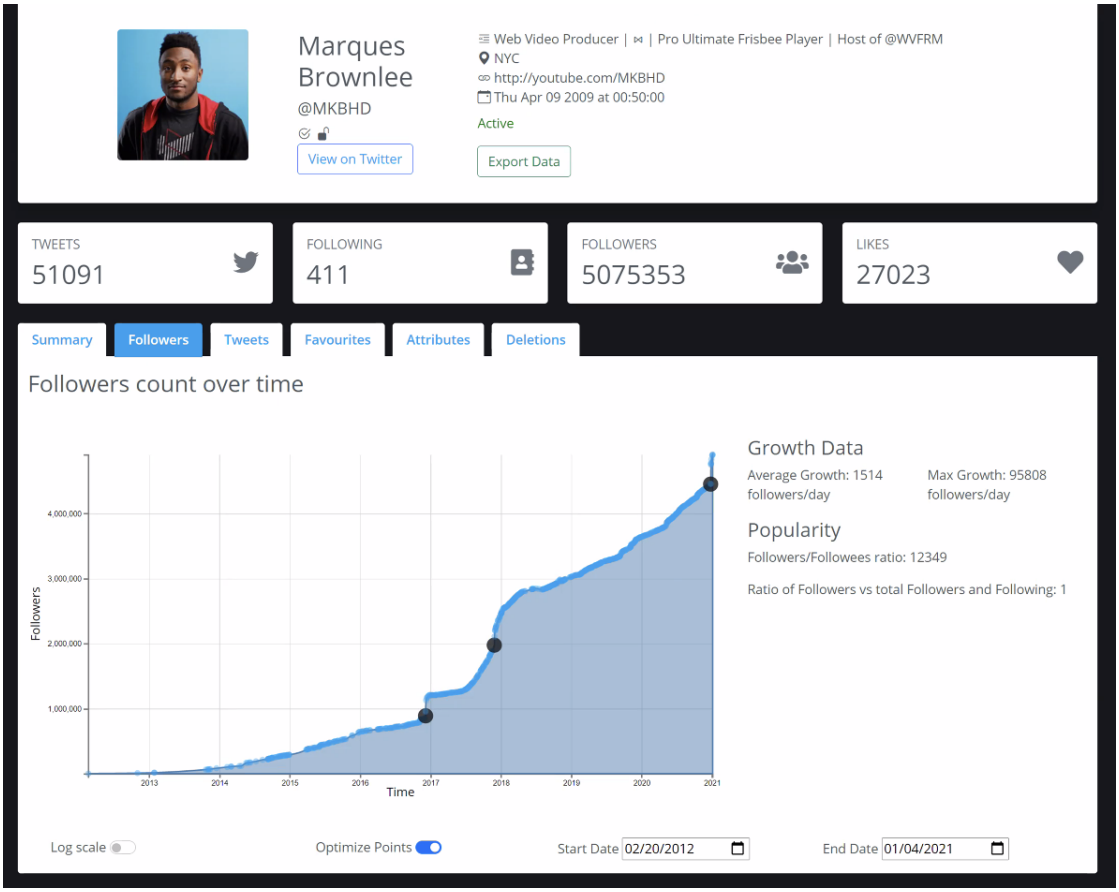

Figure 5.6: The follower growth of the account `@MKBHD`. Three points we highlighed correspond to giveaways in exchange for likes, retweets, and follows.

popular user with 9k followers. Inspection using the tool reveals that this account grew its follower count from 2,200 to 7,800 after a tweet spurred nationalist rhetoric went viral. Prior reports indeed found that this account was fake, as was the information in the viral tweet [84]. Figures 5.15 and 5.16, which we explain in detail later in this section, show how our tool can be used to uncover misleading repurposing. This account, which purports to belong to a Turkish political party, originally belonged to Juliana Knust (`@jujuknust`), a Brazilian actress. Media reports speculated that the account was likely compromised and sold [160]. Our tool corroborates these reports, showing that all of the attributes were changed together, and all of the actress's tweets were deleted.

5.11.2 System Overview

5.11.2.1 Architecture

We used Django for the server-side (backend) programming and MongoDB for managing the database. It uses the NoSQL paradigm to store hundreds of gigabytes of data and efficient querying [150]. Although the underlying database has a compressed size of 115GB, the queries using user id, screen name, or tweet ids are all instantaneous. We used Bootstrap CSS for the front-end programming. Finally, we used D3 (Data-Driven

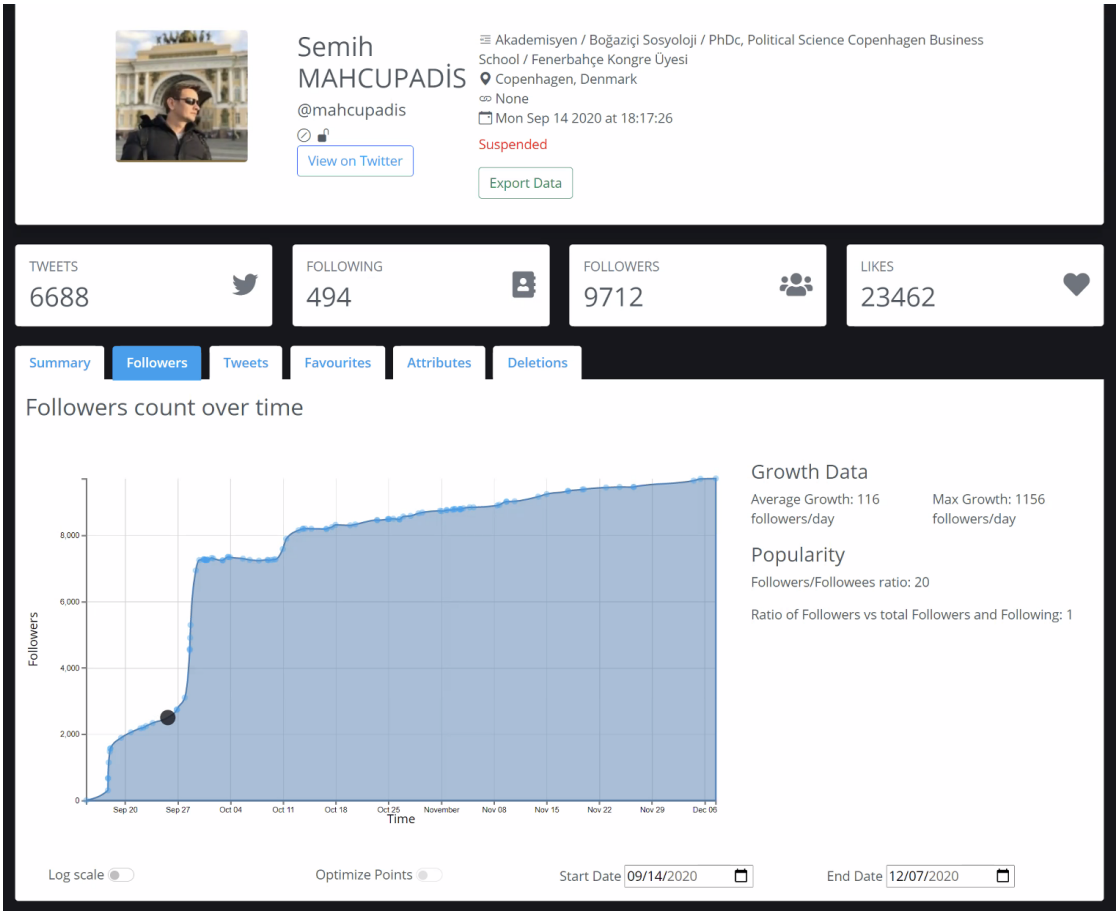

Figure 5.7: The follower growth of the account `@mahcupadis`. The point we highlighted corresponds to the time the account posted a tweet that spurred nationalist rhetoric [84]

Document) for the visualizations. It is a javascript library that allows for creating dynamic and interactive plots in the browser. The architecture is summarized in Fig. 5.8.

5.11.2.2 Data Source

As we argued previously, this tool requires a retrospective Twitter dataset, i.e., one collected in the past. To this end, we utilize archive.org's publicly available Twitter Stream Grab dataset [1]. At the time of the analysis, it consisted of tweets posted between September 2011 and December 2020. Due to the enormous size of this dataset and ethical considerations (see §5.13), we limit the tool to popular active users with more than 5,000 followers who were active in December 2020. Although we utilize this particular dataset, this tool is flexible enough to accommodate other retrospective datasets.

5.11.2.3 Data Structure and Processing

The raw tweet files in Archive's Twitter Stream Grab are quite large because it contains many extraneous attributes. The entire dataset is 4.67 terabytes. We took two important

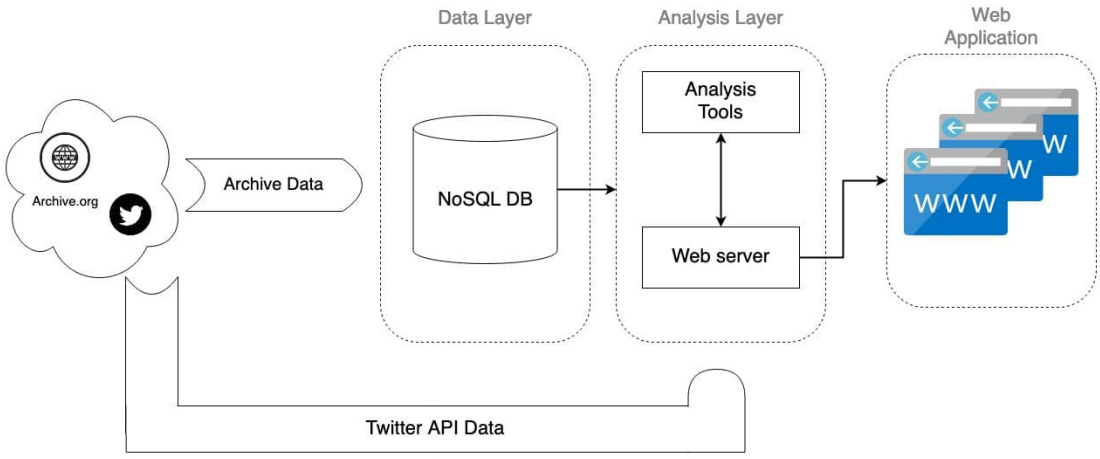

Figure 5.8: WayPop Platform architecture. First, the data is downloaded from archive.org and processed. Then in the data layer, the processed data are stored in a NoSQL database, MongoDB. The web server built using the Django framework communicates with the data layer and further analyzes the data. Lastly, it provides the necessary input for the charts in the web application, which are drawn using the D3 framework. Additionally, the webserver communicates with Twitter API to get up-to-date data.

processing steps to decrease the size. First, because each raw tweet object contains both the tweet and the posting user's information, we split each object into a tweet and a user object. Secondly, we only kept the relevant attributes for each type of object:

Tweet Object: Tweet id, creation date, user id of the poster, the tweet's content in textual form, its public metrics, its source, and, if applicable, its deletion time.

User Object: User id, account creation date, the current and historical screen names (profile handles), names, descriptions, followers, statuses, friends, and favorites counts.

Each user has multiple *data points* that indicate their past profile attributes (e.g., description) and the tweet associated with them. The past attributes are stored in a history array within the user object.

All processing was performed on a single machine using an AMD Ryzen9 3900x 12 core processor with 32 GB of memory. The final uncompressed sizes of the files for the users and the tweets were 71 GB and 200 GB, respectively. The entire process took 30 days.

5.11.3 Features

The tool has several features that provide historical information about a Twitter account. The end-user designates a Twitter account by its screen name or Twitter id. If the account exists in the database, the tool will redirect the user to the page where the account pane and the summary pane summarize the descriptive statistics of the user.

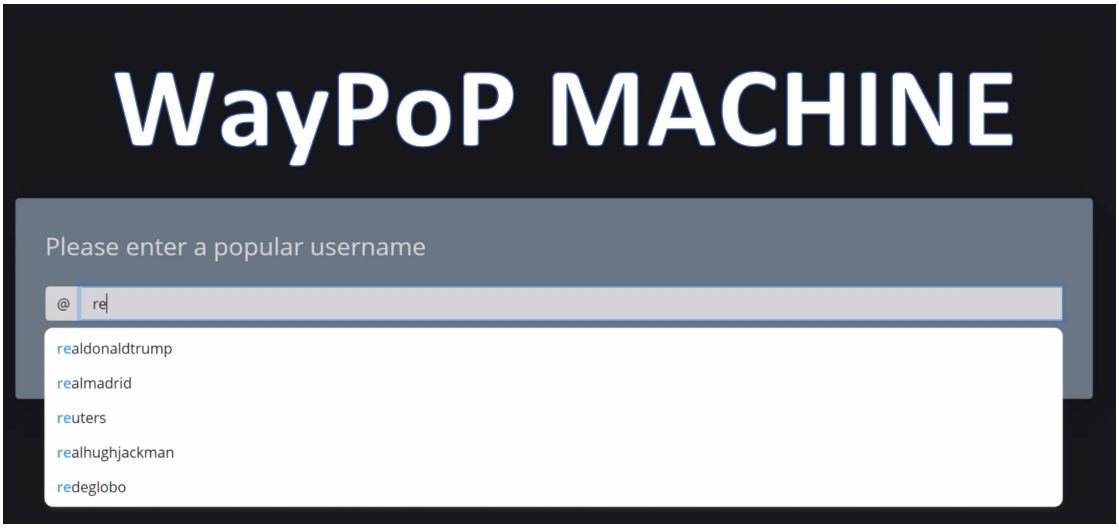

Figure 5.9: The main page of the app. The end-user can enter a screen name or an id as an input.

5.11.3.1 Account Summary

The first page the end-user will be redirected to includes both the account pane and the account summary. The account pane shows up-to-date information about the user profile via Twitter API. If that is unavailable (i.e., the user has been suspended), it shows the most recent profile in the retrospective dataset. The information provided includes the name, screen name, description, home page, location, account creation date, and whether the account is currently suspended, as seen in Fig. 5.10. The user can also export the given user's data, consisting of the tweet object and user object described in §5.11.2.3.

Next, we show the summary of the account. The summary includes the daily rhythm of the user, the number of tweets, retweets and replies, the users the account has retweeted and mentioned, and the sources of the tweets (i.e., the app used to post the tweet). Compared to the account analysis app, which shows statistics from the last 200 tweets of a given user, our tool shows the statistics using all the retrospective data. That way, the user can see statistics of old or deleted tweets.

5.11.3.2 Follower Growth

The key to the tool is the historical follower growth, which attempts to help us understand why a user is popular. It does so by visualizing the follower growth in a chart in which the x-axis denotes the time, and the y-axis denotes the follower count if the corresponding data point is available. Hovering on the data point shows the corresponding tweet for which the chart shows the time and follower count.

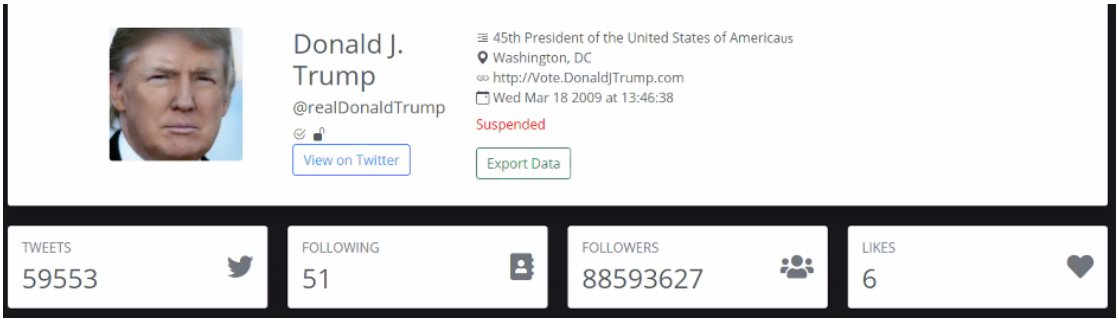

Figure 5.10: The account pane shows the attributes of the Twitter account. If the account is still active, it shows the up-to-date information collected using Twitter API. Otherwise, it uses the most recent data in the dataset.

5.11.3.3 Tweets

The tweet panel complements the follower growth to corroborate the results further and provide more information about the virality of tweets. On this panel, the tool provides the cumulative number of tweets via a chart. Displayed to the right of the chart are the most engaged tweets. Engagements are the sum of retweets, quotes, and favorites. They act as another proxy for viral tweets, providing a more complete picture of popularity.

5.11.3.4 Favorites

On this panel, the tool provides the cumulative number of the favorites (i.e., likes) that the user's tweets had over time. Large drops in the favorites count signal that an account has been repurposed. Recall that repurposing entails a new owner taking over an account and deleting the past favorites (along with other content and attributes). Since no retrospective dataset includes users' likes and there is no API call to collect a user's favorites, we do not show the favorite tweets of the users. Instead, we provide their most-liked tweets. The most liked tweets are different from the most engaged tweets. This is because engagements are not necessarily endorsements, i.e., users may quote other users to criticize them. Therefore, likes are a stronger proxy for endorsements, as the word *like* itself self-states that the user endorses the tweet.

5.11.3.5 Change of Attributes

Our tool provides the given user's past attributes on this panel: name, screen name, and description field. It indicates the attribute changed and the time the change is first observed. We use Levenshtein distance to quantify the extent of the change of a given attribute. The visualization is in the form of a graph. The x-axis indicates the time that the change is first observed. The y-axis is the sum of the Levenshtein distances between pairs of names, screen names, and the description fields. The user can hover on the bars indicating changes to see the previous and the new text of the attribute, as Fig. 5.15 shows.

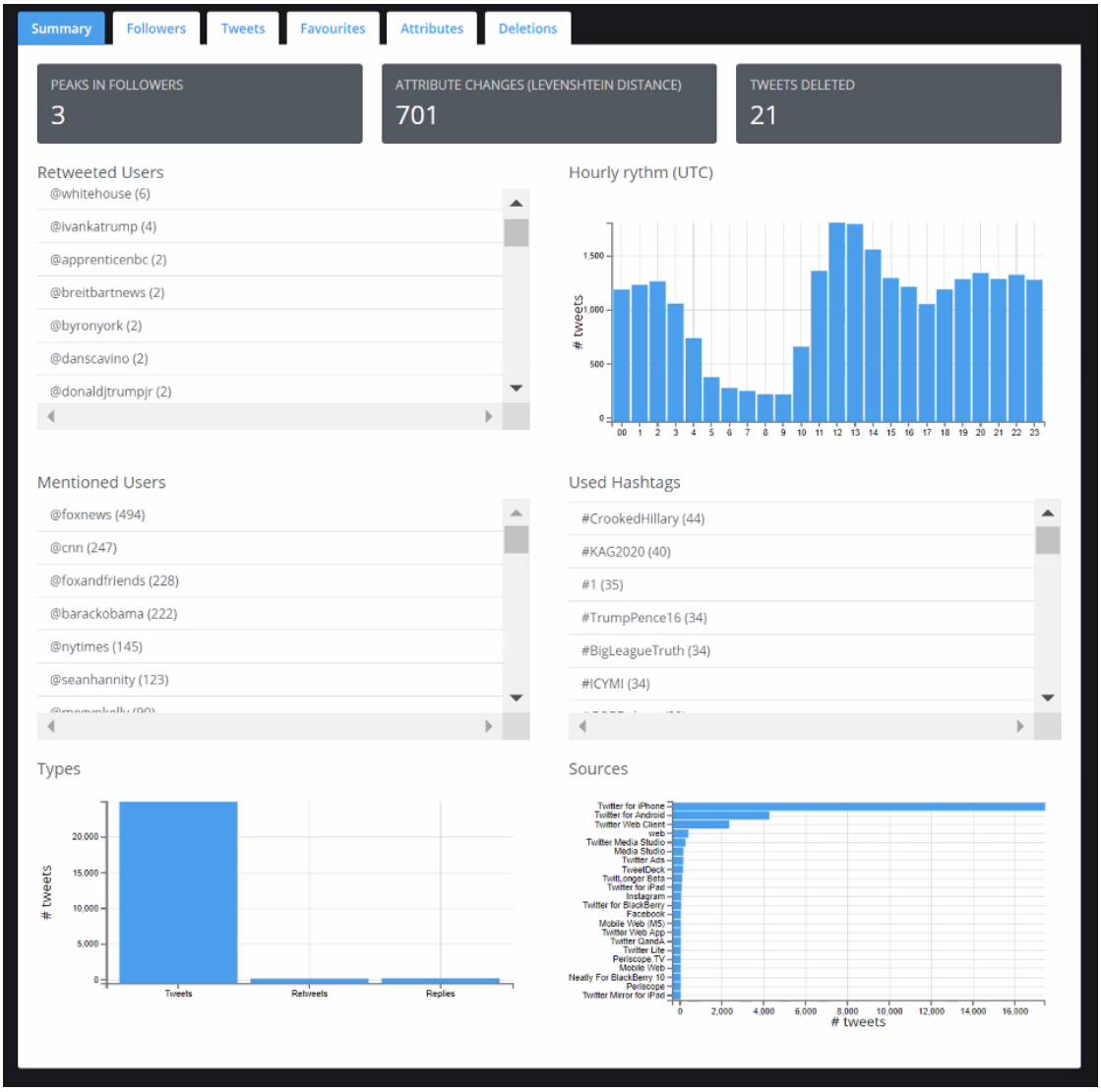

Figure 5.11: The summary including statistics of a Twitter account. The tool provides information on the daily rhythm of the user, the number of tweets, retweets, and replies, the users the account has retweeted and mentioned and the sources of the tweets (i.e. the app that is used to post the tweet) using the retrospective dataset.

5.11.3.6 Deletions

On this panel, we show the number of deleted tweets of the user over time. Sudden peaks in the number of deleted tweets indicate purging behavior, similar to sudden drops in likes. Our tool also shows the text of the deleted tweets when the user hovers over the chart. As seen in Fig. 5.16, the compromised account of Juliana Knust, now representing a Turkish political party, had deleted 30 tweets within the same day. These tweets belonged to the old owner of the account as they are in Portuguese.

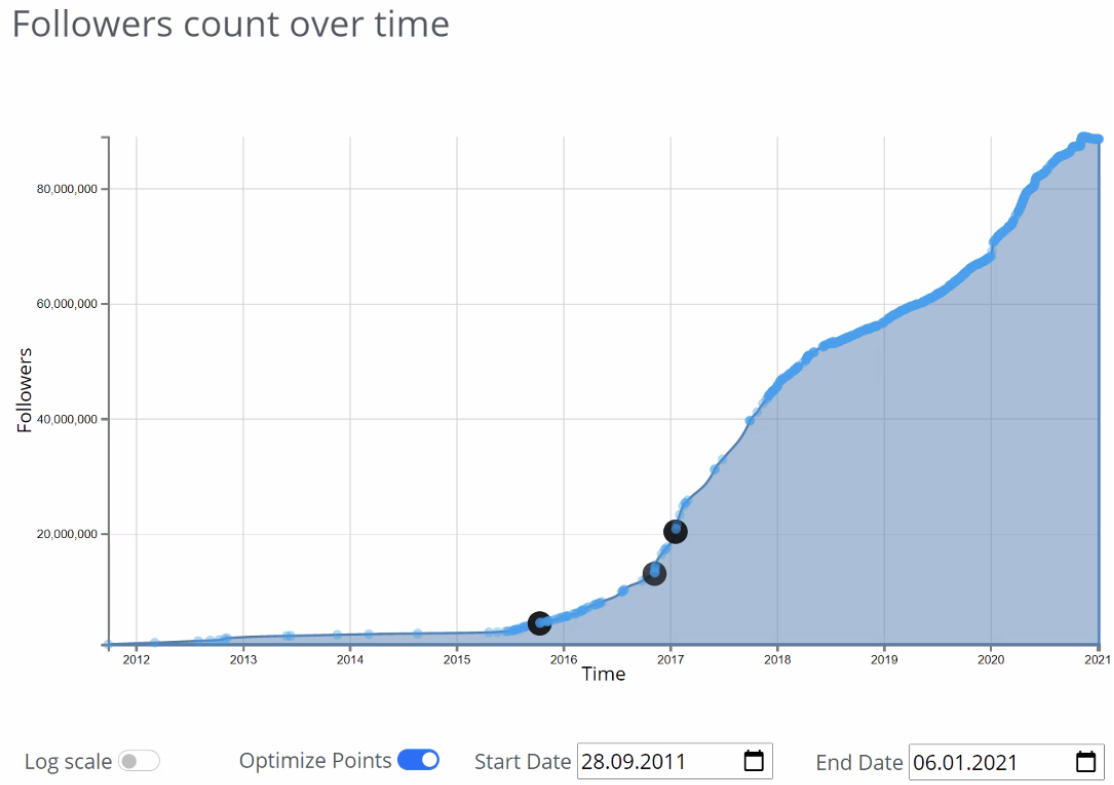

Figure 5.12: Follower growth of @realdonaldtrump. We highlight three points: him comments on the Democratic Debate in 2015, his winning of the election in 2016, and him entering the White House.

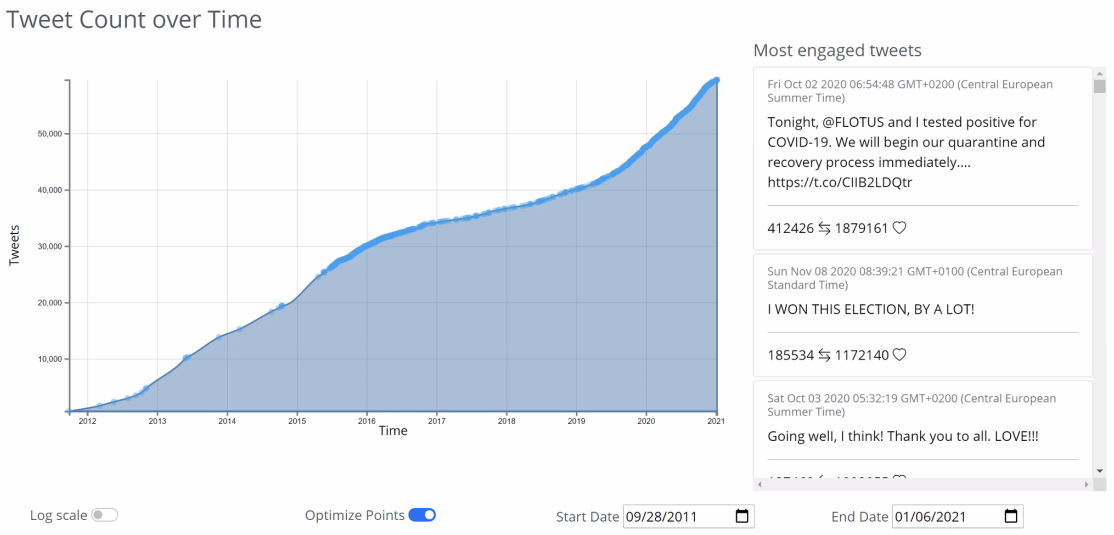

Figure 5.13: Trump's tweet count over time and his most engaged tweets.

5.12 Limitations

The primary caveat of this work is that it is limited by the choice of dataset, which contains only 1% of all tweets. This makes the study biased towards popular users in

Figure 5.14: Favorite count of Trump's Twitter account over time. The favourites are purged several times.

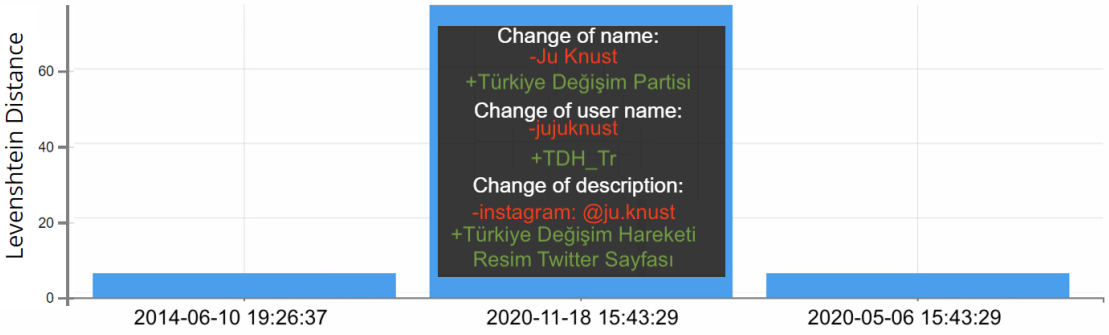

Figure 5.15: Changes of attributes of the old account of Juliana Knust, now owned by a political party. Its attributes changed dramatically on November 18, 2020.

terms of preciseness (those who are retweeted more often show up more often in the 1% sample, as retweets contribute to the sampling). Since we already limit our study to popular users, this limitation does not introduce a critical problem, but still has to be acknowledged.

Another caveat is that the tool we propose is limited to social media data. Interactions on social media are not the only factor in a user's popularity: external factors also, or even often, play a role. For instance, our tool found that Donald Trump had peak points / unusual growths when he officially won the 2016 election and when he entered the white house. Although he had highly engaged tweets at this peak point, the main factor of follower growth was caused by external events. Additionally, the social media is vulnerable to manipulation and both the follower counts and the public metrics of tweets can be inflated. Follower growth can be due to fake followers and engaged tweets

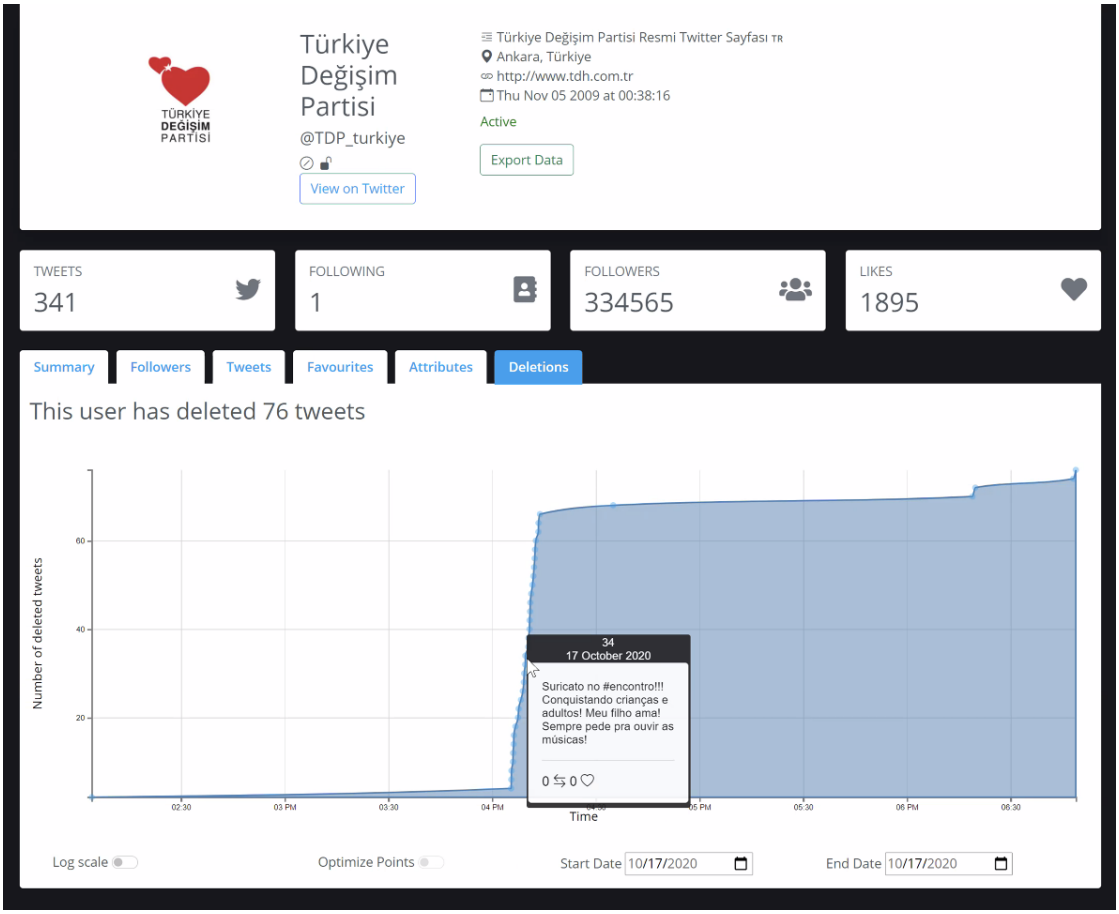

Figure 5.16: The deletion statistics of the old account of Juliana Knust, now owned by a political party. 76 tweets were purged on October 17, 2020.

can be due to bots hired to like and retweet a target tweet.

5.13 Ethical Implications

5.13.1 Data Collection and Management

This study only uses the public data provided by Twitter and the Internet Archive, both of which have been analyzed extensively by previous work. We do not use or store any other data. To comply with the Twitter Terms of Service and protect the privacy of Twitter users, we do not share the data of repurposed accounts from the popular dataset. However, we will share the code and the ids of the repurposed accounts from the integrity dataset, since these accounts have already been made public by Twitter and, as such, there is no risk of further harms in their release.

To further prevent violations of user privacy, we limit our tool to internal use. We only use it to study malicious users that disrupts the public dialogue. Researchers who would like to use our tool are responsible for downloading the data from the Internet Archive and processing it. Providing safe and responsible access to our tool is our future

work. Our tool will be available and usable only as long as the Internet Archive continues to provide Twitter Stream Grab.

5.13.2 Threats to User Anonymity and Privacy

We additionally mitigate any privacy loss to normal Twitter users by limiting our study to only two types of accounts: 1) accounts in the civic integrity dataset which have been designated by Twitter as harmful to public dialogue and released by Twitter, and 2) popular accounts which can influence the public. For an account to be considered "popular", we follow Twitter's lead in choosing a threshold of 5,000 followers, the threshold Twitter uses in the civic integrity dataset to determine if a user's profile be made public. This group of accounts does include legitimate users who do not intend to mislead others or participate in malicious activity, and in the course of our study, we uncovered their former account names/old profiles via parsing publicly available data. This may include accidental deanonymization of a currently pseudonymized account if the user self-stated their identity in an old version of their profile and posted enough tweets from the old version of their account to appear in the 1% archive sample. We mitigated this risk to the best of our availability by not releasing the data publicly, performing the annotation ourselves to not expose the data to crowd workers, and not reading their tweets.

5.13.3 Further Potential Impacts of Our Work

We must also consider the impact of publishing such a study and making this type of platform manipulation known to the general public and academic community. First, we hope that this work raises awareness among Twitter users that accounts that they follow may be repurposed for malicious purposes so that they can notice such accounts when they see them, and possibly even report them as malicious. We also hope that pointing out and studying this phenomenon urges academics and Twitter alike to put more resources into mitigation methods that do not have negative impacts on normal users, especially those from already marginalized groups.

Awareness goes both ways, though, and this study could also lead to malicious users learning about repurposing. This could lead to some who did not know that repurposing was possible to maliciously repurpose more accounts. However, we know from the widespread use of malicious repurposing that this phenomenon is already known by many who wish to use it maliciously. By bringing this problem to light, we hope to mitigate this risk by promoting user and platform awareness, thus discouraging its use.

We must also examine how our methodology could be used misused. Although the goal of this study is to uncover malicious repurposing, parts of our methodology could be repurposed to deanonymize users who want to remain anonymous, as long as at one point in the past their account had an identifiable attribute. Users should be made aware that if they wish to remain anonymous, a new account should be created from scratch rather than repurposing a non-anonymous account.

Finally, this work further illustrates that deletion privacy is important for users [296], but that it also can prevent malicious activity from being discovered. While users need to be able to delete and hide their prior activities and accounts, this study underlines how such mechanisms can be misused to mislead and deceive users. This balance is difficult to find, and further research is needed to understand users' opinions and understandings of deletion privacy.

5.14 Summary

In this chapter, we defined and describe misleading repurposing, a social media manipulation that was yet to be studied. We proposed a methodology that consists of collecting, annotating, and detecting the repurposed accounts on Twitter. Our detection methodology contributes to the detection of compromised accounts as it presents a solution to discovering compromised accounts that do not observe anomalous behavior. We characterized misleading repurposing and found that adversaries target accounts with high follower counts. We found over 100,000 accounts that may have been repurposed and should be investigated. We also present a tool to assist the investigation of repurposed accounts by visualizing their profile changes and follower growth. We believe our study can assist in detecting and analyzing misleading repurposing in other contexts.

Chapter 6

Conclusion

In this chapter, we first summarize our contributions. We then discuss new directions to work presented here. We conclude with recommendations for the platforms and the researchers.

6.1 Summary of Contributions

The advance of disinformation instigated platforms and researchers to understand and mitigate social media manipulation. In this thesis, we provided background on social media manipulation and contributed to its research by introducing the role of compromised accounts. We now briefly summarize our contributions in this thesis.

Analysis In Chapter 3, we presented a case study in which we defined and analyzed how compromised accounts were used to manipulate social media. We define and describe a new attack on social media platforms. We propose a new bot dataset. We discuss the implications of this attack on platform security, user security, and society in general. We also present our counter-measure. This was one of the first case studies that analyzed how adversaries employ compromised accounts to manipulate social media.

Implications In Chapter 4, we analyzed a dataset of retweet bots that were compromised accounts. We characterized these bots by comparing them to known humans. We discuss our findings' implications for bot research and detection systems that used this dataset. This was the first study focusing on retweet bots and how their characteristics have an impact on bot studies.

Detection In Chapter 5, we proposed a novel detection approach to compromised accounts that are repurposed. We defined and described a new social media manipulation, "misleading repurposing". We proposed a framework to annotate accounts that are repurposed. We proposed a new, labeled dataset of repurposed accounts. We then detected misleading repurposing in the wild and described their characteristics. We presented a tool to study such accounts. This was the first study that proposed an approach to detect compromised accounts that were repurposed and do not observe anomalous behavior.

6.2 Future Work

Detecting Manipulations: Studies propose systems to detect bots and other kinds of malicious accounts. However, only a few propose systems to detect "manipulations": what message adversaries promote and how they do that. Additionally, such studies do not analyze if the proposed system can cope with scalability, performance, and data collection limits and work in real life. In Chapter 3, we made the first attempt to deploy a Twitter bot to detect and announce fake Twitter trends in Turkey to Twitter users. In the future, we will extend this bot to report trend manipulations in other countries and other types of manipulations.

Detecting Retweet Bots: Retweet bots promote Twitter content artificially and may be used to amplify malicious content. In Chapter 4, we propose a dataset of retweet bots to characterize by comparing them to known humans. We found that classifying retweet bots using the current datasets is trivial: the percentage of retweets in a user's full activity clearly distinguishes the retweet bots in our new dataset and the known humans in the datasets shared by previous studies. Our caveat is that legitimate users may also have a high retweet percentage. None of the human datasets we acquired from previous students were collected due to accounts' high retweet percentage. Since some of these studies rely on human annotation, it may also be possible that the annotators may have classified users with a high retweet percentage as bots. Thus, to detect retweet bots as reliable, we either need to build a dataset of humans with a high retweet percentage or prove that humans generally do not have a high retweet percentage.

Detecting Viral Social Media Posts: In Chapter 5, we propose a tool that visualizes the follower growth of users. We will extend this work to detect viral tweets in social media to better what kind of tweets go viral and how this affects the users. To this end, we will propose a more reliable detection strategy.

Analysis of Platform Amplifications: Many social media platforms amplify popular content (i.e., trends) by publishing on the main page like Spotify and Twitter, or through Explore page like Instagram, or by keeping a curated list in a dedicated Trends section such as Youtube. They expose every user on the platform to these amplifications. However, only a few studies analyze these trends' content and impact. Our work in Chapter 3 was our first attempt to analyze the trends. We will continue with an extensive analysis of the trends in the future and study how they may influence the public.

Feedback Loops due to Manipulations: In Chapter 3, we showed that adversaries create fake trends and bring them to public attention. In some cases, the public has picked up the trend and discussed it thoroughly, keeping it trending even though adversaries stopped attacking. One trend even became the agenda and made to social science studies. In other words, the adversaries create a *feedback loop* by sparking the public

interest in the topic artificially. In future work, we will analyze such feedback loops and what type of trends can create them.

Other Platforms: Our work was focused on Twitter due to its ease of data collection and analysis and more flexible terms of service. Thus, we introduce and describe novel social media techniques used by adversaries only on Twitter. However, we believe it is of public interest to test if and how these techniques translate to other platforms, which could be future work.

6.3 Recommendations

We conclude with recommendations to researchers and platforms on the directions to further enhance the research in this area.

6.3.1 Impact

Social media manipulation targets individuals, the public, and platforms. However, its impact is currently understudied. Most of the current work is quantitative or theoretical. For instance, Ross et al. argue that bots can shape people's behavior on opinion expression by aggressively promoting the minority opinion and deceiving the majority, making them believe they are the minority. This would silence the majority, called a "spiral of silence". The authors demonstrate that social media bots can make an impact by creating the spiral of silence using simulations over synthetic social networks [234]. However, whether those simulations would translate to real networks is unclear. Researchers previously analyzed the public engagements with the bots they deployed [12, 127]. However, to the best of our knowledge, there is no large-scale analysis of the impact of bots in general.

We also recommend that researchers focus on detecting manipulative messages and bots' amplifications towards them instead of tackling the task of detecting bots. Bots may not impact the public if they only interact with each other on social media. More important is which harmful messages are brought to public attention due to adversaries employing bots or other social media manipulation techniques.

6.3.2 Transparency

Increased transparency on the type of data the platforms share may facilitate research on the impact. Researchers were limited to analyzing the engagements of the bots they deployed because they do not have access to others' engagements that are passive. In other words, the platforms provide the data of active engagements (like, share, retweet) but do not provide the data of passive engagements such as who screened or read the content. This makes it challenging to map the true extent of social media manipulation [214]. For instance, Twitter notified 1.4 million users exposed to the content amplified by Russian trolls during the 2016 U.S. elections but did not disclose those users. Dutta et

al. could identify 860,000 users who were exposed to such accounts. However, their research is limited to users that actively engage with those accounts (through replies and retweets) [91]. This may discard the users who passively engage with (i.e., only read) the content amplified by Russian trolls and may introduce a bias.

In chapter 3, we analyzed fake trends created by bots. Although we argued that at least some of those trends had an impact as they made to the media, we could not perform a reliable quantitative analysis on the impact of fake trends by analyzing engagements. This is because Twitter does not provide the information on which or how many users see or engage with trends or the posts mentioning the trends. Enhanced access to such data would help us extend our work.

We also advise platforms to increase transparency on sharing the data of malicious actors. Currently, platforms only share the reports or data of state-sponsored actors. However, adversaries may not always be affiliated with governments. Additionally, non-state-sponsored adversaries' behaviors and techniques may differ from their state-sponsored counterparts, and thus, their analysis may be insightful to the research community. For instance, Twitter announced that they suspended more than 70,000 accounts that "were engaged in sharing harmful QAnon-associated content at scale and were primarily dedicated to the propagation of this conspiracy theory across the service." [281]. However, the platform did not disclose the data of those accounts as it did with the accounts originating from Russia. Thus, analyzing the QAnon activity by collecting the data after Twitter suspensions may be significantly limited.

6.3.3 Public Awareness

One counter-measure to social media manipulation is raising public awareness by informing the public of manipulation instances. Studies propose classifiers, visualizations, tools, and defenses toward this goal. However, we observe that some of these tools are not accessible to the public or are not used. We advise researchers and journalists to better collaborate on using tools to analyze and report social media manipulation for the public. We also advise researchers to turn to more visible solutions that require less effort from the end-user. For instance, reports on social media manipulation can be communicated through social media instead of a tool deployed on an independent website. We observe that our reports on fake trends reach up to 500,000 views on Twitter, which may be hard to accumulate on independent websites.

6.3.4 Fairness in Content Moderation

We also emphasize that the platforms should be fair and enforce their policies in all regions equally. We observe unequal attention to regions where platforms make the most profit from or have the linguistic competence, such as the United States. The lack of content moderation in other regions had and may continue to have a detrimental impact on the public, especially the vulnerable populations. For instance, In Myanmar, ultranationalists targeted Muslims and framed them as potential terrorists. Their

Facebook posts that spread hate and disinformation about Rohingya Muslims incited violence offline. Facebook was late to react manipulation of its platform [116]. According to Reuters, Facebook had only four Burmese speakers reviewing content in Myanmar which had 7.3 million active users, and they were working outside of the country [254]. Similarly, Facebook and Instagram's automated systems confused Nigeria's #EndSARS movement against the police violence with the disease SARS and flagged the posts containing #EndSARS as false information [267]. In chapter 3, we reported ongoing attacks on Twitter trends which we only observe in Turkey. The attacks began in 2015 and are not prevented as of 2022, demonstrating the problem with content moderation on Twitter Turkey. We also suggest that researchers study understudied communities to encourage platforms to take precautions and spark public and media attention. Our work, published in 2021, was the first literary work to bring up the issue to the public six years after the attacks began.

Bibliography

[1] Archive team: The twitter stream grab. https://archive.org/details/twitterstream. 33, 66, 96, 113

[2] L. B. Aal, J. N. Parmar, V. R. Patel, and D. J. Sen, "Whatsapp, skype, wickr, viber, twitter and blog are ready to asymptote globally from all corners during communications in latest fast life," *Research Journal of Science and Technology*, vol. 6, no. 2, p. 101, 2014. 12

[3] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Transactions on Information Systems (TOIS)*, 2008. 93

[4] J. Abbruzzese and B. Zadrozny, "After epstein's suicide, trump boosts conspiracy theories flourishing online," *NBC News*, 2019. 26

[5] H. Abdelhaq, C. Sengstock, and M. Gertz, "Eventweet: Online localized event detection from twitter," *Proceedings of the VLDB Endowment*, vol. 6, no. 12, pp. 1326–1329, 2013. 29

[6] F. Abdulrahman and A. Subedar, "How much to fake a trend on twitter? in one country, about £150," *BBC News*, 2020. 29

[7] N. Abu-El-Rub and A. Mueen, "Botcamp: Bot-driven interactions in social campaigns," in *The World Wide Web Conference, WWW*, 2019. 18

[8] A. Addawood, A. Badawy, K. Lerman, and E. Ferrara, "Linguistic cues to deception: Identifying political trolls on social media," in *Proceedings of the international AAAI conference on web and social media*, vol. 13, 2019, pp. 15–25. 18

[9] S. Afroz, A. C. Islam, A. Stolerman, R. Greenstadt, and D. McCoy, "Doppelgänger finder: Taking stylometry to the underground," in *2014 IEEE Symposium on Security and Privacy*. IEEE, 2014. 93

[10] R. Agrawal, R. Srikant *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215. Santiago, Chile, 1994, pp. 487–499. 29

[11] J. Aguerri, M. Santisteban, and F. Miró-Llinares, "The fight against disinformation and its consequences: Measuring the impact of "russia state-affiliated media" on twitter," 2022. 20

[12] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, "People are strange when you're a stranger: Impact and influence of bots on social networks," in *Proceedings of the international AAAI conference on web and social media*, vol. 6, no. 1, 2012, pp. 10–17. 125

[13] N. Albadi, M. Kurdi, and S. Mishra, "Hateful people or hateful bots? detection and characterization of bots spreading religious hatred in arabic social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–25, 2019. 12

[14] Alex Hern, "Anti-porn filters stop dominic cummings trending on twitter," https://www.theguardian.com/politics/2020/may/27/anti-porn-filters-stop-dominic-cummings-trending-on-twitter, 2020, accessed: 2022-11-07. 21, 40

[15] S. Ali Alhosseini, R. Bin Tareaf, P. Najafi, and C. Meinel, "Detect me if you can: Spam bot detection using inductive representation learning," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 148–153. 18

[16] M. Alizadeh, F. Gilardi, E. Hoes, K. J. Klüser, M. Kubli, and N. Marchal, "Content moderation as a political issue: The twitter discourse around trump's ban," *University of Zurich*, 2021. 20

[17] Anonymous, "Turks manipulate trending topics in spite of twitter," *Hurriyet Daily News*, 2014. 29, 32, 46, 48

[18] ——, "Başkanlık tweeti rekor kırdı," *Odatv*, 2015. 109

[19] ——, "$129 million turkish farmville-clone scammer found in uruguay," *Ahval News*, 2018. 54

[20] ——, "Erdoğan'a botlar "tamam" millet "devam" diyor," *Yeni Akit*, 2018. 32

[21] ——, "Vatandaşlar sosyal medyayı salladı: Süresiz nafaka zulümdür!" *Yeni Akit*, 2018. 55

[22] ——, "Adnancı ve furkancılar twitter'ı istila etti!" *Medya Radar*, 2019. 48

[23] ——, "Akp launches social media campaign to counter opposition hashtag for istanbul election," *Turkish Minute*, 2019. 55

[24] ——, ""Çünkü Çaldılar" sosyal medyada dev destek," *A Haber*, 2019. 55

[25] ——, "Ekrem İmamoğlu: They campaign by making children say 'they stole'," *Bianet*, 2019. 55

[26] ——, "Türk telekom bot hesaplarla algı operasyonu mu yapıyor?" *Medya Faresi*, 2019. 32

[27] ——, "Twitter'da #mazbatamızıverin ve #hırsızekrem savaşı," *BBC Türkçe*, 2019. 55

[28] ——, "'because they stole' statement by yıldırım: I had to, i cannot express myself," *Bianet*, 2019. 55

[29] E. Bal, "Twitter'da nasıl tt olunur?" *BBC Türkçe*, 2014. 32, 46

[30] P. Barberá, "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data," *Political analysis*, vol. 23, no. 1, pp. 76–91, 2015. 17

[31] S. Barbon, R. A. Igawa, and B. B. Zarpelão, "Authorship verification applied to detection of compromised accounts on online social networks," *Multimedia Tools and Applications*, 2017. 28

[32] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Third international AAAI conference on weblogs and social media*, 2009. xv, 49

[33] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J Inf Eng Appl*, 2013. 107

[34] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, 2010. 2, 92, 109

[35] D. M. Beskow and K. M. Carley, "Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter," in *Conference paper. SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 2018. 78

[36] M. Bhargava, P. Mehndiratta, and K. Asawa, "Stylometric analysis for authorship attribution on twitter," in *International Conference on Big Data Analytics*. Springer, 2013. 93

[37] Bir1iki2üç3456, "Twitterda şu an gündem olan #suriyelilerdefolsun hashtag'ı hakkında ne düşünüyorsunuz?" https://www.kizlarsoruyor.com/toplum-sosyal-iliskiler/q10099852-twitterda-su-an-gundem-olan-suriyelilerdefolsun-hashtag-i-hakkinda, 2018, accessed on 2020-07-27. 55

[38] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, 2008. xv, 49, 50

[39] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, pp. 295–298, 2012. 9

[40] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in *Proceedings of the 27th annual computer security applications conference*, 2011. 64

[41] H. Boyacıoğlu, "Turkey's trade ministry launches probe into 11 companies over fraudulent ponzi scheme," *Hurriyet Daily News*, 2019. 54

[42] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104. 29, 60

[43] J. Bright, "Explaining the emergence of echo chambers on social media: the role of ideology and extremism," *Available at SSRN 2839728*, 2017. 9

[44] M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, "Authorship verification for short messages using stylometry," in *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*. IEEE, 2013. 93

[45] C. Cadwalladr and E. Graham-Harrison, "Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach," *The Guardian*, 2018. 28

[46] camaraobscura, "My instagram keeps following people i don't know on its own. pls help :(. accessed on 2022-11-07," http://tiny.cc/randomfollowers, reddit. 84

[47] B. Cansu, "Suriyeliler hedef halinde," *Birgün Gazetesi*, 2018. 55

[48] J. Cao, P. Qi, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the role of visual content in fake news detection," *Disinformation, Misinformation, and Fake News in Social Media*, pp. 141–161, 2020. 16

[49] M. Carman, M. Koerber, J. Li, K.-K. R. Choo, and H. Ashman, "Manipulating visibility of political and apolitical threads on reddit via score boosting," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2018, pp. 184–190. 14, 26

[50] J. M. Carrascosa, R. González, R. Cuevas, and A. Azcorra, "Are trending topics useful for marketing? visibility of trending topics vs traditional advertisement," in *Proceedings of the first ACM conference on Online social networks*, 2013. 25

[51] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 675–684. 17

[52] R. Çakır, "#suriyelilerdefolsun: Peki, sonra?" *Medyascope*, 2018. 55

[53] D. Çakır Demirhan, "Türkiye'de geçici koruma kapsamında bulunan suriyelilerin ulusal vatandaşlık rejimine etkisi," *Sosyal Bilimler Enstitüsü*, 2020. 54, 55

[54] O. Cengiz and C. Cengiz, "Political economy of anti-refugee in turkey and cyber-racism: Thematic analysis of #suriyelilerdefolsun hashtag," in *Handbook of Research on the Political Economy of Communications and Media*, 2020. 54, 55

[55] N. Chavoshi, H. Hamooni, and A. Mueen, "Debot: Twitter bot detection via warped correlation." in *ICDM*, 2016. 18, 28, 65

[56] ——, "Temporal patterns in bot activities," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017. 28

[57] Y.-R. Chen and H.-H. Chen, "Opinion spammer detection in web forum," in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015. 56

[58] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?" in *Proceedings of the 26th annual computer security applications conference*, 2010. 81

[59] ——, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Dependable Sec. Comput.*, 2012. 2, 28

[60] Z. Chu, I. Widjaja, and H. Wang, "Detecting social spam campaigns on twitter," in *International Conference on Applied Cryptography and Network Security*, 2012. 28, 64

[61] B. C. Cohen, *Press and foreign policy*. Princeton university press, 2015, vol. 2321. 7

[62] L. Cohen, "#untrendoctober wants to end twitter's trending topics before the election," *Daily Dot*, 8 2020. 26

[63] N. Cohen, "Will california's new bot law strengthen democracy?" *The New Yorker*, vol. 2, p. 2019, 2019. 20

[64] N. I. Council, "Assessing russian activities and intentions in recent us elections," 2017. 15

[65] M. Coxall, *Human Manipulation-A Handbook*. Malcolm Coxall-Cornelio Books, 2013. 8

[66] Craig Timberg, Drew Harwell, Hamza Shaban, Andrew Ba Tran and Brian Fung, "The new zealand shooting shows how youtube and facebook spread hate and violent images — yet again,"

https://www.washingtonpost.com/technology/2019/03/15/facebook-youtube-twitter-amplified-video-christchurch-mosque-shooting/, 2020, accessed: 2022-11-07. 21

[67] K. Crawford and T. Gillespie, "What is a flag for? social media reporting tools and the vocabulary of complaint," *New Media & Society*, vol. 18, no. 3, pp. 410–428, 2016. 20

[68] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake twitter followers," *Decision Support Systems*, 2015. 1, 15, 28, 63, 64, 65, 67, 78, 81, 92

[69] ——, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th international conference on world wide web companion*, 2017. 2, 67, 71

[70] ——, "Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling," *IEEE Transactions on Dependable and Secure Computing*, 2017. 65, 67, 80

[71] S. Cresci, F. Lillo, D. Regoli, S. Tardelli, and M. Tesconi, "Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter," *ACM Transactions on the Web (TWEB)*, 2019. 65, 67, 81

[72] S. Cresci, M. Petrocchi, A. Spognardi, and S. Tognazzi, "On the capability of evolved spambots to evade detection via genetic engineering," *Online Social Networks and Media*, 2019. 28

[73] ——, "On the capability of evolved spambots to evade detection via genetic engineering," *Online Social Networks and Media*, 2019. 64

[74] S. Cresci, R. D. Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017. 28

[75] E. D. Cristofaro, A. Friedman, G. Jourjon, M. A. Kâafar, and M. Z. Shafiq, "Paying for likes?: Understanding facebook like fraud using honeypots," in *Proceedings of the 2014 Internet Measurement Conference, IMC*, 2014. 26

[76] M. Şafak Sarı, "Dezenformasyon nedir? sosyal medya savaş meydanı oldu," *Journo*, 2020. 44

[77] Şükrü Oktay Kılıç, "Twitter'da gündem oyunları," *Al Jazeera Türk*, 2014. 32, 48

[78] K. Darwish, D. Alexandrov, P. Nakov, and Y. Mejova, "Seminar users in the arabic twitter sphere," in *International Conference on Social Informatics*. Springer, 2017, pp. 91–108. 11

[79] K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov, "Unsupervised user stance detection on twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 141–152. 17

[80] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in *Proceedings of the 25th international conference companion on world wide web*, 2016, pp. 273–274. 22

[81] ——, "Botornot: A system to evaluate social bots," *Proceedings of the 25th International Conference on World Wide Web, WWW*, 2016. 28

[82] A. De, D. Bandyopadhyay, B. Gain, and A. Ekbal, "A transformer-based approach to multilingual fake news detection in low-resource languages," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–20, 2021. 16

[83] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, 2018. 93, 106

[84] DFRLab, "Suspicious twitter accounts claiming to be armenians from turkey spur nationalist rhetoric," *Medium*, 2021. xvii, 112, 113

[85] V. Doğantekin, "Ak party candidate: Votes clearly stolen in local polls," *Anadolu Agency*, 2019. 55

[86] Duke Today Staff, "The digital ban on political ads: Only the small guys got hurt," https://today.duke.edu/2021/08/digital-ban-political-ads-only-small-guys-got-hurt, 2021, accessed: 2022-11-07. 21

[87] H. S. Dutta, U. Arora, and T. Chakraborty, "Abome: A multi-platform data repository of artificially boosted online media entities," *arXiv preprint arXiv:2103.15250*, 2021. 65

[88] H. S. Dutta, A. Chetan, B. Joshi, and T. Chakraborty, "Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 242–249. 15

[89] ——, "Retweet us, we will retweet you: Spotting collusive retweeters involved in blackmarket services," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2018. 65, 71

[90] H. S. Dutta, V. R. Dutta, A. Adhikary, and T. Chakraborty, "Hawkeseye: Detecting fake retweeters using hawkes process and topic modeling," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2667–2678, 2020. 65, 80

[91] U. Dutta, R. Hanscom, J. S. Zhang, R. Han, T. Lehman, Q. Lv, and S. Mishra, "Analyzing twitter users' behavior before and after contact by the russia's internet research agency," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–24, 2021. 126

[92] P. G. Efthimion, S. Payne, and N. Proferes, "Supervised machine learning bot detection techniques to identify social twitter bots," *SMU Data Science Review*, vol. 1, no. 2, p. 5, 2018. 18

[93] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks." in *NDSS*, 2013. 13, 83, 91, 92

[94] ——, "Towards detecting compromised accounts on social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 4, pp. 447–460, 2015. 19

[95] ——, "Towards detecting compromised accounts on social networks," *IEEE Trans. Dependable Sec. Comput.*, 2017. 28

[96] M. Egele, G. Stringhini, C. Krügel, and G. Vigna, "COMPA: detecting compromised accounts on social networks," in *20th Annual Network and Distributed System Security Symposium, NDSS*, 2013. 4, 28

[97] T. Elmas, K. Hardi, R. Overdorf, and K. Aberer, "Can celebrities burst your bubble?" *arXiv preprint arXiv:2003.06857*, 2020. 9

[98] T. Elmas, R. Overdorf, and K. Aberer, "A dataset of state-censored tweets." in *ICWSM*, 2021, pp. 1009–1015. 20, 66

[99] ——, "Tactical reframing of online disinformation campaigns against the istanbul convention," *arXiv preprint arXiv:2105.13398*, 2021. 11

[100] ——, "Characterizing retweet bots: The case of black market accounts," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 171–182. 92

[101] ——, "Misleading repurposing on twitter," *arXiv preprint arXiv:2010.10600*, 2022. 16, 66

[102] T. Elmas, R. Overdorf, A. F. Özkalay, and K. Aberer, "Ephemeral astroturfing attacks: The case of fake twitter trends," *arXiv preprint arXiv:1910.07783*, 2019. 10, 14, 16

[103] ——, "Lateral astroturfing attacks on twitter trending topics," *arXiv preprint arXiv:1910.07783*, 2019. 82, 83

[104] ——, "Ephemeral astroturfing attacks: The case of fake twitter trends," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 403–422. 63, 64, 87

[105] ——, "Ephemeral astroturfing attacks: The case of fake twitter trends," in *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2021, pp. 403–422. 92

[106] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018. 12, 17

[107] R. M. Entman, "Framing: Towards clarification of a fractured paradigm," *McQuail's reader in mass communication theory*, vol. 390, p. 397, 1993. 7

[108] ——, "Framing bias: Media in the distribution of power," *Journal of communication*, vol. 57, no. 1, pp. 163–173, 2007. 7

[109] U. Etudo, V. Y. Yoon, and N. Yaraghi, "From facebook to the streets: Russian troll ads and black lives matter protests," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019. 11

[110] E. C. Ezgi Erkoyun, "Istanbul taxi drivers go to court to seek shutdown of uber," *Reuters*, 2018. 54

[111] Facebook Transparency Center, "Inauthentic behavior," https://transparency.fb.com/policies/community-standards/inauthentic-behavior/, 2022, accessed: 2022-15-07. 94

[112] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 french presidential election," *arXiv preprint arXiv:1707.00086*, 2017. 11, 29

[113] E. Ferrara, S. Cresci, and L. Luceri, "Misinformation, manipulation, and abuse on social media in the era of covid-19," *Journal of Computational Social Science*, vol. 3, no. 2, pp. 271–277, 2020. 11

[114] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, 2016. 2, 64

[115] S. Fiegerman, "Report: Twitter now charges $200,000 for promoted trends," *Mashable*, 2013. 26

[116] C. Fink, "Dangerous speech, anti-muslim violence, and facebook in myanmar," *Journal of International Affairs*, vol. 71, no. 1.5, pp. 43–52, 2018. 127

[117] L. Foster, S. Riddell, D. Mainor, and G. Roncone, "'ghostwriter'influence campaign: Unknown actors leverage website compromises and fabricated content to push narratives aligned with russian security interests," *FireEye, Mandiant*, 2020. 10

[118] C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso, "Reverse engineering socialbot infiltration strategies in twitter," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2015. 28

[119] P. d. Freitas Melo, C. C. Vieira, K. Garimella, P. O. Melo, and F. Benevenuto, "Can whatsapp counter misinformation by limiting message forwarding?" in *International conference on complex networks and their applications*. Springer, 2019, pp. 372–384. 20

[120] R. J. Gallagher, M. R. Frank, L. Mitchell, A. J. Schwartz, A. J. Reagan, C. M. Danforth, and P. S. Dodds, "Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts," *EPJ Data Science*, vol. 10, no. 1, p. 4, 2021. 69

[121] F. Gallwitz and M. Kreil, "The rise and fall of 'social bot' research," *SSRN: https://ssrn.com/abstract=3814191*, 2021. 80, 82

[122] M. Gao, H. J. Do, and W.-T. Fu, "Burst your bubble! an intelligent system for improving awareness of diverse social opinions," in *23rd International Conference on Intelligent User Interfaces*, 2018, pp. 371–383. 9, 22

[123] K. Garimella, G. De Francisc iMorales, A. Gionis, and M. Mathioudakis, "Mary, mary, quite contrary: Exposing twitter users to contrarian news," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 201–205. 9

[124] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Reducing controversy by connecting opposing views," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, pp. 81–90. 9

[125] ——, "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 913–922. 9

[126] M. Giatsoglou, D. Chatzakou, N. Shah, C. Faloutsos, and A. Vakali, "Retweeting activity on twitter: Signs of deception," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2015. 65

[127] Z. Gilani, R. Farahbakhsh, G. Tyson, L. Wang, and J. Crowcroft, "Of bots and humans (on twitter)," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 2017, pp. 349–354. 67, 125

[128] O. Goga, G. Venkatadri, and K. P. Gummadi, "The doppelgänger bot attack: Exploring identity impersonation in online social networks," in *Proceedings of the 2015 Internet Measurement Conference*, 2015. 92

[129] J. Golbeck, "Benford's law can detect malicious social bots," *First Monday*, 2019. xvi, 15, 65, 68

[130] F. J. Gómez-Fernandez and F. Terroso-Sáenz, "Towards a web tool for the analysis of twitter profiling information," in *Intelligent Environments 2020*. IOS Press, 2020, pp. 391–399. 22

[131] C. Grier, K. Thomas, V. Paxson, and C. M. Zhang, "@spam: the underground on 140 characters or less," in *Proceedings of the 17th ACM Conference on Computer and Communications Security, CCS*, 2010. 28

[132] S. Grossman, F. A. Akış, A. Alemdaroğlu, J. Goldstein, and K. Jonsson, "Political retweet rings and compromised accounts: A twitter influence operation linked to the youth wing of turkey's ruling party," Stanford University, Tech. Rep., 2020, stanford Internet Observatory. 10, 18, 21, 46, 48

[133] S. Grossman, K. H., E. Ross, and D. Thiel, "Royal sockpuppets and handle switching: How a saudi arabia-linked twitter network stoked rumors of a coup in qatar," *Stanford Internet Observatory*, 2020. 95

[134] A. Guille and C. Favre, "Event detection, tracking, and visualization in twitter: a mention-anomaly-based approach," *Social Network Analysis and Mining*, vol. 5, no. 1, pp. 1–18, 2015. 29

[135] S. Gupta, P. Kumaraguru, and T. Chakraborty, "Malreg: Detecting and analyzing malicious retweeter groups," in *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2019. 65

[136] J. Guzman and B. Poblete, "On-line relevant anomaly detection in the twitter stream: an efficient bursty keyword detection model," in *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, 2013, pp. 31–39. 29

[137] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *ACM sigmod record*, vol. 29, no. 2, pp. 1–12, 2000. 29, 59

[138] D. Hangartner, G. Gennaro, S. Alasiri, N. Bahrich, A. Bornhoft, J. Boucher, B. B. Demirci, L. Derksen, A. Hall, M. Jochum *et al.*, "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment," *Proceedings of the National Academy of Sciences*, vol. 118, no. 50, p. e2116310118, 2021. 17

[139] K. Hao, "Google is finally admitting it has a filter-bubble problem," *Quarz*, 2018. 9

[140] P. Harrison and S. Akyol, "Founder accused of defrauding gamers," *BBC*, 2018. 54

[141] R. Help, "What constitutes vote cheating or vote manipulation? accessed on 2022-15-07," https://www.reddithelp.com/hc/en-us/articles/360043066412. 10

[142] Y. Help, "Verification badges on channels. accessed on 2022-15-07," https://support.google.com/youtube/answer/3046484?hl=en. 110

[143] W. Herzallah, H. Faris, and O. Adwan, "Feature engineering for detecting spammers on twitter: Modelling and analysis," *Journal of Information Science*, 2018. 64, 92, 109

[144] M. Hindman and V. Barash, "Disinformation,'fake news' and influence campaigns on twitter," 2018. 11

[145] Holger Schwenk, "Zero-shot transfer across 93 languages: Open-sourcing enhanced laser library," https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/, 2019, accessed: 2022-11-07. 21

[146] P. N. Howard and B. Kollanyi, "Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum," *Available at SSRN 2798311*, 2016. 74, 80

[147] K. Hristakieva, S. Cresci, G. Da San Martino, M. Conti, and P. Nakov, "The spread of propaganda by coordinated communities on social media," in *14th ACM Web Science Conference 2022*, 2022, pp. 191–201. 18

[148] IcyBug8, "#suriyelilerdefolsun is trending in turkey," https://tinyurl.com/redditsyrians, 2018, accessed on 2020-07-27. 55

[149] M. Ienca and E. Vayena, "Ethical requirements for responsible research with hacked data," *Nature Machine Intelligence*, vol. 3, no. 9, pp. 744–748, 2021. 87

[150] L. P. Issac, "Sql vs nosql database differences explained with few example db," https://www.thegeekstuff.com/2014/01/sql-vs-nosql-db/, 2014, the Geek Stuff. 112

[151] A. Iyer and S. Vosoughi, "Style change detection using bert." in *CLEF (Working Notes)*, 2020. 93, 106

[152] P. Jain and P. Kumaraguru, "On the dynamics of username changing behavior on twitter," in *Proceedings of the 3rd IKDD Conference on Data Science, 2016*, 2016. 92

[153] J. Jia, B. Wang, and N. Z. Gong, "Random walk based fake account detection in online social networks," in *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2017. 28

[154] M. O. Jones, "The gulf information war— propaganda, fake news, and fake trends: The weaponization of twitter bots in the gulf crisis," *International journal of communication*, vol. 13, p. 27, 2019. 29

[155] Jules Wang, "Massive twitter breach made possible by social engineering," https://www.androidpolice.com/2020/07/20/verified-twitter-accounts-hijacked-to-promote-bitcoin-scam/, 2020, accessed: 2022-11-07. 13

[156] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, "exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert)," *Applied Sciences*, vol. 9, no. 19, p. 4062, 2019. 16

[157] S. Kaçar, "Adnan oktar sahte twitter hesaplarıyla gündeme getiriliyor!" *Haber Ne Diyor*, 2020. 48

[158] F. Kafka, "Die verwandlung," *Die Weißen Blätter. Eine Montasschrift*, 1915. 89

[159] Kaparoz, "Akp'den savunma: Hesaplar bot değil arkadaşlarımızın türkçesi bu kadar," *Kaparoz*, 2018. 32

[160] Karar, "Tdp'nin twitter hesabının brezilyalı aktrise ait olduğu iddia edildi," *Karar*, 2020. 112

[161] H. Karimi, C. VanDam, L. Ye, and J. Tang, "End-to-end compromised account detection," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 314–321. 4, 19, 83, 91, 92

[162] R. Kaur, S. Singh, and H. Kumar, "Tb-coauth: Text based continuous authentication for detecting compromised accounts in social networks," *Applied Soft Computing*, vol. 97, p. 106770, 2020. 4, 19, 83, 91, 92

[163] D. Kekulluoglu, K. Vaniea, and W. Magdy, "Understanding privacy switching behaviour on twitter," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–14. 23

[164] F. B. Keller, D. Schoch, S. Stier, and J. Yang, "Political astroturfing on twitter: How to coordinate a disinformation campaign," *Political Communication*, 2019. 29

[165] ——, "Political astroturfing on twitter: How to coordinate a disinformation campaign," *Political Communication*, vol. 37, no. 2, pp. 256–280, 2020. 15

[166] K. A. Kırkıç, A. P. Kırkıç, and c. Berberoğlu, "The educational needs of refugees in a multicultural world: An innovative solution to the problem," *Uluslararası Eğitim Programları ve Öğretim Çalışmaları Dergisi*, 2018. 54, 55

[167] Q. Kong, R. Ram, and M.-A. Rizoiu, "A toolkit for analyzing and visualizing online users via reshare cascade modeling," *arXiv e-prints*, pp. arXiv–2006, 2020. 22

[168] A. Korbani and J. LaBrie, "Toxic tiktok trends," *Journal of Student Research*, vol. 10, no. 2, 2021. 23

[169] S. Kotsiantis, D. Kanellopoulos, P. Pintelas *et al.*, "Handling imbalanced datasets: A review," *GESTS international transactions on computer science and engineering*, 2006. 107

[170] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Information Sciences*, vol. 467, pp. 312–322, 2018. 18

[171] M. Kutlu, K. Darwish, C. Bayrak, A. Rashed, and T. Elsayed, "Embedding-based qualitative analysis of polarization in turkey," *arXiv preprint arXiv:1909.10213*, 2019. 32

[172] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *2013 IEEE 13th international conference on data mining*. IEEE, 2013, pp. 1103–1108. 17

[173] P. Lahoti, K. Garimella, and A. Gionis, "Joint non-negative matrix factorization for learning ideological leaning on twitter," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 351–359. 17

[174] R. Layton, P. Watters, and R. Dazeley, "Authorship attribution for twitter in 140 characters or less," in *2010 Second Cybercrime and Trustworthy Computing Workshop*. IEEE, 2010. 93

[175] G. Lebanon and M. El-Geish, "Thoughts on system design for big data," in *Computing with Data*. Springer, 2018, pp. 495–541. 12

[176] K. Lee, B. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on twitter," in *Proceedings of the international AAAI conference on web and social media*, vol. 5, no. 1, 2011, pp. 185–192. 18

[177] A. L. M. Lemos, E. C. Bitencourt, and J. G. B. dos Santos, "Fake news as fake politics: the digital materialities of youtube misinformation videos about brazilian oil spill catastrophe," *Media, Culture & Society*, p. 0163443720977301, 2020. 23

[178] E. Lex, M. Wagner, and D. Kowald, "Mitigating confirmation bias on twitter by recommending opposing views," *arXiv preprint arXiv:1809.03901*, 2018. 9

[179] H. Li, B. Hecht, and S. Chancellor, "All that's happening behind the scenes: Putting the spotlight on volunteer moderator labor in reddit," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 584–595. 20

[180] C. Ling, K. P. Gummadi, and S. Zannettou, "" learn the facts about covid-19": Analyzing the use of warning labels on tiktok videos," *arXiv preprint arXiv:2201.07726*, 2022. 20

[181] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422. 29, 60

[182] S. Liu, B. Hooi, and C. Faloutsos, "Holoscope: Topology-and-spike aware fraud detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017. 65

[183] C. Llewellyn, L. Cram, R. L. Hill, and A. Favero, "For whom the bell trolls: Shifting troll behaviour in the twitter brexit debate," *JCMS: Journal of Common Market Studies*, vol. 57, no. 5, pp. 1148–1164, 2019. 95

[184] H. Lu, J. Caverlee, and W. Niu, "Biaswatch: A lightweight system for discovering and tracking topic-sensitive opinion bias in social media," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 2015, pp. 213–222. 17

[185] L. Luceri, A. Deb, A. Badawy, and E. Ferrara, "Red bots do it better: Comparative analysis of social bot partisan behavior," in *Companion proceedings of the 2019 World Wide Web conference*, 2019, pp. 1007–1012. 17

[186] L. Luceri, S. Giordano, and E. Ferrara, "Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 417–427. 18

[187] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1751–1754. 17

[188] maharishi, "#suriyelilerdefolsun. accessed on 2020-07-27," 2018, ekşi Sözlük. 55

[189] E. Mariconti, J. Onaolapo, S. S. Ahmad, N. Nikiforou, M. Egele, N. Nikiforakis, and G. Stringhini, "What's in a name? understanding profile name reuse on twitter," in *Proceedings of the 26th International Conference on World Wide Web*, 2017. 92

[190] S. Martin, W. M. Brown, R. Klavans, and K. W. Boyack, "Openord: an open-source toolbox for large graph layout," in *Visualization and Data Analysis 2011*. International Society for Optics and Photonics, 2011. xv, 49

[191] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "Rtbust: Exploiting temporal patterns for botnet detection on twitter," in *Proceedings of the 10th ACM conference on web science*, 2019, pp. 183–192. 65, 67, 80

[192] J. McCoy, T. Rahman, and M. Somer, "Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities," *American Behavioral Scientist*, 2018. 32

[193] K. Meleshevich and B. Schafer, "Online information laundering: The role of social media," *Alliance for Securing Democracy, January*, vol. 9, 2018. 8

[194] J. Mendelsohn, C. Budak, and D. Jurgens, "Modeling framing in immigration discourse on social media," *arXiv preprint arXiv:2104.06443*, 2021. 11, 17

[195] Meta, "Updating our data access tools. accessed on 2022-07-11," https://about.fb.com/news/2020/03/data-access-tools, 2020, meta Newsroom. 20

[196] 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*, 2012, pp. 646–655. 29

[197] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Information Sciences*, vol. 260, pp. 64–73, 2014. 18

[198] M. Minaei, S. C. Mouli, M. Mondal, B. Ribeiro, and A. Kate, "Deceptive deletions for protecting withdrawn posts on social media platform," in *NDSS*, 2021. 22

[199] V. Miz, B. Ricaud, K. Benzi, and P. Vandergheynst, "Anomaly detection in the dynamics of web and social networks using associative memory," in *The World Wide Web Conference*, 2019, pp. 1290–1299. 29

[200] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *Proceedings of the 28th ACM conference on hypertext and social media*, 2017, pp. 85–94. 12

[201] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing? understanding microblog credibility perceptions," in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 441–450. 14, 109

[202] D. C. Moyer, S. L. Carson, T. K. Dye, R. T. Carson, and D. Goldbaum, "Determining the influence of reddit posts on wikipedia pageviews," in *Ninth international AAAI conference on web and social media*, 2015. 9

[203] Nathaniel Gleicher, "Coordinated inauthentic behavior explained," https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/, 2018, accessed: 2022-11-07. 15, 18

[204] ——, "Removing coordinated inauthentic behavior," https://about.fb.com/news/2020/10/removing-coordinated-inauthentic-behavior-september-report/, 2020, accessed: 2022-11-07. 13

[205] K. Neha, S. Srikanth, S. Singhal, S. Singh, A. B. Buduru, and P. Kumaraguru, "Is change the only constant? profile change perspective on# loksabhaelections2019," *arXiv preprint arXiv:1909.10012*, 2019. 92

[206] C. Nyst and N. Monaco, "State-sponsored trolling: how governments are deploying disinformation as part of broader digital harassment campaigns," *Institute for the Future*, 2018. 28

[207] A. Olteanu, C. Castillo, J. Boy, and K. Varshney, "The effect of extremist violence on hateful speech online," in *Proceedings of the international AAAI conference on web and social media*, vol. 12, no. 1, 2018. 12

[208] R. Overdorf and R. Greenstadt, "Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution." *Proc. Priv. Enhancing Technol.*, 2016. 93

[209] O. Ozduzen and U. Korkut, "Post-'refugee crisis' social media: the unbearable lightness of sharing racist posts," *Discover Society*, 2020. 54, 55

[210] D. Pacheco, A. Flammini, and F. Menczer, "Unveiling coordinated groups behind white helmets disinformation," in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 611–616. 11

[211] D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, and F. Menczer, "Uncovering coordinated networks on social media: Methods and case studies." *ICWSM*, vol. 21, pp. 455–466, 2021. 18

[212] M. Parenti, "Methods of media manipulation," *The humanist*, vol. 57, no. 4, p. 5, 1997. 8

[213] E. Pariser, *The filter bubble: What the Internet is hiding from you.* Penguin UK, 2011. 9

[214] I. V. Pasquetto, B. Swire-Thompson, M. A. Amazeen, F. Benevenuto, N. M. Brashier, R. M. Bond, L. C. Bozarth, C. Budak, U. K. Ecker, L. K. Fazio *et al.*, "Tackling misinformation: What researchers could do with social media data," *The Harvard Kennedy School Misinformation Review*, 2020. 125

[215] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 60

[216] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, 2011. 107

[217] Pedro Kanahuati, "Keeping passwords secure," https://about.fb.com/news/2019/03/keeping-passwords-secure, 2019, accessed: 2022-11-07. 12

[218] S. Pekkendir, "Discursive identifaction of syrian refugees in turkey as a way of governmentality," Ph.D. dissertation, University of London, 2018. 54, 55

[219] R. M. Perloff, *The dynamics of political communication: Media and politics in a digital age*. Routledge, 2021. 7

[220] N. Perlroth, "A conspiracy made in america may have been spread by russia," *New York Times*, 2020. 26

[221] F. Policies, "Pages, groups, and events. accessed on 2022-15-07," https://www.facebook.com/policies_center/pages_groups_events. 91

[222] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, "Where the truth lies: Explaining the credibility of emerging claims on the web and social media," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 1003–1012. 16

[223] T. Porter, "How diehard trump fans transformed their twitter accounts into bots which spread conspiracies in a vast russia-style disinformation network," *Business Insider*, 2020. 28, 56

[224] M. Potter, "Bad actors never sleep: content manipulation on reddit," *Continuum*, pp. 1–13, 2021. 23

[225] R. Ram, Q. Kong, and M.-A. Rizoiu, "Birdspotter: A tool for analyzing and labeling twitter users," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 918–921. 22

[226] J. Ratkiewicz, M. D. Conover, M. R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, "Detecting and tracking political abuse in social media," in *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 2011. 29

[227] A. Rauchfleisch and J. Kaiser, "The false positive problem of automatic bot detection in social science research," *Berkman Klein Center Research Publication*, 2020. 85

[228] S. Raza and C. Ding, "Fake news detection based on news content and social contexts: a transformer-based approach," *International Journal of Data Science and Analytics*, vol. 13, no. 4, pp. 335–362, 2022. 16

[229] Reddit, "Quarantined subreddits. accessed on 2022-11-07," https://reddit.zendesk.com/hc/en-us/articles/360043069012-Quarantined-Subreddits, reddit Help. 21

[230] A. Richterich, "'karma, precious karma!'karmawhoring on reddit and the front page's econometrisation," *Journal of Peer Production*, vol. 4, no. 1, pp. 1–12, 2014. 15

[231] L. Risso, "Harvesting your soul? cambridge analytica and brexit," *Brexit Means Brexit*, vol. 2018, pp. 75–90, 2018. 13

[232] S. T. Roberts, "Commercial content moderation: Digital laborers' dirty work," 2016. 20

[233] K. Roose, "Qanon followers are hijacking the #savethechildren movement," *New York Times*, 2020. 26

[234] B. Ross, L. Pilz, B. Cabrera, F. Brachten, G. Neubaum, and S. Stieglitz, "Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks," *European Journal of Information Systems*, vol. 28, no. 4, pp. 394–412, 2019. 125

[235] A. Rother, U. Niemann, T. Hielscher, H. Völzke, T. Ittermann, and M. Spiliopoulou, "Assessing the difficulty of annotating medical data in crowdworking with help of experiments," *PloS one*, 2021. 98

[236] N. Ruwandika and A. Weerasinghe, "Identification of hate speech in social media," in *2018 18th international conference on advances in ICT for emerging regions (ICTer)*. IEEE, 2018, pp. 273–278. 17

[237] M. H. Saeed, S. Ali, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, "Trollmagnifier: Detecting state-sponsored troll accounts on reddit," *arXiv preprint arXiv:2112.00443*, 2021. 18

[238] 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*, 2010, pp. 851–860. 29

[239] A. Salihefendic, "How reddit ranking algorithms work," *medium*, 2015. 56

[240] M. Sayyadiharikandeh, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "Detection of novel social bots by ensembles of specialized classifiers," in *Proceedings of the 29th ACM international conference on information & knowledge management*, 2020, pp. 2725–2732. 18

[241] C. O. Schneble, B. S. Elger, and D. Shaw, "The cambridge analytica affair and internet-mediated research," *EMBO reports*, vol. 19, no. 8, p. e46579, 2018. 1

[242] C. Semercioğlu, "Dünya gündeminde nasıl 1 numara oldum," *Hürriyet*, 2014. 32, 46, 48

[243] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Hoaxy: A platform for tracking online misinformation," in *Proceedings of the 25th international conference companion on world wide web*, 2016, pp. 745–750. 22

[244] R. Shay, I. Ion, R. W. Reeder, and S. Consolvo, "" my religious aunt asked why i was trying to sell her viagra" experiences with account hijacking," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014. 92

[245] J. Shima, M. Yoshida, and K. Umemura, "When do users change their profile information on twitter?" in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017. 92

[246] K. Shu, D. Mahudeswaran, and H. Liu, "Fakenewstracker: a tool for fake news collection, detection, and visualization," *Computational and Mathematical Organization Theory*, vol. 25, no. 1, pp. 60–71, 2019. 22

[247] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, 08 2017. 16

[248] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 312–320. 16

[249] M. Singhal, C. Ling, N. Kumarswamy, G. Stringhini, and S. Nilizadeh, "Sok: Content moderation in social media, from guidelines to enforcement, and research to practice," *arXiv e-prints*, pp. arXiv–2206, 2022. 20

[250] P. Smeros, C. Castillo, and K. Aberer, "Scilens: Evaluating the quality of scientific news articles using social media and scientific literature indicators," in *The World Wide Web Conference*, 2019. 17, 98

[251] S. N. Soroka, "The gatekeeping function: Distributions of information in media and the real world," *The Journal of Politics*, vol. 74, no. 2, pp. 514–528, 2012. 7

[252] E. K. Sozeri, "How pro-government trolls are using a sexy twitter bot to sway turkey's election," *The Daily Dot*, 2015. 95

[253] M. Stella, M. Cristoforetti, and M. De Domenico, "Influence of augmented humans in online interactions during voting events," *PloS one*, 2019. 28

[254] Steve Stecklow, "Why facebook is losing the war on hate speech in myanmar," https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/, 2018, accessed: 2022-11-07. 127

[255] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th annual computer security applications conference*, 2010. 81

[256] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao, "Follow the green: growth and dynamics in twitter follower markets," in *Proceedings of the 2013 conference on Internet measurement conference*, 2013, pp. 163–176. 14

[257] E. Strøm, "Multi-label style change detection by solving a binary classification problem," in *CLEF*, 2021. 93

[258] I. J. Strudwicke and W. J. Grant, "# junkscience: Investigating pseudoscience disinformation in the russian internet research agency tweets," *Public Understanding of Science*, vol. 29, no. 5, pp. 459–472, 2020. 1

[259] R. Takacs and I. McCulloh, "Dormant bots in social media: Twitter and the 2018 us senate election," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2019, pp. 796–800. 92

[260] T. Y. Team, "An update to dislikes on youtube. accessed on 2022-15-07," https://blog.youtube/news-and-events/update-to-youtube/. 10

[261] R. Tekumalla, J. R. Asl, and J. M. Banda, "Mining archive. org's twitter stream grab for pharmacovigilance research gold," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 909–917. 66

[262] D. Thom, H. Bosch, S. Koch, M. Wörner, and T. Ertl, "Spatiotemporal anomaly detection through visual analysis of geolocated twitter messages," in *2012 IEEE Pacific Visualization Symposium*. IEEE, 2012, pp. 41–48. 29

[263] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar *et al.*, "Sok: Hate, harassment, and the changing landscape of online abuse," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 247–267. 11

[264] K. Thomas, C. Grier, and V. Paxson, "Adapting social spam infrastructure for political censorship," in *5th USENIX Workshop on Large-Scale Exploits and Emergent Threats, LEET*, 2012. 28

[265] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson, "Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse," in *22nd {USENIX} Security Symposium ({USENIX} Security 13)*, 2013. 92, 95

[266] J. Tidy, "Youtube accused of not tackling musk bitcoin scam streams," *BBC News*, 2022. 110

[267] Tomiwa Ilori, "Facebook's content moderation errors are costing africa too much," https://slate.com/technology/2020/10/facebook-instagram-endsars-protests-nigeria.html, 2020, accessed: 2022-11-07. 127

[268] C. Torres-Lugo, M. Pote, A. C. Nwala, and F. Menczer, "Manipulating twitter through deletions," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 1029–1039. 16

[269] C. Torres-Lugo, K.-C. Yang, and F. Menczer, "The manufacture of partisan echo chambers by follow train abuse on twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 1017–1028. 16

[270] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," in *Eighth international AAAI conference on weblogs and social media*, 2014. 23

[271] ——, *Twitter and tear gas*. Yale University Press, 2017. 1, 8, 14, 19

[272] Twitter, "About government and state-affiliated media account labels on twitter. accessed on 2022-11-07," https://help.twitter.com/en/rules-and-policies/state-affiliated, twitter Help Center. 21

[273] ——, "About the notifications timeline. accessed on 2022-11-07," https://help.twitter.com/en/managing-your-account/understanding-the-notifications-timeline, twitter Help Center. 21

[274] ——, "Political content," https://business.twitter.com/en/help/ads-policies/ads-content-policies/political-content.html, accessed: 2022-11-07. 21

[275] Twitter Inc., "About automated account labels. accessed on 2022-07-11," https://help.twitter.com/en/using-twitter/automated-account-labels, twitter Help Center. 20

[276] ——, "Information operations. accessed on 2022-11-07," https://transparency.twitter.com/en/reports/information-operations.html. 21, 97

[277] ——, "Twitter trends faq. accessed on 2022-15-07," https://help.twitter.com/en/using-twitter/twitter-trending-faqs, twitter Help Center. 39

[278] ——, "Disclosing networks of state-linked information operations we've removed," 2020. 46

[279] ——, "Platform manipulation and spam policy. accessed on 2022-07-11," 2020, twitter Help Center. 15

[280] ——, "Platform manipulation and spam policy. accessed on 2022-07-11," https://help.twitter.com/en/rules-and-policies/coordinated-harmful-activity, 2020, twitter Help Center. 18

[281] Twitter Safety, "An update following the riots in washington, dc," https://blog.twitter.com/en/topics/company/2021/protecting–the-conversation-following-the-riots-in-washington–, 2021, accessed: 2022-11-07. 20, 126

[282] T. Uren, E. Thomas, and J. Wallis, "Tweeting through the great firewall: Preliminary analysis of prc-linked information operations on the hong kong protest," *Australia Strategic Policy Institute: International Cyber Policy Center, Barton*, 2019. 95, 102

[283] J. Uyheng and K. M. Carley, "Bots and online hate during the covid-19 pandemic: case studies in the united states and the philippines," *Journal of computational social science*, vol. 3, no. 2, pp. 445–468, 2020. 11

[284] I. Van Ooijen and H. U. Vrabec, "Does the gdpr enhance consumers' control over personal data? an analysis from a behavioural perspective," *Journal of consumer policy*, vol. 42, no. 1, pp. 91–107, 2019. 20

[285] C. VanDam and P. Tan, "Detecting hashtag hijacking from twitter," in *Proceedings of the 8th ACM Conference on Web Science, WebSci*, 2016. 28

[286] C. VanDam, J. Tang, and P.-N. Tan, "Understanding compromised accounts on twitter," in *Proceedings of the International Conference on Web Intelligence*, 2017. 92

[287] J. Vanian, "Facebook is testing this new feature to fight 'filter bubbles'," *Fortune*, 2017. 9

[288] L. Vargas, P. Emami, and P. Traynor, "On the detection of disinformation campaign activity with network analysis," in *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2020. 65

[289] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017. 67, 78

[290] O. Varol, E. Ferrara, F. Menczer, and A. Flammini, "Early detection of promoted campaigns on social media," *EPJ Data Science*, 2017. 29

[291] B. Vidgen and T. Yasseri, "Detecting weak and strong islamophobic hate speech on social media," *Journal of Information Technology & Politics*, vol. 17, no. 1, pp. 66–78, 2020. 17

[292] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. 60

[293] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *science*, vol. 359, no. 6380, pp. 1146–1151, 2018. 9

[294] V. V. Vydiswaran, C. Zhai, D. Roth, and P. Pirolli, "Overcoming bias to learn about controversial topics," *Journal of the Association for Information Science and Technology*, vol. 66, no. 8, pp. 1655–1672, 2015. 9

[295] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. J. Metzger, H. Zheng, and B. Y. Zhao, "Social turing tests: Crowdsourcing sybil detection," in *20th Annual Network and Distributed System Security Symposium, NDSS*, 2013. 28

[296] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor, "" i regretted the minute i pressed share" a qualitative study of regrets on facebook," in *Proceedings of the seventh symposium on usable privacy and security*, 2011. 122

[297] C. Warzel, "New charts show what the russian troll @ten_gop account was tweeting this summer," *BuzzFeed*, 2017. 95

[298] J. Weerasinghe, B. Flanigan, A. Stein, D. McCoy, and R. Greenstadt, "The pod people: Understanding manipulation of social media popularity via reciprocity abuse," in *Proceedings of The Web Conference 2020*, 2020, pp. 1874–1884. 15

[299] F. Wei and U. T. Nguyen, "Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings," in *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. IEEE, 2019, pp. 101–109. 18

[300] J. Weng and B.-S. Lee, "Event detection in twitter," in *Proceedings of the international aaai conference on web and social media*, vol. 5, no. 1, 2011, pp. 401–408. 29

[301] R. Wesslen, S. Nandu, O. Eltayeby, T. Gallicano, S. Levens, M. Jiang, and S. Shaikh, "Bumper stickers on the twitter highway: Analyzing the speed and substance of profile changes," in *Twelfth International AAAI Conference on Web and Social Media*, 2018. 92, 97

[302] T. Wilson and K. Starbird, "Cross-platform information operations: Mobilizing narratives & building resilience through both'big'&'alt'tech," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–32, 2021. 23

[303] S. C. Woolley, "Automating power: Social bot interference in global politics," *First Monday*, 2016. 28

[304] S. C. Woolley and P. N. Howard, *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press, 2018. 10

[305] x0rz, "Uncovering foreign trolls (trying) to influence french elections on twitter," *Medium*, 2018. 94

[306] W. Xiong and R. Lagerström, "Threat modeling–a systematic literature review," *Computers & security*, vol. 84, pp. 53–69, 2019. 10

[307] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," *IEEE Transactions on Information Forensics and Security*, 2013. 78

[308] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 71–80. 15

[309] C. Yang, R. C. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," *IEEE Trans. Information Forensics and Security*, 2013. 28

[310] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *Proceedings of the ACM SIGKDD workshop on mining data semantics*, 2012, pp. 1–7. 17

[311] K.-C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, "Arming the public with artificial intelligence to counter social bots," *Human Behavior and Emerging Technologies*, 2019. 28

[312] K.-C. Yang, O. Varol, P.-M. Hui, and F. Menczer, "Scalable and generalizable social bot detection through data selection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. 67, 78

[313] S. Yardi, D. Romero, G. Schoenebeck *et al.*, "Detecting spam in a twitter network," *First Monday*, 2010. 2, 28, 63, 64

[314] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman, "Sybilguard: defending against sybil attacks via social networks," in *Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, 2006. 28

[315] M. J. Zaki, "Scalable algorithms for association mining," *IEEE transactions on knowledge and data engineering*, vol. 12, no. 3, pp. 372–390, 2000. 29

[316] E. Zangerle, M. Mayerl, M. Potthast, and B. Stein, "Overview of the style change detection task at pan," 2021. 93

[317] E. Zangerle, M. Mayerl, G. Specht, M. Potthast, and B. Stein, "Overview of the style change detection task at pan 2020." in *CLEF (Working Notes)*, 2020. 93

[318] E. Zangerle and G. Specht, “” sorry, i was hacked” a classification of compromised twitter accounts,” in *Proceedings of the 29th annual ACM symposium on applied computing*, 2014. 69, 92

[319] ——, “”sorry, I was hacked”: a classification of compromised twitter accounts,” in *Symposium on Applied Computing, SAC*, 2014. 28

[320] S. Zannettou, “” i won the election!”: An empirical analysis of soft moderation interventions on twitter.” in *ICWSM*, 2021, pp. 865–876. 20

[321] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn, “Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web,” in *Companion proceedings of the 2019 world wide web conference*, 2019, pp. 218–226. 15, 16, 18

[322] ——, “Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web,” in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019. 92

[323] S. Zannettou, M. ElSherief, E. Belding, S. Nilizadeh, and G. Stringhini, “Measuring and characterizing hate speech on news websites,” in *12th ACM Conference on Web Science*, 2020, pp. 125–134. 17

[324] J. Zarocostas, “How to fight an infodemic,” *The lancet*, vol. 395, no. 10225, p. 676, 2020. 11

[325] J. Zhang, R. Zhang, Y. Zhang, and G. Yan, “The rise of social botnets: Attacks and countermeasures,” *IEEE Trans. Dependable Sec. Comput.*, 2018. 28

[326] Y. Zhang, X. Yuan, J. Li, J. Lou, L. Chen, and N.-F. Tzeng, “Reverse attack: Black-box attacks on collaborative recommendation,” in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 51–68. 22

[327] Y. Zhang, X. Ruan, H. Wang, H. Wang, and S. He, “Twitter trends manipulation: A first look inside the security of twitter trending,” *IEEE Trans. Information Forensics and Security*, 2017. 29

[328] Z. Zhang, Z. Han, L. Kong, X. Miao, Z. Peng, J. Zeng, H. Cao, J. Zhang, Z. Xiao, and X. Peng, “Style change detection based on writing style similarity,” 2021. 93, 106

[329] S. Zhao, L. Zhong, J. Wickramasuriya, and V. Vasudevan, “Human as real-time sensors of social and physical events: A case study of twitter and sports games,” *arXiv preprint arXiv:1106.4300*, 2011. 29

[330] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD explorations newsletter*, vol. 21, no. 2, pp. 48–60, 2019. 17

[331] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, and P. Nakov, "An ensemble-rich multi-aspect approach for robust style change detection," *CLEF 2018 Working Nots of CLEF*, 2018. 93

[332] A. Zubiaga, D. Spina, E. Amigó, and J. Gonzalo, "Towards real-time summarization of scheduled events from twitter streams," in *Proceedings of the 23rd ACM conference on Hypertext and social media*, 2012, pp. 319–320. 29

[333] C. Zuo, Y. Zhao, and R. Banerjee, "Style change detection with feed-forward neural networks." in *CLEF (Working Notes)*, 2019. 93