EPFL

Thèse n° 8283

# On the use of applied machine learning and digital infrastructure to leverage social media data in health and epidemiology

Présentée le 23 février 2021

Faculté des sciences de la vie
Unité du Prof. Salathé (SV/IC)
Programme doctoral en biologie computationnelle et quantitative

pour l'obtention du grade de Docteur ès Sciences

par

**Martin Mathias MÜLLER**

Acceptée sur proposition du jury

Prof. J. Fellay, président du jury
Prof. M. Salathé, directeur de thèse
Prof. H. Larson, rapporteuse
Prof. M. Santillana, rapporteur
Prof. R. West, rapporteur

École polytechnique fédérale de Lausanne

2021

# Acknowledgements

First and foremost, I would like to thank Marcel Salathé for giving me the opportunity to conduct this research. I am grateful for the level of trust and freedom he gave me in this endeavour as well as for providing me with key advice which allowed me to see the bigger picture in this work. His optimism, clarity of mind, and humor allowed me to overcome difficult phases of my PhD.

I would like to acknowledge my collaborators, in particular Per Egil Kummervold, with whom I spent countless hours debugging BERT models and dived into the depths of natural language processing.

My gratitude goes to all present and past members of the Digital Epidemiology Lab, in particular Yannis Jacquet and Sean Carroll who introduced me to the craft of building web applications in the beginning of my PhD. I thank Djilani Kebaili for countless discussions about programming, crypto, and politics over late-night chess games or Taco Rico dinners. I also would like to thank Gianrocco Lazzari for the many discussions about science, academia, and PhD life in general. Special thanks go to Marina Secat for her adamant support during these four years. Thanks to Chloé Allémann, Talia Salzmann, S.P. Mohanty, and Sylvain Bernard for having been great office mates.

Special thanks go to Harry Anderson for the many laughs, cold dips in the lake, cooking enormous amounts of food, complaining about cars, and just generally sharing some of the best times I've had during my PhD. Further, I would like to thank my life-long friends the "junks", Christian Somody, and Philipp Indlekofer.

And finally, as others will be able to attest, I'm incredibly lucky to have Burcu Tepekule by my side, who has helped me in countless ways over these years both as a partner and friend but also scientifically. It is clear that I wouldn't have finished this work without her support. My gratitude also goes to my dad Toni Müller and my sister Susanne Müller for their continuous and unconditional support.

# Preface

All code, data, as well as all machine learning models which have been developed in the context of this thesis, are openly available on GitHub, GitLab, or Zenodo. The link to the respective repository is provided at the end of each chapter.

# Abstract

The quantification of population-level health behaviors is crucial for guiding public health policy. However, traditional methods for measuring such health behaviors have several shortcomings. In recent years social media data has been successfully used to measure health behaviors and may be used as a low-cost and real-time addition to traditional data sources. Methods from the field of natural language processing are increasingly used to automatically process, filter and categorize the rapidly growing amount of publicly available social media data. However, a number of methodological challenges limit the rate at which we can generate insight from such data.

In this work I will argue that long-term investment into digital infrastructure and open source tooling is required in order to overcome these challenges. In chapter 2 we introduce the Crowdbreaks platform which is the basis of this thesis. Crowdbreaks is an open source framework for real-time data collection, continuous crowdsourced annotation, and continuous re-training of machine learning classifiers. In contrast to traditional research workflows, projects on Crowdbreaks run over an extended period of time, allowing for the observation of health trends over multiple years while keeping algorithms up-to-date. In chapter 3 we quantify the occurrence of concept drift in vaccine-related Twitter data, which further validates the need for the Crowdbreaks platform. In chapter 4 we use the Crowdbreaks platform to trace sentiment towards the novel gene-editing technology CRISPR/Cas9 back to its first application in 2013 and investigate how public opinion may have been affected in context of recent scandals surrounding the technology. In chapter 5 we turn our attention to the COVID-19 pandemic and analyze who was speaking and who was heard in the early months of the pandemic. Chapter 6 builds on this work and explores the dynamics of Twitter communities during the COVID-19 pandemic. Lastly, in chapter 7 we introduce COVID-Twitter-BERT, a domain-specific language model which has been used in various downstream natural language processing applications on COVID-19-related Twitter data.

**Keywords**: Social media, natural language processing, digital epidemiology, Twitter, health

# Zusammenfassung[I]

Die Quantifizierung des Gesundheitsverhaltens auf Bevölkerungsebene ist für die Steuerung der öffentlichen Gesundheitspolitik von entscheidender Bedeutung. Herkömmliche Methoden zur Messung solchen Gesundheitsverhaltens haben jedoch mehrere Mängel. In den letzten Jahren wurden Daten aus sozialen Medien erfolgreich zur Messung des Gesundheitsverhaltens eingesetzt und können als kostengünstige und zeitnahe Ergänzung zu traditionellen Datenquellen verwendet werden. Methoden aus dem Bereich der natürlichen Sprachverarbeitung werden zunehmend eingesetzt, um die schnell wachsende Menge an öffentlich verfügbaren Social-Media-Daten automatisch zu verarbeiten, zu filtern und zu kategorisieren. Eine Reihe von methodischen Herausforderungen begrenzt jedoch die Geschwindigkeit, mit der wir aus solchen Daten Erkenntnisse gewinnen können. In dieser Arbeit werde ich

argumentieren, dass langfristige Investitionen in digitale Infrastruktur und Open-Source-Tools erforderlich sind, um diese Herausforderungen zu überwinden. In Kapitel 2 wird die Crowdbreaks-Plattform vorgestellt, die die Grundlage für diese Arbeit bildet. Crowdbreaks ist ein Open-Source-Framework für Echtzeit-Datensammlung, kontinuierliche Crowdsourced Annotation und kontinuierliches Re-Training von Machine-Learning-Klassifikatoren. Im Gegensatz zu traditionellen Forschungsabläufen laufen die Projekte auf Crowdbreaks über einen längeren Zeitraum, was die Beobachtung von Gesundheitstrends über mehrere Jahre hinweg ermöglicht und die Algorithmen auf dem neuesten Stand hält. In Kapitel 3 quantifizieren wir das Auftreten von Konzeptdrift in impfstoffbezogenen Twitter-Daten, was den Bedarf an der Crowdbreaks-Plattform weiter untermauert. In Kapitel 4 verwenden wir die Crowdbreaks-Plattform, um die Stimmung gegenüber der neuartigen Gen-Editing-Technologie CRISPR/-Cas9 bis zu ihrer ersten Anwendung im Jahr 2013 zurückzuverfolgen und zu untersuchen, wie sich die öffentliche Meinung im Zusammenhang mit den jüngsten Skandalen rund um die Technologie verändert haben könnte. In Kapitel 5 wenden wir uns der COVID-19-Pandemie zu und analysieren, wer in den ersten Monaten der Pandemie zu Wort kam und wer gehört wurde. Kapitel 6 baut auf dieser Arbeit auf und untersucht die Dynamik der Twitter-Communities

[I] Translated with `deepl.com/translator`

während der COVID-19-Pandemie. Schließlich stellen wir in Kapitel 7 COVID-Twitter-BERT vor, ein domänenspezifisches Sprachmodell, das in verschiedenen nachgelagerten Anwendungen zur Verarbeitung natürlicher Sprache auf COVID-19-bezogenen Twitter-Daten verwendet wurde.

**Stichwörter**: Soziale Medien, Natürliche Sprachverarbeitung, Digitale Epidemiologie, Twitter, Gesundheit

# Contents

# 1 Introduction

Although Epidemiology is a relatively young field it has already experienced several revolutions. Be it the mapping of cholera cases in London by John Snow, the link between smoking and lung cancer, or the discovery of population-level genetic risk factors. These revolutions were driven by a novel combination of data sources and analysis techniques. The massive flood of data generated by electronic devices and digital means of communication combined with the advances made in the field of machine learning seem like the obvious ingredients for the next revolution in Epidemiology. This thesis is about the challenges we face while harnessing these novel data sources. It underlines the investment in digital infrastructure and open source tooling as important factors for this revolution to succeed.

## 1.1 Research context

### 1.1.1 The need to quantify health behaviors

Changing health behaviors is at the core of reducing death from many communicable and non-communicable diseases around the world (Bartholomew, Parcel, and Kok 1998; Glanz, Rimer, and Viswanath 2008). Health behaviors can be intentional or unintentional and include behaviors such as smoking, diet, physical activity, or getting vaccinated, just to name a few. Health behaviors are driven by attitudes, beliefs, norms, and cultural practices and are therefore often discussed on the level of the individual, e.g. by building on theories from the psychology of decision making. However, they can also be summarized for groups or entire populations. In this context, population-level health behaviors can be seen as dynamically changing, both temporally and spatially and can be traced and mapped like any other epidemiological or socio-economic parameter (Cohen, Scribner, and Farley 2000; Marmot 1998). Quantifying these factors may therefore help to guide public health policy and understand which strategies succeed in positively influencing health behaviors (Mokdad and Remington

2010).

Direct parallels have been drawn between the way infectious diseases and behaviors spread (Christakis and Fowler 2007; Bauch and Galvani 2013). In this context behaviors are also referred to as social contagion (Le Bon 1897). Adherence to control measures, such as the wearing of masks, is a typical example in which a social contagion influences a biological contagion. The link between behavior and disease spread is well-known, however their incorporation into mathematical models has been slow (Christakis and Fowler 2007). An important reason for this is the difficulty to quantify health behaviors. However, as we will see in this work, online social media provides new opportunities for measuring health behaviors in practice.

### 1.1.2 The case for social media data

Traditionally, the method of choice for understanding the attitudes that give rise to health behaviors has been the use of questionnaires or surveys. However, there are several shortcomings in using surveys as a basis to study health behavior. The shortcomings that are relevant in the context of this work are the following: (i) survey participants frequently answer questions falsely or at least inaccurately, a well studied effect known as response bias (Furnham 1986) (ii) at scale, surveys quickly become expensive and time-consuming (iii) singular surveys reflect a snapshot of the situation at a particular time and, on their own, are unable to show a trend (iv) they are limited by the the need to formulate an answerable survey question and therefore outcomes heavily reflect survey design. Arguably, some of these issues can be mitigated through establishing proper survey guidelines (Passmore et al. 2002), however others are inherent limitations of the method. In this thesis I will address how these limitations may be overcome with techniques from the field of digital epidemiology.

Almost all subfields of epidemiology now use at least some amount of digital methods (e.g. most surveys are conducted online) and frequently make use of digital data (e.g. simulation data). The key difference is that digital epidemiology makes use of digital data which was not created with the primary purpose of doing epidemiology and is usually collected outside of the public health system (Salathé et al. 2012; Salathé 2018). Examples for such data sources which can be repurposed for epidemiology are physical activity trackers, search engines, web logs, mobile phones, and social media services. Although all of these data sources are now used to study health behaviors, social media data allows the study of health behaviors in the context of publicly expressed opinions and attitudes (Paul and Dredze 2017).

Social media has been in a decade-long sprint of adoption and, as of October 2020, has reached the impressive mark of 4 billion active users, or 53% of the world's population, with 2 million new users joining every day (Kemp 2020). Although initially skewed towards adoption among the younger generation in the western world, these numbers show that social media usage has

become cross-generational and truly global. The COVID-19 pandemic has further accelerated these trends with the average user spending 2.5 h per day on social platforms. Every post and interaction on these platforms can be seen like a rich recording of the attitudes, beliefs, and behaviors in the context of a vast number of different topics. As it turns out, many health related issues are surprisingly often discussed online and some of them are rarely reported in a doctor's office. Harnessing such data for public good while respecting the user's expectation of privacy remains an important balancing act for the field (Vayena et al. 2015).

### 1.1.3 The nature of Twitter data

Throughout this thesis, publicly available data from the microblogging platform Twitter will be analyzed. For people outside the field, it might be surprising as to why Twitter data is so commonly used in research, given that Twitter is by far not the most popular social media platform. In fact, as of October 2020, there are 16 social media platforms with a higher monthly active user count than Twitter (Kemp 2020). The simple answer to this question lies in the access to this data for researchers, as for example Facebook and Instagram have severely limited access to their APIs (Application Programming Interface). However, the advantages of Twitter go beyond the trivial question of access. Although private accounts exist on Twitter, content is usually meant to be broadcasted widely and therefore available to the public. This makes Twitter fundamentally different from a platform such as Facebook, also from the perspective of expectation of privacy.

Another feature that makes Twitter interesting for research on health behaviors is the availability of real-time streaming API endpoints. As of 2020, Twitter data can be streamed from a 1% random sample stream or from the so-called filter stream, which delivers only tweets matching a given list of keywords. The filter stream is complete with respect to these keywords as long as the volume does not exceed 1% of the complete data stream. Beyond that threshold, the keyword stream is randomly subsampled to match the 1%-sample. This means that Twitter works well for answering targeted research questions (using the filter stream), but also allows for a more discovery-driven research workflow (using the sample stream).

Twitter data consists of individual posts (called "tweets"), which are semi-structured objects containing a text field of maximum 280 characters in length, as well as media (images and videos) and numerous metadata fields. Among others, the metadata includes creation time, user-related information, and geolocation information. Twitter users can interact with other users by either following them (i.e. being exposed to more of their content) or replying, liking and retweeting (i.e. re-sharing other people's content verbatim) other users' content. Although these are relatively simple mechanics, they allow us to study a wide variety of phenomena. Due to all these reasons Twitter has been called the Drosophila Melanogaster of social media research, highlighting the model characteristics of the platform (Tufekci 2014).

Research using Twitter data is fundamentally asking what was said, when, where and by whom. Additionally, the ability to retweet content provides another unique feature of Twitter, which allows to study attention mechanisms and virality patterns. Frequently, retweets are also studied in the context of a network, allowing the abstraction of users into groups or higher-level communities which may share common opinions or interests.

### 1.1.4 Understanding natural language

As the number of users on social media platforms increases, there is an ever-expanding diversity of topics discussed. With more than 4M tweets per day in the 1% sample stream, it is clear that tools from natural language processing (NLP) are required to automatically derive insight from this data. Although there have been massive advances in NLP in extracting semantic information from arbitrary text, significant challenges remain. These challenges led to the slightly paradoxical situation that it is easier to analyse tweets on the level of metadata (e.g. who retweeted whom) compared to what was actually being said. However, a better understanding of content is often crucial for the interpretation of results derived from metadata and may unlock multiple new avenues for research.

Supervised multi-label text classification is the method of choice for analysing social media content. This method takes as input unknown text and is able to classify text into a number of pre-defined subclasses. A classifier can be trained on human-annotated data and establishes a mapping between the given input text and the labels. In social media analysis, text classification is most commonly used for stance prediction in the context of opinion mining. However, due to its superiority over simple keyword matching methods, text classification can be used for any filtering or categorization process.

The recent advancements in the field of natural language understanding (NLU) have led to major improvements in text classification. Two major milestones can be identified: (1) the ability to generate meaningful vector representations of words (also known as word embeddings), fuelled by the method word2vec (Mikolov, Sutskever, et al. 2013; Mikolov, Chen, et al. 2013) and (2) the ability to generate context-aware word and sentence representations, fuelled by the BERT model (Bidirectional Encoder Representations from Transformers) (Vaswani et al. 2017; Devlin et al. 2018). Here, "fuelled" refers to the fact that key developments were made before the introduction of these models. Both approaches are able to learn word and sentence representations from massive bodies of raw text, scraped from the web and from books, which in turn lowers their dependency on task-specific human-annotated data.

With model performance surpassing human performance on multi-task benchmarks, such as GLUE (Wang et al. 2018), the question whether most challenges in NLP have effectively been "solved" arises. While the performance of transformer-based models are impressive, it

is important to acknowledge that some limitations of previous methods remain and that a number of new practical and ethical challenges have been introduced (Hovy and Spruit 2016). Most of all, significant work is still required for so-called low-resource languages, i.e. languages with fewer existing data sources and tools.

## 1.2 Problem

As discussed, state-of-the-art classifiers have a lower dependency on human-annotated data. Nevertheless, annotation data is still required, especially in order to validate classifiers for each new task. The dependency on task-specific human-annotated data therefore severely limits the speed at which we can generate new insight from social media data.

On the one hand, collecting annotation data is expensive, time-consuming, and to some, it may seem as a tedious but necessary step in the analysis of social media data. On the other hand, it is a craft, which may expose serious problems in the design of the research question. Investing effort into building a systematic annotation pipeline will therefore pay off both in terms of quality of results but also in terms of understanding of the underlying data. The questions of how to choose the annotation classes, how to annotate, and which data to select for annotation remain important challenges in the field (Kovashka et al. 2016).

At first glance, it seems like the process of collecting annotation data and fitting a model only needs to be completed once for a specific task and, if done well, the task can be considered as successfully "learnt" for the future. In fact, most of the development in machine learning operates under this assumption and it lays the basis for performance benchmarks to reflect real-world model performance. In fully observable, static environments this might indeed be the case, however social media data streams, are classical non-stationary systems (Costa et al. 2014). It is therefore to be expected that over time the underlying data distribution will change. Even worse, the definitions of the class labels, as encoded in the training data (i.e. the "ground truth" data) may change. For certain categorization problems it is also conceivable that new categories appear or existing categories may disappear. These phenomena are commonly referred to as concept drift.

Multiple theoretical approaches to detecting and overcoming concept drift have been proposed but in practice their adoption in NLP systems is limited. Concept drift may therefore affect classifier performance outside of the observation period the classifier was trained on, which severely limits the classifier's usefulness over time. Given the efforts required to train these algorithms this is an important challenge to be addressed in order for research activities to become more sustainable.

In this work I will argue that due to the complexity of phenomena like concept drift, a holistic

framework is required to address them in practice. Tasks like stance prediction on social media data can therefore not be reduced to the singular problem of training an algorithm, but should be seen as an interplay of continuous data collection, annotation and (re-)training of algorithms, a process which requires careful tuning and supervision.

## 1.3 Contribution

The contribution of this work lies in (i) building the open-source platform Crowdbreaks which may be used as a framework to overcome the aforementioned issues of concept drift by leveraging a crowdsourcing approach (ii) multiple applications of this platform on research questions in the field of public health and bioethics and (iii) a domain-specific machine learning model which has facilitated research in the context of COVID-19 and Twitter.

## 1.4 Outline

Chapter 2 will introduce the Crowdbreaks platform and present vaccine stance prediction on Twitter data as a core use case of the platform. In chapter 3 we will investigate the occurrence of concept drift in vaccination-related Twitter data and study the impact of concept drift on the interpretation of analysis results. We also report on preliminary findings regarding vaccine sentiment during the COVID-19 pandemic. In chapter 4 we present an application of the Crowdbreaks platform for studying public opinion on the novel gene technology CRISPR/Cas-9. We highlight the potential impact of scandals on long-term opinion making in the public. Chapter 5 investigates the role of experts during the COVID-19 crisis and analyses who was speaking and who was being heard during the pandemic. In chapter 6 we build on the methods developed in the previous chapter and study the dynamics of the retweet network that was shaped by the COVID-19 pandemic. Lastly, in chapter 7 we present COVID-Twitter-BERT, a domain-specific machine learning model which can improve the performance on several NLP tasks when used in the context of COVID-19 and Twitter.

## References

Bartholomew, L Kay, Guy S Parcel, and Gerjo Kok (1998). "Intervention mapping: a process for developing theory and evidence-based health education programs". In: *Health education & behavior* 25.5, pp. 545–563.

Glanz, Karen, Barbara K Rimer, and Kasisomayajula Viswanath (2008). *Health behavior and health education: theory, research, and practice*. John Wiley & Sons.

Cohen, Deborah A, Richard A Scribner, and Thomas A Farley (2000). "A structural model of health behavior: a pragmatic approach to explain and influence health behaviors at the population level". In: *Preventive medicine* 30.2, pp. 146–154.

Marmot, Michael G (1998). "Improvement of social environment to improve health". In: *The Lancet* 351.9095, pp. 57–60.

Mokdad, Ali H and Patrick Remington (2010). "Measuring health behaviors in populations". In: *Preventing chronic disease* 7.4.

Christakis, Nicholas A and James H Fowler (2007). "The spread of obesity in a large social network over 32 years". In: *New England journal of medicine* 357.4, pp. 370–379.

Bauch, Chris T and Alison P Galvani (2013). "Social factors in epidemiology". In: *Science* 342.6154, pp. 47–49.

Le Bon, Gustave (1897). *The crowd: A study of the popular mind*. T. Fisher Unwin.

Furnham, Adrian (1986). "Response bias, social desirability and dissimulation". In: *Personality and individual differences* 7.3, pp. 385–400.

Passmore, Cindy et al. (2002). "Guidelines for constructing a survey". In: *FAMILY MEDICINE-KANSAS CITY-* 34.4, pp. 281–286.

Salathé, Marcel et al. (2012). "Digital epidemiology". In: *PLoS Comput Biol* 8.7, e1002616.

Salathé, Marcel (2018). "Digital epidemiology: what is it, and where is it going?" In: *Life sciences, society and policy* 14.1, p. 1.

Paul, Michael J and Mark Dredze (2017). "Social monitoring for public health". In: *Synthesis Lectures on Information Concepts, Retrieval, and Services* 9.5, pp. 1–183.

Kemp, S (2020). *Digital 2020: October Global Statshot*. URL: `https://wearesocial.com/blog/2020/10/social-media-users-pass-the-4-billion-mark-as-global-adoption-soars` (visited on 12/03/2020).

Vayena, Effy et al. (2015). "Ethical challenges of big data in public health". In: *PLoS Comput Biol* 11.2, e1003904.

Tufekci, Zeynep (2014). "Big questions for social media big data: Representativeness, validity and other methodological pitfalls". In: *arXiv preprint arXiv:1403.7400*.

Mikolov, Tomas, Ilya Sutskever, et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Mikolov, Tomas, Kai Chen, et al. (2013). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30, pp. 5998–6008.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Wang, Alex et al. (2018). "Glue: A multi-task benchmark and analysis platform for natural language understanding". In: *arXiv preprint arXiv:1804.07461*.

Hovy, Dirk and Shannon L Spruit (2016). "The social impact of natural language processing". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 591–598.

Kovashka, Adriana et al. (2016). "Crowdsourcing in computer vision". In: *arXiv preprint arXiv:1611.02145*.

Costa, Joana et al. (2014). "Concept drift awareness in twitter streams". In: *2014 13th International Conference on Machine Learning and Applications*. IEEE, pp. 294–299.

# 2 Crowdbreaks: Tracking Health Trends Using Public Social Media Data and Crowdsourcing

**Published as:**
Müller M., and Salathé M. (2019). Crowdbreaks: Tracking Health Trends Using Public Social Media Data and Crowdsourcing *Frontiers in public health* 7 (2019): 81.

**Abstract**

In the past decade, tracking health trends using social media data has shown great promise, due to a powerful combination of massive adoption of social media around the world, and increasingly potent hardware and software that enables us to work with these new big data streams. At the same time, many challenging problems have been identified. First, there is often a mismatch between how rapidly online data can change, and how rapidly algorithms are updated, which means that there is limited reusability for algorithms trained on past data as their performance decreases over time. Second, much of the work is focusing on specific issues during a specific past period in time, even though public health institutions would need flexible tools to assess multiple evolving situations in real time. Third, most tools providing such capabilities are proprietary systems with little algorithmic or data transparency, and thus little buy-in from the global public health and research community. Here, we introduce Crowdbreaks, an open platform which allows tracking of health trends by making use of continuous crowdsourced labelling of public social media content. The system is built in a way which automatizes the typical workflow from data collection, filtering, labelling and training of machine learning classifiers and therefore can greatly accelerate the research process in the public health domain. This work describes the technical aspects of the platform, thereby covering the functionalities at its current state and exploring its future use cases and extensions.

## 2.1 Introduction

In the past years, data derived from public social media has been successfully used for capturing diverse trends about health and disease-related issues, such as flu symptoms, sentiments towards vaccination, allergies, and many others (Culotta 2010; Paul and Dredze 2011; Salathé and Khandelwal 2011; Paul and Dredze 2012; Parker et al. 2013). Most of these approaches are based on natural language processing (NLP) and share a common workflow. This workflow involves data collection, human annotation of a subset of this data, training of a supervised classifier, and subsequent analysis of the remaining data. The approach has proven promising in many cases, but it also shares a few shortcomings. A major drawback of this type of research process is that a model, which was trained on data from previous years, might not generalize well into the future. This issue, commonly known as concept drift (Widmer and Kubat 1996), may not necessarily be only related to overfitting, but may simply be a consequence of how language and content, especially on the internet, evolve over time. A similar effect has been suggested to be the main reason for the increasing inaccuracy of Google Flu Trends (GFT), one of the most well-known flu surveillance systems in the past (Ginsberg et al. 2009). After launching the platform in 2003, GFT's model had been retrained in 2009, which led to a significant improvement of its performance in the following years. However, during the influenza epidemic in 2012/13, the model's performance decreased again and overestimated the extent of the epidemic by a large margin. Shortly after, it was discontinued (Lazer et al. 2014; Butler 2013).

Apart from the issue of model drift, a second issue associated with current NLP models is that the collection of large amounts of labelled data, usually through platforms such as Amazon Turk[I] (MTurk), is very costly. Labelling a random subset of the collected social media data may be inefficient, as depending on the degree of filtering applied, large fractions of the collected data are possibly not relevant to the topic, and therefore have to be discarded.

Lastly, there is a growing interest in the public health field to capture more fine-grained categorizations of trends, opinions or emotions. Such categorizations could allow to paint a more accurate picture of the nature of the health issue at hand. However, multi-class annotations of a large sample of data again exponentially increases costs.

Here, we introduce Crowdbreaks[II], a platform targeted at tackling some of these issues. Crowdbreaks allows the continuous labelling of public social media content in a crowdsourced way. The system is built in a way which allows algorithms to improve as more labelled data is collected. This work describes the functionalities of the platform at its current state as well as its possible use cases and extensions.

[I] `https://mturk.amazon.com/`
[II] `https://www.crowdbreaks.org`

In recent years, a number of platforms have been launched which allow the public to contribute to solving a specific scientific problem. Among many others, examples of successful projects include the Zooniverse platform (formerly known as Galaxy Zoo) (Simpson, Page, and De Roure 2014), Crowdcrafting (*Crowdcrafting* 2018), eBird (a platform for collecting ornithological data) (Wood et al. 2011), and FoldIt (a platform to solve protein folding structures) (Khatib et al. 2010). Many of these projects have shown that citizen science can be used to help solve complex scientific problems. At the same time, there is a growing number of platforms which offer monetary compensations to workers for the fulfillment of microtasks (the most prominent example being MTurk). These platforms gain importance as the need for large amounts of labelled data for the training of supervised machine learning algorithms increases. Previous work focused mostly on efficiency improvement of large-scale human annotation of images, e.g. in the context of the ImageNet project (Russakovsky et al. 2015). Most of these improvements include better ways to select *which* data to annotate, *how* to annotate (which is a UI specific problem) and what type of annotations (classes and subclasses) should be collected (Kovashka et al. 2016). Online task assignment algorithms have been suggested which may consider both label uncertainty as well as annotator uncertainty during the annotation process (Welinder and Perona 2010; Ho and Vaughan 2012). Results suggest that this allows for a more efficient training of algorithms. More recently, a crowd-based scientific image annotation platform called Quantius has been proposed, showing decreased analysis time and cost (Hughes et al. 2017). To our knowledge, no similar work has been proposed with the regard to the human annotation of textual data, such as tweets.

## 2.2 Methods and tools

Crowdbreaks is a platform which aims at automatizing the whole process from data collection (currently through Twitter), filtering, crowdsourced annotation and training of Machine Learning classifiers. Eventually these algorithms can help evaluate trends in health behaviours, such as vaccine hesitancy or the risk potential for disease outbreaks.

Crowdbreaks consists of a data collection pipeline[III] ("streaming pipeline") and a platform for the collection of labelled data[IV] ("user interface"), connected through an API (Application Programming Interface), as schematized in figure 2.1.

[III] `https://github.com/crowdbreaks/crowdbreaks-streamer`
[IV] `https://github.com/crowdbreaks/crowdbreaks`

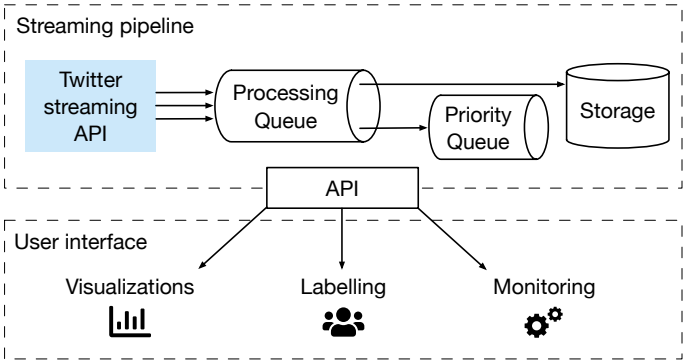



Figure 2.1: Overview of the architecture of the Crowdbreaks platform. The platform consists of a streaming pipeline (a message queueing system) and a user interface, linked through an API.

### 2.2.1 Streaming pipeline

Currently Crowdbreaks consumes data from the Twitter streaming API only, therefore the rest of this work will focus on tweets as the only data source. However, it could be extended to any textual data which can be collected in the form of a data stream through an API. The Twitter API allows for the filtering of tweets by a specific set of keywords in real-time. Tweets collected contain at least one exact match within certain fields of the tweet object. Incoming tweets are put on a background job queue for filtering, pre-processing, geo-tag enrichment, and annotation with metadata, such as estimated relevance or sentiment (more on this in section 2.4). After these processing steps, tweets are stored in a database. Based on a priority score (e.g. the uncertainty of a predicted label, see section 2.2.3) the tweet IDs are also pushed into a priority queue for subsequent labelling. Once the priority queue has reached a certain size, older items with low priority are removed from the queue and replaced with more recent items. Therefore the queue keeps a pool of recent tweets which are prioritized for labelling. Once a tweet has been labelled, it is ensured that the same tweet will be labelled by a certain number of distinct users in order to reach a consensus.

### 2.2.2 User interface

The user interface allows labelling of tweets based on answering of a sequence of questions. Arbitrary question sequences can be defined, which allow the annotation of multiple classes and subclasses to a single tweet. Most commonly, different follow-up questions would be asked depending on the answers given previously, e.g. whether or not the tweet is relevant to the topic at hand (see figure 2.2a). In the beginning of a question sequence an API call is made to the streaming pipeline to retrieve a new tweet ID from the priority queue (see section 2.2.1). Every question a user answers creates a new row in a database table, containing the respective user, tweet, question and answer IDs. After the user has successfully finished the question sequence the respective user ID is then added to a set, in order to ensure that the same tweet is not labelled multiple times by the same user.

Crowdbreaks supports multiple projects, each project may be connected to its own data stream from Twitter. New projects can be created through an admin interface, making it possible to control both the data collection, as well as to define project-specific question sequences. Eventually, visualizations, such as sentiment trends over time, may be presented to the public user, allowing the users to see the outcomes of their work. Crowdbreaks also features an integration of the question sequence interface with Amazon Turk, allowing the collection of labelled data through paid crowdworkers as an alternative to public users.

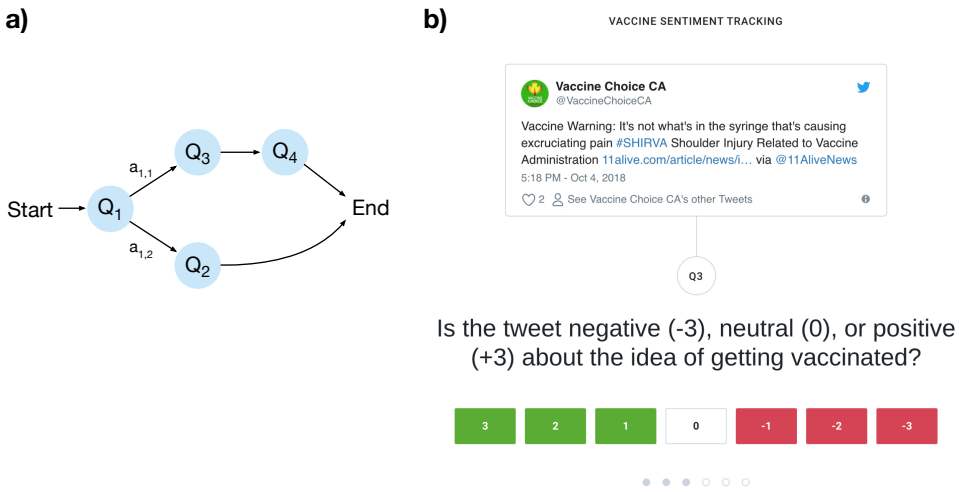



Figure 2.2: a) An example of a question sequence. Questions are denoted by $Q$, answers by $a$ and the arrows designate the possible transitions between questions. In the given example, different questions are reached depending on whether an annotator answers $Q_1$ with $a_{1,1}$ or $a_{1,2}$ allowing for an efficient and fine-grained annotation of the data. b) Screenshot of the annotation interface. Shown is a question for determining the vaccine sentiment of a tweet which has been deemed relevant to the topic.

### 2.2.3 Sentiment analysis

#### Algorithms

In recent years, algorithms for sentiment analysis based on word embeddings have become increasingly more popular compared to traditional approaches which rely on manual feature engineering (Bengio et al. 2003; Mikolov et al. 2013; Joulin et al. 2016). Word embeddings give a high-dimensional vector representation of the input text, usually based on a pre-trained language model. Although these approaches may not consistently yield better results compared to traditional approaches, they allow for an easier automatization of the training workflow and are usually more generalizable to other problems. This is a desirable property in the context of Crowdbreaks, as it aims to further automatize this process and retrain classifiers automatically as more labelled data arrive. Furthermore, pre-trained word embeddings based on large Twitter corpora are available in different languages, which also make them interesting for following health trends in languages other than English (Deriu et al. 2017). At its current state, the platform makes use of a baseline fastText classifier (Joulin et al. 2016), which is trained on a small set of labelled data. FastText allows for fast re-training and small model sizes which are desirable properties for active learning production environments.

#### Active Learning

Active learning frameworks have been proposed for a more efficient training of classifiers in the context of word embeddings (Kholghi et al. 2017; Zhang and Wallace 2016). These frameworks allow algorithms to be trained with a much smaller number of annotated data, compared to a standard supervised training workflow (see figure 2.3). The query strategy, which is usually related to label uncertainty, is generally the critical component for the relative performance speed-up of these methods. In the context of Crowdbreaks, we are not only prioritizing data with higher label uncertainty, but also data which is more recent in time. Therefore, we are faced with a trade-off between exploration and exploitation with regard to label uncertainty and timeliness of data. Crowdbreaks can serve as a framework to explore these challenges and find the right balance.

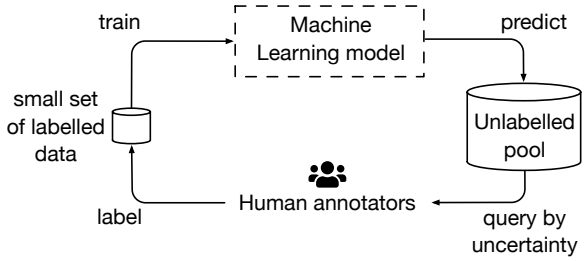



Figure 2.3: Crowdbreaks can be seen as an active learning framework which allows to improve algorithms as more labels are collected. In this example, an algorithm tries to learn sentiments from tweets and is given an initial small set of labelled data to be trained on. This algorithm may then be used to predict the labels and label uncertainty of newly collected tweets. Subsequently, tweets which the algorithm is most uncertain about will be presented to human annotators. As new labelled data is generated, the algorithm is retrained to further improve in performance.

### 2.2.4 Technologies used

Crowdbreaks uses a Python Flask API to interface between the components of the streaming pipeline and the user interface. The streaming pipeline makes use of Redis for the message queuing of the processing queue as well as the priority queue (see figure 2.1). Filtering and data processing, as well as NLP-related tasks are written in Python using the standard data analysis toolchain (numpy, scipy, nltk). Tweet objects are stored as flat files as well as in JSON format on Elasticsearch, which allows for an easier exploration and visualization of the data using Kibana. The user interface is built using Ruby on Rails with a postgres database backend in order to store the annotations, as well as user-related data.

All tools in the Crowdbreaks stack are open source and easy to deploy using Docker. The choice of tools was influenced by their long-term availability, community support and openness.

## 2.3 Results

The intensity, spread and effects of public opinion towards vaccination on social media and news sources has been explored in previous work (Seeman, Ing, and Rizo 2010; Salathé and Khandelwal 2011). Declines in vaccine confidence and boycotts of vaccination programs could sometimes be linked to disease outbreaks or set back efforts to eradicate certain diseases, such as polio or measles (Heidi J Larson and Ghinai 2011; Yahya 2007). In particular, the potential benefits of real-time monitoring of vaccine sentiments as a tool for the improved planning of public health intervention programs has been highlighted (Heidi J. Larson et al. 2013; Pananos et al. 2017; Bahk et al. 2016). Tracking of such sentiments towards vaccines is a primary use case of Crowdbreaks.

Between July 2018 and January 2019 tweets were collected through the Twitter Streaming API using a list of vaccine-related keywords[V] and predicted using a supervised bag-of-words fastText classifier[VI]. The classifier was trained on annotated data (collected through MTurk) provided in recent work by Pananos et al. (Pananos et al. 2017), resulting in micro-averaged precision and recall scores of 77.0%. The collected annotations include the label classes “positive”, “negative” and “other” (in this work denoted as “neutral”) with regard to the attitude towards vaccinations the tweets express. For a detailed reasoning of how and why these specific labels and keywords were selected, please refer to the work by Pananos et al. As shown in figure 2.4, we observe most of the discussion surrounding vaccination to be either neutral or positive. The fraction of data classified as “anti-vaccine” is below 10% and remains relatively constant at that level. Furthermore, we observe that the weekly tweet count exhibits

[V] The keywords include “vaccine”, “vaccination”, “vaxxer”, “vaxxed”, “vaccinated”, “vaccinating”, “vacine”

[VI] Data and code of the analysis are provided under `https://github.com/salathegroup/crowdbreaks-paper`

a large variance in terms of volume over time. This effect can be mitigated by calculating a normalized ratio os positive and negative counts in a rolling window of one month, which we call “sentiment index” in figure 2.4 (black curve). The sentiment index is calculated as $(r-\mu)/\sigma$, in which $r$ is the fraction of tweets predicted as positive among positive and negative tweets, and $\mu$ and $\sigma$ are the mean and standard deviation of this ratio, respectively. This value remains largerly constant over time and then increases after August 2018, due to an increase in the number of tweets predicted as “pro-vaccine” and stays at that level. Further investigation will be needed in order to understand the nature of this change. Although these results are only of preliminary nature they illustrate the potential of the platform to track health trends over time.

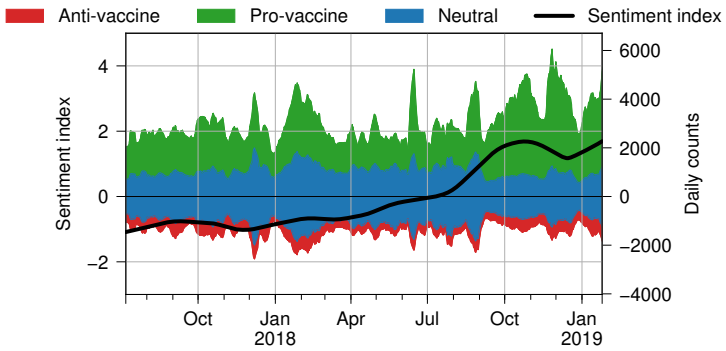



Figure 2.4: Real-time predictions of vaccine sentiments using Crowdbreaks. The data is based on a Twitter data stream filtered by vaccine-related keywords. Colored values indicate the stacked 1-week moving averages of tweet counts of the respective label class. The black curve denotes a sentiment index which reflects a lowess fit of the normalized ratio of counts of tweets predicted as postive and negative, aggregated in a 1 month window. The sentiment index reveals certain long-term trends irrespective of the high variance in volume over time.

## 2.4 Discussion

Here we introduced Crowdbreaks, an open tool allowing any researcher to start measurements of health trends in real-time from public social media content. As illustrated in the use case on vaccine sentiments, the platform can be used to monitor such sentiments and detect long-term shifts in health trends. Further analysis will be needed in order to reveal spatial sentiment distributions of the predicted vaccine sentiment as well as the correlation with vaccination coverage or disease outbreak data. Such analysis would however go beyond the scope of this work. Unlike in traditional settings of measuring vaccine sentiment, the platform involves crowdworkers as well as the general public to collect new annotations continuously over time. This allows to re-train models and counteract the problem of concept drift. In the future, we may use the plaftorm to measure more fine-grained categorizations of this data, hence improving our understanding of attitudes towards vaccination.

A major goal of the platform is the eventual incorportation of similar models into the public health decision-making process. In order to achieve this, there is a need for proper validation and benchmarking of machine learning models, which in turn increases both trust and transparency of algorithms used for such purpose (Salathé, Wiegand, and Wenzel 2018). In the future, annotation data generated on Crowdbreaks may be released in public challenges, thereby creating an open benchmark for a specific problem.

Although the platform focuses on the measurement of health trends, Crowdbreaks may also be used with regard to tracking flu or other infectious diseases in the future. However, disease prediction solely from Twitter data remains to be a hard problem. This is due to the fact that a precise understanding of the content (e.g. whether a tweet just raises awareness vs. actually reporting an infection) is crucial for the robustness of the model. Previous work has suggested hybrid models between Twitter and less volatile data sources (such a Wikipedia page rate clicks) to be superior for the purpose of outbreak tracking (McIver and Brownstein 2014; Santillana et al. 2015). Such hybrid models may serve as a future direction for disease prediction projects on Crowdbreaks.

**Acknowledgments.** We thank Sean Carroll, Yannis Jaquet, Djilani Kebaili and S.P. Mohanty for valuable discussions and help regarding the technical aspects of this project. Thanks also to Chloé Allémann for comments and Laura Symul for advice on visualization.

**Author Contributions.** MM built the platform, did the analysis and wrote large parts of the papers and made the figures. MS had the initial idea for the project, drafted the initial design of the platform and wrote the abstract of the paper. All authors revised the manuscript and made corrections.

**Data availability.** Data and code of the analysis are provided under `https://github.com/salathegroup/crowdbreaks-paper`.

**Conflict of Interest Statement .** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

Culotta, Aron (2010). "Towards detecting influenza epidemics by analyzing Twitter messages". In: *Proceedings of the first workshop on social media analytics*. ACM, pp. 115–122.

Paul, Michael J and Mark Dredze (2011). "You are what you Tweet: Analyzing Twitter for public health." In: *Icwsm* 20, pp. 265–272.

Salathé, Marcel and Shashank Khandelwal (2011). "Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control". In: *PLoS Computational Biology* 7.10. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1002199. arXiv: 1105.4502.

Paul, Michael J and Mark Dredze (2012). "A model for mining public health topics from Twitter". In: *Health* 11, pp. 16–6.

Parker, Jon et al. (2013). "A framework for detecting public health trends with twitter". In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, pp. 556–563.

Widmer, Gerhard and Miroslav Kubat (1996). "Learning in the presence of concept drift and hidden contexts". In: *Machine learning* 23.1, pp. 69–101.

Ginsberg, Jeremy et al. (2009). "Detecting influenza epidemics using search engine query data". In: *Nature* 457.7232, p. 1012.

Lazer, David et al. (2014). *The parable of google flu: Traps in big data analysis*. DOI: 10.1126/science.1248506.

Butler, Declan (2013). "When Google got flu wrong". In: *Nature* 494.February, pp. 155–156. ISSN: 1476-4687. DOI: 10.1038/494155a.

Simpson, Robert, Kevin R Page, and David De Roure (2014). "Zooniverse: observing the world's largest citizen science platform". In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 1049–1054. DOI: 10.1145/2567948.2579215.

*Crowdcrafting* (2018). URL: https://crowdcrafting.org (visited on 04/01/2018).

Wood, Chris et al. (2011). "eBird: Engaging birders in science and conservation". In: *PLoS Biology* 9.12. ISSN: 15449173. DOI: 10.1371/journal.pbio.1001220.

Khatib, Firas et al. (2010). "Crystal structure of a monomeric retroviral protease solved by protein folding game players". In: *Nature Structural and Molecular Biology* 18.10, pp. 1175–1177. ISSN: 15459993. DOI: 10.1038/nsmb.2119. arXiv: NIHMS150003.

Russakovsky, Olga et al. (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3, pp. 211–252. ISSN: 15731405. DOI: 10.1007/s11263-015-0816-y. arXiv: 1409.0575.

Kovashka, Adriana et al. (2016). "Crowdsourcing in computer vision". In: *arXiv preprint arXiv:1611.02145*.

Welinder, Peter and Pietro Perona (2010). "Online crowdsourcing: Rating annotators and obtaining cost-effective labels". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pp. 25–32. ISBN: 9781424470297. DOI: 10.1109/CVPRW.2010.5543189.

Ho, Chien-Ju and Jennifer Wortman Vaughan (2012). "Online Task Assignment in Crowdsourcing Markets". In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence* Kuhn 1955, pp. 45–51. arXiv: 1508.03593. URL: http://arxiv.org/abs/1508.03593.

Hughes, Alex J. et al. (2017). "Quantius: Generic, high-fidelity human annotation of scientific images at $10^5$ - clicks-per-hour". In: *bioRxiv (preprint)*. DOI: 10.1101/164087.

Bengio, Yoshua et al. (2003). "A Neural Probabilistic Language Model". In: *The Journal of Machine Learning Research* 3, pp. 1137–1155. ISSN: 15324435. DOI: 10.1162/153244303322533223. arXiv: arXiv:1301.3781v3.

Mikolov, Tomas et al. (2013). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781*.

Joulin, Armand et al. (2016). "Bag of tricks for efficient text classification". In: *arXiv preprint arXiv:1607.01759*.

Deriu, Jan et al. (2017). "Leveraging large amounts of weakly supervised data for multi-language sentiment classification". In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1045–1052.

Kholghi, M. et al. (2017). "Clinical information extraction using small data: An active learning approach based on sequence representations and word embeddings". In: *Journal of the Association for Information Science and Technology* 68.September, pp. 2543–2556. ISSN: 23301643. DOI: 10.1002/asi.23936.

Zhang, Ye and Byron Wallace (2016). "Active Discriminative Word Embedding Learning". In: *NAACL*. arXiv: 1606.04212. URL: http://arxiv.org/abs/1606.04212.

Seeman, Neil, Alton Ing, and Carlos Rizo (2010). "Assessing and responding in real time to online anti-vaccine sentiment during a flu pandemic". In: *Healthc Q* 13.Sp, pp. 8–15.

Larson, Heidi J and Isaac Ghinai (2011). "Lessons from polio eradication". In: *Nature* 473.7348, p. 446.

Yahya, Maryam (2007). "Polio vaccines—"no thank you!" barriers to polio eradication in Northern Nigeria". In: *African Affairs* 106.423, pp. 185–204.

Larson, Heidi J. et al. (2013). "Measuring vaccine confidence: Analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines". In: *The Lancet Infectious Diseases* 13.7, pp. 606–613. ISSN: 14733099. DOI: 10.1016/S1473-3099(13)70108-7.

Pananos, A. Demetri et al. (2017). "Critical dynamics in population vaccinating behavior". In: *Proceedings of the National Academy of Sciences*, p. 201704093. ISSN: 0027-8424. DOI: 10.1073/pnas.1704093114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1704093114.

Bahk, Chi Y. et al. (2016). "Publicly available online tool facilitates real-time monitoring of vaccine conversations and sentiments". In: *Health Affairs* 35.2, pp. 341–347. ISSN: 15445208. DOI: 10.1377/hlthaff.2015.1092.

Salathé, Marcel, Thomas Wiegand, and Markus Wenzel (2018). "Focus Group on Artificial Intelligence for Health". In: *arXiv preprint arXiv:1809.04797*.

McIver, David J. and John S. Brownstein (2014). "Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time". In: *PLoS Computational Biology* 10.4. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1003581.

Santillana, Mauricio et al. (2015). "Combining Search , Social Media , and Traditional Data Sources to Improve Influenza Surveillance". In: pp. 1–15. DOI: 10.1371/journal.pcbi.1004513.

# 3 Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic

**Published as:**
Müller M., and Salathé M. (2020). Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic *arXiv preprint* arxiv:2012.02197

**Abstract**

Social media analysis has become a common approach to assess public opinion on various topics, including those about health, in near real-time. The growing volume of social media posts has led to an increased usage of modern machine learning methods in natural language processing. While the rapid dynamics of social media can capture underlying trends quickly, it also poses a technical problem: algorithms trained on annotated data in the past may underperform when applied to contemporary data. This phenomenon, known as concept drift, can be particularly problematic when rapid shifts occur either in the topic of interest itself, or in the way the topic is discussed. Here, we explore the effect of machine learning concept drift by focussing on vaccine sentiments expressed on Twitter, a topic of central importance especially during the COVID-19 pandemic. We show that while vaccine sentiment has declined considerably during the COVID-19 pandemic in 2020, algorithms trained on pre-pandemic data would have largely missed this decline due to concept drift. Our results suggest that social media analysis systems must address concept drift in a continuous fashion in order to avoid the risk of systematic misclassification of data, which is particularly likely during a crisis when the underlying data can change suddenly and rapidly.

## 3.1 Introduction

Supervised and semi-supervised Machine Learning algorithms are now ubiquitous in the analysis of social media data. At the core of these algorithms is their ability to make sense of a vast amount of semi-structured real-time data streams, allowing them to automatically categorize or filter new data examples into, usually pre-defined, classes. Multi-class text classification has been successfully used in public health surveillance, election monitoring, or vaccine stance prediction (Salathé and Khandelwal 2011; Bermingham and Smeaton 2011; Brownstein, Freifeld, and Madoff 2009). In recent years such algorithms have also been developed to mitigate the negative effects of social media, such as in the detection of cyber-bullying, hate speech, misinformation, and automated accounts (bots) (Reynolds, Kontostathis, and Edwards 2011; Davidson et al. 2017; Shu et al. 2017; Davis et al. 2016).

The microblogging service Twitter has played a central role in these efforts, as it serves as a public medium and provides easy access to real-time data through its public APIs, making it the primary focus of this work. Twitter is well described as a classical example of a non-stationary system with frequently emerging and disappearing topical clusters (Costa et al. 2014). This poses problems for the aforementioned applications, as the underlying data distribution is different between training time and the time of the algorithm's application in the real world. This phenomenon is known as concept drift (Schlimmer and Granger 1986) and can lead to a change in performance of the algorithm over time.

It is important to distinguish concept drift from other reasons for performance differences between training and testing, such as random noise due to sampling biases or differences in data preprocessing (Žliobaitė 2010; Webb et al. 2016). A classic example of concept drift is the change in the meaning of classes, which requires an update of the learned class decision boundaries in the classifier. This is sometimes also referred to as real concept drift. Often, however, an observed performance change is a consequence of a change in the underlying data distribution, leading to what is known as virtual drift (Widmer and Kubat 1996; Tsymbal 2004). Virtual drift can be overcome by supplemental learning, i.e. collecting training data from the new environment. A good example are periodic seasonality effects, which may not be fully represented in the initial training data and only become fully visible over time. However, in practice it is usually very difficult (if not impossible) to disentangle virtual from real concept drift, and as a consequence they are treated as the same effect (Žliobaitė 2010).

On Twitter concept drift might appear on very different time scales and at different rates. Sudden shifts in a debate might be triggered by a quickly evolving news cycle or a catastrophic event. Concept drift may also be a slow process in which the way a topic is discussed gradually changes over time. A substantial amount of work has been dedicated to detecting and overcoming concept drift (Widmer and Kubat 1996; Žliobaitė 2010; Elwell and Polikar 2011). Three basic re-training procedures for overcoming concept drift have been proposed: (i) a

time-window approach, (ii) an incremental model, and (iii) an ensemble model (Costa et al. 2014). In the time-window approach, a sliding window of recent training examples is used to train an algorithm. In this approach, the algorithm ignores training data collected outside of that time window. The incremental model, in contrast, uses all previously collected training examples to re-train the model. Lastly, the ensemble model trains a model for each time window and uses the consensus of all previous models for future predictions. As found in (ibid.), in the case of hashtag prediction on Twitter data, the incremental method gave the best results.

Although sophisticated methods have been proposed to estimate concept drift in an unsupervised way (Katakis, Tsoumakas, and Vlahavas 2010; Yang, Wu, and Zhu 2008), in practice, a certain amount of re-annotation for both the detection and re-training of models seems unavoidable. The decision about which of the newly collected data to annotate points to an exploration-exploitation dilemma, which is usually addressed in the context of an active learning framework (Settles 2009). The Crowdbreaks platform (M. M. Müller and Salathé 2019) is an example of such a framework and has been built with the goal of exploring optimal solutions to this problem in order to overcome concept drift.

A change in the underlying data distribution might not necessarily have a negative impact on classifier performance. It is conceivable, for example, that a polarisation in a debate on Twitter about a topic could even lead to an improvement in classifier performance. It is therefore important to ask how much we should be worried about concept drift: even if model performance were to decrease, the real impacts on our analysis or interpretation might be negligible.

The consequences of concept drift are task-, environment-, and model-dependent (Žliobaitė, Pechenizkiy, and Gama 2016). Here, we will address concept drift in the specific case of vaccine stance classification. Vaccine stance classification on Twitter data has been widely studied and has shown promising links to vaccination decision making and vaccine uptake rates in different countries (Salathé and Khandelwal 2011; Bello-Orgaz, Hernandez-Castro, and Camacho 2017). The COVID-19 pandemic further emphasizes its importance, as evolving concerns about vaccines may significantly influence their effect (Johnson et al. 2020; Burki 2020).

To the best of our knowledge, only one study directly addressed concept drift in vaccine stance classification. In this study (D'Andrea et al. 2019) on tweets posted between September 2016 and January 2017 in Italian language, the authors did not find a substantial improvement of their model from incremental re-training before specific events. Re-training was performed on 60 newly annotated tweets from seven manually selected events. The authors conclude that either their original algorithm was already quite robust towards concept change, or that the newly collected training data was too small to see an effect.

Here, we use FastText (Joulin et al. 2016) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018), two commonly used models in social media text classification. Most work on the topic of concept drift was conducted using classical machine learning models, to which also FastText belongs. These types of models are very reliant on high-quality annotation data. More recently, models of the transformer family, such as BERT (ibid.), have been proposed, which require significantly less annotation data. In what follows, we will examine whether these two models also share different concept drift characteristics.

The goal of this work is to emulate a typical social media analysis study, in which data is collected for a certain period of time, and a supervised machine learning model is trained on a subset of annotated data. The model is then published and used to predict newly collected data. First, we will try to answer whether or not concept drift can be observed, and if so, at what rate it occurs. Second, we will investigate the influence of the study duration and the amount of annotation data used. Lastly, we will examine to what extent concept drift influences the final analysis outcomes, in this case a sentiment index.

## 3.2 Results

### 3.2.1 Observing concept drift

Throughout the 1188 day observation period, starting on July 1st, 2017 and ending on October 1st, 2020, a total of 57.5M English vaccination-related tweets were collected. A random subset of 11,893 tweets were annotated with respect to stance towards vaccines, which resulted in 5482 (46%) positive, 4270 neutral (36%), and 2141 negative (18%) labels (for further details see methods section 3.4.2). The dataset therefore bears clear label imbalance.

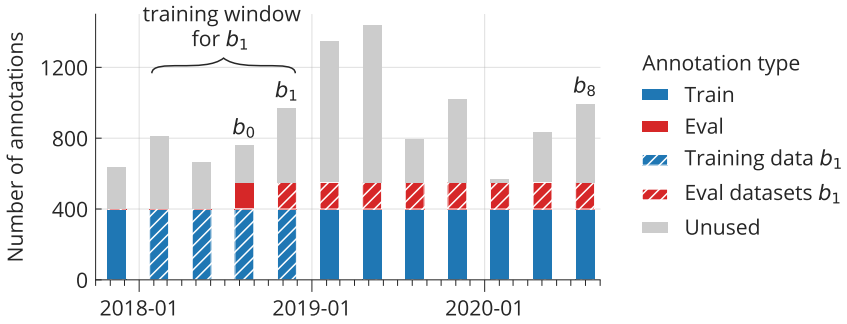



Figure 3.1: Training and evaluation datasets. Each 90 day bin consists of 400 samples of training data (blue) and 150 samples of evaluation data (red). Each trained model is using the most recent 1600 samples for training, which is an equivalent of 4 bins or 360 days. For illustration purposes, the training data for the second bin $b_1$ is indicated as blue with white stripes. The $b_1$ model is then evaluated on all future evaluation datasets, indicated as red with white stripes.

In order to observe whether classifiers experience drift in our dataset, we analysed the performance change of a model when predicting newly collected labelled data. For this we used a sliding time window approach, as first proposed in (Costa et al. 2014). We dissected the collected 11,893 annotations into 13 bins of 90 days each. From each bin we sampled 550 examples and split them into a train ($n$ = 400, 72%) and evaluation ($n$ = 150, 27%) set (see Figure 3.1). Each model was trained on a window of 4 bins of training data, which is equivalent to 1600 samples and a time span of 360 days. The models are subsequently evaluated on the evaluation set corresponding to the bin at the end of their training window as well as on all future evaluation sets. We repeat the process of binning, splitting, training and evaluating 50 times in order to yield a measure of confidence to our results.

Figure 3.2 shows the classifier performance at training time (square symbol) and the performance at each future evaluation dataset (circle symbol) for classifiers trained on different training windows (color). The upper left panel shows the results of these experiments for the FastText models. We will first compare the initial performance in terms of F1-macro score (i.e. the arithmetic mean of the class-level F1 scores) of the classifiers on a test dataset which was sampled from the last bin of the corresponding training window (square symbols). The initial performance of the first model is at 0.42, the subsequent models plateau at around 0.50, followed by a peak in fall 2019 with an abrupt decline in January 2020. This variability in the initial performance of models points to considerable differences between training datasets over time. The performance of the FastText models is quite low in general, which may be a consequence of the relatively small training dataset of 1600 examples and the lack of hyperparameter tuning.

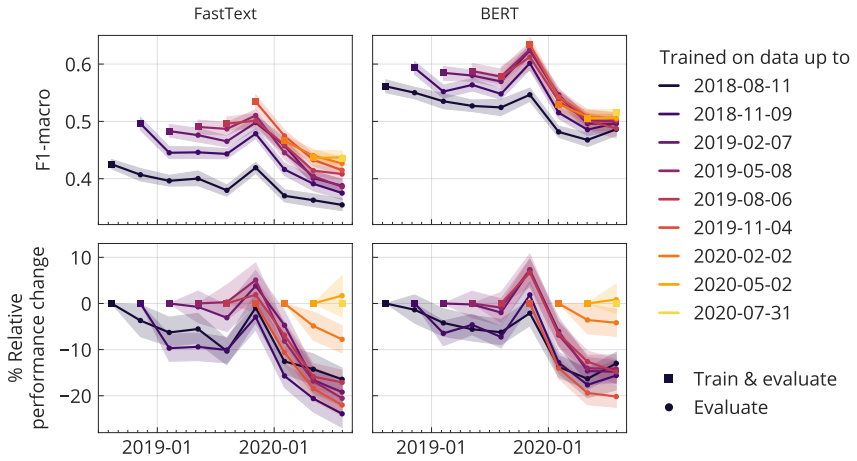



Figure 3.2: Model performance over time. The top row shows absolute performance (in terms of F1-macro), and the bottom row shows the relative performance change of models compared to training time. The columns show the result for the two different classifier types FastText and BERT. The square indicates the performance at training time, the circles correspond to performance of that same model on future evaluation sets (compare with Figure 3.1). Bands correspond to bootstrapped 95% confidence intervals resulting from 50 repeats.

Comparing the performance scores on future evaluation sets (circles) between models, we observe that the oldest models (black) generally perform worse than newer models (yellow) and that the ordering between models is preserved at all times. However, in order to disentangle this effect from the variability in initial performance, we compute the relative change in performance with respect to performance at training time (lower left panel). Starting from zero, the first model's performance drops quickly by around 5 %–10 %, followed by a rebound to initial performance in fall 2019, and ending in a sudden drop of approximately –20% in early 2020. The last drop indicates a very abrupt shift in concepts, twice as strong as during a comparable time window in 2019. In fall 2019, changes in the data distribution allowed all models to rebound to initial performance, with some even "over-performing" by 5% compared to training time. This is a sign that the data distribution was particularly easy to predict.

Further investigation of the F1-scores by class reveals that concept drift is especially impactful on the negative class, whereas the positive and neutral classes do not experience a significant drift (see Figure A.1). This could either indicate that anti-vaccine concepts are changing faster than pro-vaccine concepts or that the negative class is harder to learn due to label imbalance (cf. Figure A.4) and might, as a consequence, be more affected by virtual drift. We will further investigate this difference in the next section.

Comparing these results to the BERT models (upper right panel), the models show higher absolute performance but they experience a similar level of relative performance loss and similar drift patterns. This confirms that the observations are not model-specific but are likely to be observed in state-of-the-art semi-supervised machine learning models.

As previously stated, each model was trained on 1600 training examples over the previous 360 days. Experiments were conducted under fewer training examples (Figure A.2) and smaller training windows (Figure A.3) for FastText. As expected, training on fewer training examples leads to lower model performance, but we find the same drift patterns irrespective of the training data size. Reducing the training window while keeping the number of training examples constant does seem to have an impact on performance or drift patterns.

### 3.2.2 Explaining concept drift

Next, we will try to explain both the variance in initial performance, as well as the different rates of drift observed. We will investigate the effects of label imbalance, annotator agreement, and corpus variability on initial performance of models (Figure 3.3a-c). Additionally, we compare corpus similarity over time and discuss it in the context of concept drift (Figure 3.3d). In particular, we consider the first sampling (repeat) of the combined training ($n$ = 1600) and first evaluation set ($n$ = 150) for each training window. The provided measures therefore correspond to what the model "saw" during training and in the first bin of evaluation. Figure A.3

shows the equivalent metrics when limited to only the individual 90 day bins.

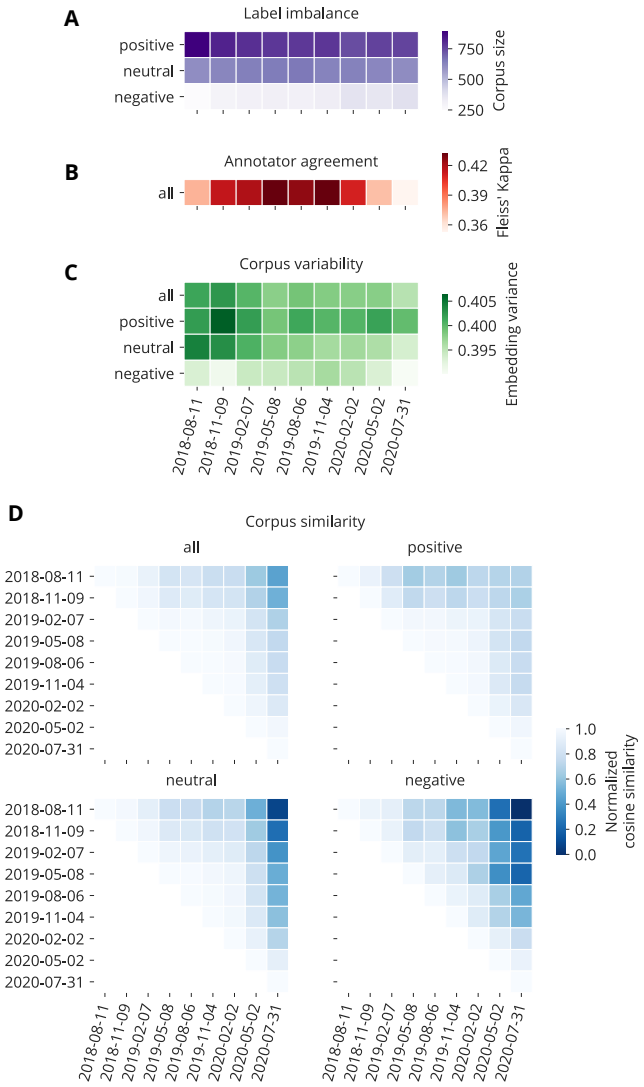



Figure 3.3: Properties of the combined training data and first evaluation dataset for each trained model. **A.** Distribution in the number of labels per class. **B.** Annotator agreement, measured by Fleiss' Kappa. **C.** Corpus variability in terms of the variance of sentence embeddings within a corpus. Variability is shown for the full corpus as well by class. **D.** Normalized cosine similarity between the mean corpus vectors (i.e. the mean of all sentence vectors in each corpus) for all data as well as by class.

### Label imbalance

Although the used training datasets are always of equal absolute size, they vary in the number of examples per class over time (see Figure 3.3a). It is commonly known that label imbalance can negatively impact model performance, which is also observed here (see Figure A.1). However, we note that label imbalance was highest in the very beginning of the observation period and continuously decreased towards a more balanced situation. Given the drop in initial performance in 2020, we conclude that label imbalance alone does not explain the observed variability in initial performance.

### Annotator agreement

We measure annotator agreement by computing the Fleiss' Kappa (Fleiss 1971) values for each dataset. Annotator agreement is initially low at 0.37, then increases to almost 0.45 and drops again to 0.36 in mid-2020. This overlaps very well with the initial performance trend observed in Figure 3.2. Variation in inter-annotator agreement may be a consequence of differences in annotation quality or difficulty of the annotation task, possibly hinting at semantic ambiguity of the text, as discussed next.

### Corpus variability

We use the BERT-large-uncased model to generate a 1024-dimensional sentence embedding vector (i.e. the vector of the CLS token) for each tweet text in the datasets. Note that this BERT model has not been trained on any of our datasets, but it is able to generate rich sentence embeddings due to having been pre-trained on large amounts of English text. Figure 3.3c shows the variance in the generated sentence embeddings across time. We note that overall, corpus variability is highest in the beginning of our observation period, and then decreases towards the end. Also, when considering the corpus variability by label class, we observe that negative samples have consistently lower variability compared to text labelled as positive. The neutral class seems to undergo a shift from high to low variability. In general, we may hypothesize that a lower variability points to lower separability in embeddings space, and therefore lower model performance. This hypothesis aligns with the observations made in terms of initial performance.

### Corpus similarity

Similarity was measured by calculating the cosine similarity between the mean vectors for each corpus. Low cosine similarity points to large semantic differences between datasets, which in turn could be an indicator for concept drift. In the top left panel ("all"), the datasets

are compared with each other. We observe that over time, corpus vectors are moving further away from each other. The biggest difference was observed between the two datasets furthest from each other in time (2018-08-11 and 2020-07-31). We also observe a bright area in the middle of the heatmap, which reveals that datasets between February 2019 and February 2020 are more similar to each other compared to datasets before (2018) or after (May & July 2020). This aligns well with the results in Figure 3.2: Most of the concept drift was observed in 2018 and following 2020, whereas models in 2019 didn't drift by a lot. When considering the corpus similarity by class, we can attribute most of these effects to the neutral and negative class. We therefore show that anti-vaccine content "drifts" faster than pro-vaccine content.

In conclusion, our observations point to the fact that the differences in initial performance of models are likely a consequence of low annotator agreement. The reason for this low agreement could be rooted in semantic ambiguity, as expressed by annotator agreement and corpus variability. The degree of concept drift on the other hand is best explained by our measure of corpus similarity.

### 3.2.3 Consequences of concept drift on real-time monitoring

Lastly, but perhaps most importantly, we highlight the impact of concept drift on the inference of the previously trained models when used for real-time monitoring of new data. We compare the predictions of a legacy model, which was trained in August 2018 and used for the two subsequent years, to a model we update (re-train) every 90 days. We compute the sentiment index $s$, which corresponds to the weekly mean of positive, neutral and negative predictions, when mapped to the numerical values of 1, 0 and −1, respectively. Figure 3.4 shows these sentiment trends for both the FastText and BERT model variants. We observe that, in the case of FastText, the sentiment predicted by the legacy model increased slightly until 2019 and then remained static. The updated models, however, show a downwards trend starting in mid-2019 and dropping further in 2020. By the end of our observation period the legacy model predicts a 0.3 points higher sentiment than the up-to-date models, while completely missing out on the downwards trend.

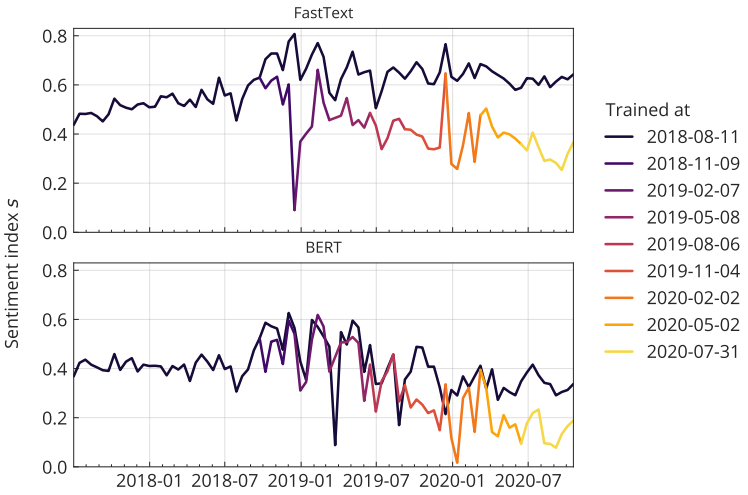



Figure 3.4: Impact of concept drift on the predictions made by FastText and BERT models. Each panel shows the comparison of a model which was trained in August 2018 (black) to a model which was continuously updated every 90 days (colored).

BERT models show a similar but smaller error, which is in agreement with our previous analysis. The legacy BERT model was in agreement with the updated models at the time of the first drop in 2019, but then started to diverge. We can therefore conclude that due to their higher overall performance, BERT models will have less severe deviations, but are not immune to effects of concept drift in the long run. We also note a large difference in the extent of positive and negative spikes between the legacy and re-trained models. Drift may therefore not only affect the mean sentiment trend but also sensitivity on shorter time scales, which could be problematic for real-time event or anomaly detection.

As previously stated, the sentiment trend of both the updated BERT and FastText models show a negative trend of the vaccine sentiment. Given the current debate surrounding novel vaccines for the Sars-CoV-2 virus, this finding is concerning from an epidemiological perspective. Note however, that the BERT models used for these predictions are of mediocre performance and future studies will be needed to confirm and interpret these trends.

## 3.3 Discussion

In this work, we investigated the effects of concept drift in vaccination-related Twitter data streams over a duration of three years. Using a sliding time window approach, we emulate a social media study in which (i) data is collected for one year, (ii) an algorithm is trained, and (iii) the algorithm is used in real-time monitoring of new data. While this may correspond to a common setup in social media analytics, we demonstrate here that without taking concept drift into account, the quality of the results will decay. Using a vaccine-related dataset from 2018–2020, we demonstrate how failing to take concept drift into account would have largely missed a rather dramatic decay in vaccine sentiment during the COVID-19 pandemic in 2020.

We find that overall, concept drift indeed occurred, which led to a decline in model performance of over 20% in the course of three years. However, most of this decline happened in only ten months. Concept drift therefore affected model performance at different rates throughout the observation period. Furthermore, the relative performance loss was not consistently negative but reverted to initial levels, or even slightly above that. These findings are consistent with the various ways real and virtual concept drift can occur. Although BERT models yielded higher performance scores, they are not immune to issues related to concept drift. On a relative scale, BERT models show the same degree of drift as the much less sophisticated FastText models.

In order to better understand the reasons for these phenomena, we investigate the properties of the used datasets. We can explain the large differences in initial performance of models with differences in semantic ambiguity of the text, as indicated by low inter-annotator agreement and low corpus variability. Occurrence of concept drift could be linked to differences in corpus similarity. In particular, we find that the negative class is responsible for most of the decay in

performance over time and also shows the strongest signs of drift. Anti-vaccine content may therefore change topics at an increased rate compared to both positive or neutral content.

A caveat of this study is that the results are based on classifiers of mediocre performance. Given the fact that the negative class was most affected by concept drift and is at the same time also the smallest class in our dataset, it is a fair question to ask whether concept drift would disappear given more annotation data and higher performance of models. It is conceivable that more annotation data would lead to a better representation of the training window. However, as results in a study on automated geo-location of tweets show (Dredze, Osborne, and Kambadur 2016), concept drift will still occur also under vast amounts of annotated data and adaptive re-training on even a relatively small corpus can overcome this drift.

Our results do not overlap with a previous study on vaccination-related Twitter data (D'Andrea et al. 2019), which did not find concept drift in an observation period between September 2016 and January 2017 in Italian language. The reason for this could be that the time scale analysed was too small to see an effect, or that concept drift was much smaller in that particular dataset.

It is safe to assume that the COVID-19 pandemic led to severe topical shifts in the vaccine debate, which ultimately translated into strong concept drift and model performance loss. Based on these results, it can be expected that future crisis situations would lead to similarly strong concept drift, thereby severely undermining the utility of social media monitoring tools that do not take concept drift into account. This is especially true for applications which are intended to be used exactly in such circumstances.

Although our work focused on the singular task of vaccine stance prediction, we believe that these results stress the general importance of addressing concept drift in any real-time social media monitoring project. Overcoming concept drift is a complex task, and many algorithmic solutions have been proposed. However, in order to succeed in practice, a tightly coordinated and fine-tuned framework for both the annotation and retraining of models is required. The Crowdbreaks platform (M. M. Müller and Salathé 2019) was built with the intention to address this issue and provide solutions for it.

## 3.4 Materials and methods

### 3.4.1 Data collection

This study is based on Twitter data collected through the Crowdbreaks platform (ibid.). Between July 1st, 2017 and October 1st, 2020 a total of 57.5M tweets (including 39.7M retweets) in English language by 9.9M unique users were collected using the public filter stream endpoint of the Twitter API. The tweets matched one or more of the keywords "vaccine", "vaccination",

"vaxxer", "vaxxed", "vaccinated", "vaccinating", "vacine", "overvaccinate", "undervaccinate", "unvaccinated". The data can be considered complete with respect to these keywords.

### 3.4.2 Annotation data

Human annotation of a subset of tweets was performed through the Crowdbreaks platform (ibid.). Tweets were anonymized by replacing user mentions and URLs with placeholders. Tweets between February 2nd 2018 and November 11th 2020 were sampled for annotation if they contained at least 3 words. Exact duplicates were removed. Annotators were asked the question "What is the attitude of the author of the tweet regarding vaccines?" and given the three options "negative", "neutral", and "positive". Annotation was performed both on Amazon Turk (mTurk) and, to a smaller extent (roughly 1% of all annotations) by public users on the Crowdbreaks website. We yield a dataset of 44,843 annotations (Fleiss' kappa of 0.30), which resulted in 11,893 three-fold annotated tweets. Tweets with less than two-third agreement were excluded and conflicts were decided through majority vote.

### 3.4.3 Training of classifiers

In this work we leverage two different classifiers: FastText (Joulin et al. 2016) and BERT (Devlin et al. 2018). For both models, hyperparameters were first tuned on the full annotation data to yield optimal performance and then fixed for further experiments. For FastText we used 10 dimensions, 500 epochs, a learning rate of 0.01, and using 1-gram embeddings. Optimal results were yielded by lower casing texts, converting them to ASCII and using the tags "user" and "url" for anonymization. BERT models of the type bert-large-uncased (pretrained in English language) were trained for 20 epochs, training batch size of 32, and a learning rate $2 \times 10^{-5}$ (using 10% warmup with linear decay to zero), as recommended in recent literature (Mosbach, Andriushchenko, and Klakow 2020; Dodge et al. 2020). FastText models were trained on a university cluster using the Crowdbreaks TEXT-CLASSIFICATION library[I] and BERT models were trained using Google Cloud v3-8 TPUs and the COVID-TWITTER-BERT library[II] (M. Müller, Salathé, and Kummervold 2020). For the purpose of predictions, text was preprocessed using the respective preprocessing approach.

**Data availability.** All data and code can be found on our public GitHub repository `https://github.com/digitalepidemiologylab/concept_drift_paper`.

[I] `https://github.com/crowdbreaks/text-classification`
[II] `https://github.com/digitalepidemiologylab/covid-twitter-bert`

**Author contributions.** M.M. collected the data, designed the experiments and analysed the data. M.M. and M.S. conceptualized the work and wrote the manuscript.

**Acknowledgments.** The authors would like to acknowledge Dr. Per Egil Kummervold and Dr. Burcu Tepekule for their valuable comments and discussions.

**Competing interests.** The authors declare no competing interests.

**Funding.** This work received funding through the Versatile Emerging infectious disease Observatory (VEO) grant as a part of the European Commission's Horizon 2020 framework programme (grant agreement ID: 874735). Compute resources (Cloud TPUs) were provided through Google's TensorFlow Research Cloud and the work was supported through Google Cloud credits in the context of COVID-19-related research.

## References

Salathé, Marcel and Shashank Khandelwal (2011). "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control". In: *PLoS Comput Biol* 7.10, e1002199.

Bermingham, Adam and Alan Smeaton (2011). "On using Twitter to monitor political sentiment and predict election results". In: *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pp. 2–10.

Brownstein, John S, Clark C Freifeld, and Lawrence C Madoff (2009). "Digital disease detection—harnessing the Web for public health surveillance". In: *The New England journal of medicine* 360.21, p. 2153.

Reynolds, Kelly, April Kontostathis, and Lynne Edwards (2011). "Using machine learning to detect cyberbullying". In: *2011 10th International Conference on Machine learning and applications and workshops*. Vol. 2. IEEE, pp. 241–244.

Davidson, Thomas et al. (2017). "Automated hate speech detection and the problem of offensive language". In: *arXiv preprint arXiv:1703.04009*.

Shu, Kai et al. (2017). "Fake news detection on social media: A data mining perspective". In: *ACM SIGKDD explorations newsletter* 19.1, pp. 22–36.

Davis, Clayton Allen et al. (2016). "Botornot: A system to evaluate social bots". In: *Proceedings of the 25th international conference companion on world wide web*, pp. 273–274.

Costa, Joana et al. (2014). "Concept drift awareness in twitter streams". In: *2014 13th International Conference on Machine Learning and Applications*. IEEE, pp. 294–299.

Schlimmer, Jeffrey C and Richard H Granger (1986). "Incremental learning from noisy data". In: *Machine learning* 1.3, pp. 317–354.

Žliobaitė, Indrė (2010). "Learning under concept drift: an overview". In: *arXiv preprint arXiv:1010.4784*.

Webb, Geoffrey I et al. (2016). "Characterizing concept drift". In: *Data Mining and Knowledge Discovery* 30.4, pp. 964–994.

Widmer, Gerhard and Miroslav Kubat (1996). "Learning in the presence of concept drift and hidden contexts". In: *Machine learning* 23.1, pp. 69–101.

Tsymbal, Alexey (2004). "The problem of concept drift: definitions and related work". In: *Computer Science Department, Trinity College Dublin* 106.2, p. 58.

Elwell, Ryan and Robi Polikar (2011). "Incremental learning of concept drift in nonstationary environments". In: *IEEE Transactions on Neural Networks* 22.10, pp. 1517–1531.

Katakis, Ioannis, Grigorios Tsoumakas, and Ioannis Vlahavas (2010). "Tracking recurring contexts using ensemble classifiers: an application to email filtering". In: *Knowledge and Information Systems* 22.3, pp. 371–391.

Yang, Ying, Xindong Wu, and Xingquan Zhu (2008). "Conceptual equivalence for contrast mining in classification learning". In: *Data & Knowledge Engineering* 67.3, pp. 413–429.

Settles, Burr (2009). *Active learning literature survey*. Tech. rep. University of Wisconsin-Madison Department of Computer Sciences.

Müller, Martin M and Marcel Salathé (2019). "Crowdbreaks: Tracking health trends using public social media data and crowdsourcing". In: *Frontiers in public health* 7, p. 81.

Žliobaitė, Indrė, Mykola Pechenizkiy, and Joao Gama (2016). "An overview of concept drift applications". In: *Big data analysis: new algorithms for a new society*. Springer, pp. 91–114.

Bello-Orgaz, Gema, Julio Hernandez-Castro, and David Camacho (2017). "Detecting discussion communities on vaccination in twitter". In: *Future Generation Computer Systems* 66, pp. 125–136.

Johnson, Neil F et al. (2020). "The online competition between pro-and anti-vaccination views". In: *Nature*, pp. 1–4.

Burki, Talha (2020). "The online anti-vaccine movement in the age of COVID-19". In: *The Lancet Digital Health* 2.10, e504–e505.

D'Andrea, Eleonora et al. (2019). "Monitoring the public opinion about the vaccination topic from tweets analysis". In: *Expert Systems with Applications* 116, pp. 209–226.

Joulin, Armand et al. (2016). "Bag of tricks for efficient text classification". In: *arXiv preprint arXiv:1607.01759*.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Fleiss, Joseph L (1971). "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5, p. 378.

Dredze, Mark, Miles Osborne, and Prabhanjan Kambadur (2016). "Geolocation for twitter: Timing matters". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1064–1069.

Mosbach, Marius, Maksym Andriushchenko, and Dietrich Klakow (2020). "On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines". In: *arXiv preprint arXiv:2006.04884*.

Dodge, Jesse et al. (2020). "Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping". In: *arXiv preprint arXiv:2002.06305*.

Müller, Martin, Marcel Salathé, and Per E Kummervold (2020). "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter". In: *arXiv preprint arXiv:2005.07503*.

# 4 Assessing Public Opinion on CRISPR/-Cas9

**Published as:**
Müller M., Schneider M., Salathé M., and Vayena E. (2020). Assessing Public Opinion on CRISPR-Cas9: Combining Crowdsourcing and Deep Learning. *Journal of medical Internet research* 22.8 (2020): e17830.

**Abstract**

**Background**: The discovery of the CRISPR-Cas9–based gene editing method has opened unprecedented new potential for biological and medical engineering, sparking a growing public debate on both the potential and dangers of CRISPR applications. Given the speed of technology development and the almost instantaneous global spread of news, it is important to follow evolving debates without much delay and in sufficient detail, as certain events may have a major long-term impact on public opinion and later influence policy decisions.

**Objective**: Social media networks such as Twitter have shown to be major drivers of news dissemination and public discourse. They provide a vast amount of semistructured data in almost real-time and give direct access to the content of the conversations. We can now mine and analyze such data quickly because of recent developments in machine learning and natural language processing.

**Methods**: Here, we used Bidirectional Encoder Representations from Transformers (BERT), an attention-based transformer model, in combination with statistical methods to analyze the entirety of all tweets ever published on CRISPR since the publication of the first gene editing application in 2013.

**Results**: We show that the mean sentiment of tweets was initially very positive, but began to decrease over time, and that this decline was driven by rare peaks of strong negative sentiments. Due to the high temporal resolution of the data, we were able to associate these peaks with specific events and to observe how trending topics changed over time.

**Conclusions**: Overall, this type of analysis can provide valuable and complementary insights into ongoing public debates, extending the traditional empirical bioethics toolset.

## 4.1 Introduction

Genome editing has many potential applications, ranging from gene therapy (Rosenberg et al. 1990) to crop enhancement (Comai et al. 1985) and production of biomolecules (Johnson 1983; Ye et al. 2000). While it has been possible to modify the genomes of eukaryotic cells since the 1980s, traditional methods have proven to be rather impractical, inaccurate, or impossible to use at scale (Thomas, Folger, and Capecchi 1986; Choulika et al. 1995; Jasin 1996; Porteus and Baltimore 2003). Accurately targeted gene editing has only become possible within the last decade (Barrangou et al. 2007; Jinek et al. 2012) using a CRISPR-Cas9–based method. In 2013, the method was further developed to be used on human cells (Cong et al. 2013; Mali et al. 2013), which allowed for the first successful experiment to alter the human germline DNA of non-viable embryos in April 2015 (Liang et al. 2015). The experiment, conducted by a group of Chinese scientists, raised ethical concerns among researchers and the general public about the potential far-reaching consequences of introducing germline modifications (Caplan et al. 2015; Bosley et al. 2015). Such ethical concerns include unexpected side effects on the evolution of humans, as well as cultural and religious arguments. In November 2018, Jiankui He announced the genetic editing of two viable human embryos with the goal of introducing HIV resistance (Regalado 2018). The work came to be known to a global public under the term "CRISPR babies" and was condemned by the scientific community as unethical, unnecessary, and harmful to the two babies (Cyranoski and Ledford 2018; Normile 2018).

As the costs of the technology drop further and usage becomes more widespread, governments and policy makers are faced with the challenging task of posing adequate ethical restrictions to prevent misuse. To gain time to introduce appropriate ethical frameworks, some scientists have called for a moratorium on genetically editing the human germline (Baltimore et al. 2015; Lander 2015; Lander et al. 2019). Previous studies on opinion towards GMO plants highlight how certain events or scandals (e.g. with respect to food safety) may have a major long-term impact on public opinion and later drive policy decisions (Legge Jr and Durant 2010; Paola Ferretti 2007; Marris 2001; Geller, Bernhardt, and Holtzman 2002). Understanding the public attitudes towards topics such as CRISPR is therefore of paramount importance for policy making (Travis 2015; National Academies of Sciences, Engineering, and Medicine and others 2017).

Several surveys have been conducted with the goal of evaluating the public's perception of CRISPR and genetic engineering in general (Weisberg, Badgio, and Chatterjee 2017; McCaughey, Sanfilippo, et al. 2016; Gaskell et al. 2017; Hendriks et al. 2018; McCaughey, Budden, et al. 2019). Such surveys have found that participants are largely in favor of the technology used for somatic purposes (eg, in the context of treatment) but less so for germline editing, especially if this is not for clearly medical purposes. Additionally, the studies underline certain demographic correlations (eg, that women, people belonging to ethnic minorities, and reli-

gious communities are more critical about the potential applications of CRISPR (Weisberg, Badgio, and Chatterjee 2017; Gaskell et al. 2017)). Somewhat unsurprisingly, the surveys also show that public views are not always aligned with expert opinions (McCaughey, Budden, et al. 2019). A recent study that explored coverage of news articles on CRISPR in North America between 2012 and 2017 found CRISPR to be overwhelmingly portrayed as positive and potentially overhyped in news media compared to the public's views (Marcon et al. 2019).

Social media platforms allow people to discuss a topic online with other people around the globe, creating an abundance of semistructured conversational data. Sentiment analysis provides a way to study people's perception of a topic, based on personal statements, and to process large volumes of such data in an automated way. Sentiment analysis has been used in the past to analyze different features such as emotions and polarity in several different contexts (B. Liu 2012). While traditional methods are based on linguistic expert knowledge (eg, rule-based methods), newer methods leverage machine learning, can be trained for specific contexts, and dominate traditional methods on polarity classification tasks (K. Ravi and V. Ravi 2015). Additionally, the supervised machine learning approaches have the advantage that the performance of a model for the specific context can be evaluated. The adaption to a specific context is particularly useful for tweets, which have a very specific, informal language (Kouloumpis, Wilson, and Moore 2011). Accordingly, machine learning methods have been successfully used for Twitter sentiment analysis (Severyn and Moschitti 2015; Müller and Salathé 2019). Most classical supervised machine learning algorithms for text classification (such as Naive Bayes or support vector machines [SVMs]) rely on manual feature extraction. Recently, a type of semisupervised machine learning model called Bidirectional Encoder Representations from Transformers (BERT) has been introduced to natural language processing (Devlin et al. 2018). BERT models are pretrained on large corpuses of raw text and can be adapted to a target task in a process called transfer learning. BERT models are based on the transformer, a neural network architecture that has been shown to outperform previously mentioned models in most natural language processing tasks, including text classification and sentiment analysis (Vaswani et al. 2017; Sun, L. Huang, and Qiu 2019). BERT has also been used in top-ranking submissions in the SemEval2019 challenges on detection of hate speech and offensive language in social media data (Basile et al. 2019; Zampieri et al. 2019).

In this study, we conducted the first analysis of a complete dataset of all tweets about CRISPR published over a 6.5-year period. The analyzed timespan includes the first experiment of CRISPR on human cells in 2013 but also recent events, such as the first genetic editing of viable human babies in November 2018. Furthermore, we make use of recent advances in text classification models, such as BERT (Devlin et al. 2018), which use semisupervised machine learning to generate a high-resolution temporal signal of the sentiment towards CRISPR over the observed timespan. By combining multiple text classification methods, we obtain results that can also be linked back to previous studies conducted with traditional methods, such as

surveys.

## 4.2 Methods

### 4.2.1 Overview

Our analysis consisted of 4 different explorative approaches, all of which build upon the sentiments of the tweets. Therefore, sentiment analysis represents the core of our analysis. In order to determine the sentiment for the entirety of tweets published over the last 6.5 years, we trained a predictive model on a previously manually annotated subset of the data. The process can be divided into 5 main tasks, which we describe in the following sections (see Figure 4.1 for an overview of the process): data collection, preparation, annotation, training, and analysis.

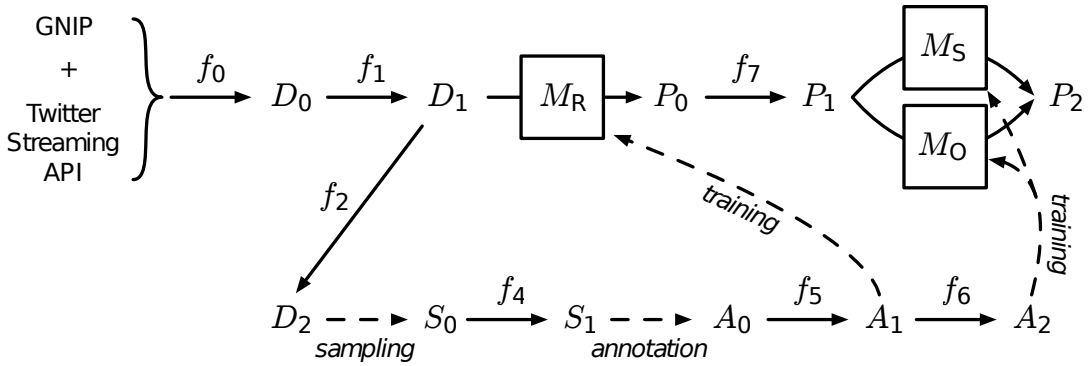



Figure 4.1: Overview of the data processing pipeline. Labels $f_{0-7}$ denote filtering steps, $D_{0-2}$ datasets, $S_{0-1}$ samples, $A_{0-2}$ annotation sets, and $P_{0-2}$ predictions. $M_R$, $M_S$, and $M_O$ represent machine learning models. API: application programming interface.

### 4.2.2 Data collection

The data set (denoted as $D_0$ in Figure 4.1) for our analysis consists of all tweets (including retweets, quoted tweets, replies, and mentions) that match the character sequence CRISPR (in any capitalization), have been detected to be in English language, and were published between January 1, 2013 and May 31, 2019. We retrieved these data either through the Twitter Streaming API or through GNIP, a Twitter subsidiary that allows access to historical data that were not retrievable through the Twitter Streaming API. The 3 aforementioned filtering conditions were used as parameters in the retrieval through Twitter APIs (denoted as $f_0$) as well as for the requested data from GNIP.

The number of tweets varied greatly over time, ranging from 4818 in 2013 to 445,744 in 2018, totaling 1,508,044 tweets by 348,502 distinct users (also refer to Multimedia Appendix B.1). Since the focus was on the overall evolution of the discourse provided by aggregated information, this study considered only the text in the tweet objects and ignored user-related information (such as location) or media content (such as photos or videos). In addition, any occurrences of Twitter handles and URLs in the text were anonymized (replaced by `@<user>` and `<url>`, respectively) to protect individuals.

### 4.2.3 Preparation

In a preparatory step, tweets suitable for annotation were selected from $D_0$. As an inclusion criterion, only tweets with ≤ 3 English words (after removal of stop words) were considered ($f_1$). Although a tweet with < 3 non-stop words may express a sentiment, we chose this threshold to ensure that the annotators had at least a minimal context to determine if the tweet was in fact relevant to the topic and what sentiment it expressed. The word count was determined by the help of NLTK's (Natural Language Toolkit, a python library for natural language processing) TweetTokenizer and English word and stop word corpora (Bird, Klein, and Loper 2009). The filtering and subsequent dataset operations and analysis were carried out using pandas, a python package for data analysis (McKinney et al. 2010). The resulting dataset $D_1$ ($n$ = 1,334,114) was used as the basis for the subsequent analysis. To avoid the annotation of duplicates, all retweets, quoted tweets, and other duplicates of tweets with the same text were removed, leading to dataset $D_2$ ($n$ = 433,930).

Next, we selected a random sample $S_0$ ($n$ = 29,238), so that we obtained a more or less evenly distributed number of tweets over the observed timespan. This was achieved by binning the data by all 77 months and selecting a constant number of tweets from each monthly bin. In contrast to a fully random sample, our sampling scheme contained no oversampling bias with regard to very recent content. Therefore, the generated sample was more representative of the whole observation period and accounted for the possibility that the nature of the tweets

changed notably over time.

### 4.2.4 Annotation

After generating the sample, the selected tweets were annotated through the Crowdbreaks platform (Müller and Salathé 2019), which uses crowdsourcing to annotate social media data. The platform allows for the creation of a question sequence that is then submitted in combination with a tweet as a task to MTurk (Amazon Mechanical Turk). The question sequence contained 3 questions for each task. The first question was on the relevance of the tweet to the topic of CRISPR-Cas9, allowing "relevant" and "not relevant" as possible answers. The second question was on the sentiment (positive, negative, or neutral), and the third question was on the organism (humans, human embryos, animals [other than human], plants, bacteria, multiple, not specified).

Before submitting the task to MTurk, the availability of the tweet was automatically checked. This was done in order to respect the user's right to either delete their content or set it to private after the time of data collection. Filtering by tweets that were still available yielded the sample $S_1$ ($n = 22{,}513$), which was subsequently annotated with regard to the 3 questions mentioned earlier. This resulted in annotation set $A_0$. To detect workers with questionable performance, the annotators' raw agreement was calculated, which denotes the fraction of the number of actual agreements over the number of possible agreements an annotator had with other annotators. An annotator was considered an outlier if this value was larger than 3 standard deviations from the mean, the annotator had less than 20 possible agreements with other annotators, or the annotator was involved in less than 3 separate tasks. All annotations by outlier annotators were subsequently removed. The resulting Fleiss' kappa agreement scores (Fleiss and J. Cohen 1973) were 0.81 and 0.28 for the questions of relevance and sentiment, respectively. Tweets for which a unanimous consensus of at least 3 independent annotators could be found were merged into dataset $A_1$. For the questions on sentiment and organism, only tweets that were labelled as relevant were considered and exported to $A_2$. This resulted in 3 cleaned datasets with annotated tweets for relevance ($n = 16{,}421$), sentiment ($n = 4718$), and organism ($n = 1196$), which we used to train 3 classifiers.

### 4.2.5 Training

In order to classify the data with regard to relevance, sentiment, and organism, we constructed 3 classifiers: $M_R$, $M_S$, and $M_O$, respectively. The classifiers tried to predict the respective labels from the text of the tweet alone. In the process, we analyzed the performance of 4 different classifier models: Bag of Words (BoW), Sent2Vec sentence embeddings (Pagliardini, Gupta, and Jaggi 2017) coupled with SVMs (Cortes and Vapnik 1995), FastText (Joulin et al. 2016), and

BERT (Devlin et al. 2018). The tokenization process was different for each model class. In order to evaluate the models, the cleaned annotation data were shuffled and split into training (80%) and test sets (20%).

For the BoW, SVM, and FastText models, we used supervised learning to train the 3 classifiers for sentiment, relevance, and organisms. A limited search of model parameters was conducted. In the case of BERT, we started from the pretrained (unsupervised) English BERT-large-uncased model provided by the Huggingface library (Wolf et al. 2019) and conducted an additional step of unsupervised, domain-specific pretraining on our raw body of tweets. This model then served as the basis for the final, supervised training step (i.e., fine-tuning the general model with classifier-specific labelled data). For this fine-tuning step, a learning rate of $1 \times 10^{-5}$ and 2 epochs of training were used. This work was conducted using PyTorch (Paszke et al. 2019) and the Huggingface library (Wolf et al. 2019).

After the training phase, we selected the classifiers for relevance, sentiment, and organism ($M_R$, $M_S$, and $M_O$ in Figure 4.1) by evaluating the performance of the models on the test set (see Multimedia Appendix B.1 for different model performances). The fine-tuned BERT model was the best performing sentiment classifier ($M_S$), with a macro-averaged F1 score of 0.727 ($F1_{\text{positive}} = 0.827$, $F1_{\text{neutral}} = 0.715$, $F1_{\text{negative}} = 0.639$). The fine-tuned BERT model was also found to be the best performing model for the relevance ($M_R$) and organism ($M_O$) classifiers with macro-averaged F1 scores of 0.91 ($F1_{\text{related}} = 0.997$, $F1_{\text{unrelated}} = 0.823$) and 0.89 ($F1_{\text{humans}} = 0.873$, $F1_{\text{embryos}} = 0.762$, $F1_{\text{animals}} = 1$, $F1_{\text{plants}} = 0.889$, $F1_{\text{bacteria}} = 0.909$, $F1_{\text{unspecific}} = 0.902$), respectively.

### 4.2.6 Prediction

For the analysis, the best performing model (fine-tuned BERT) for relevance $M_R$ was used to predict dataset $D_1$ and yield the predicted dataset $P_0$ ($n$ = 1,334,114) of the same length containing a label for relevance. Next, all tweets predicted as not relevant were removed from $P_0$, yielding the dataset $P_1$ ($n$ = 1,311,544). This dataset was then used to predict sentiment and organism using the models $M_S$ and $M_O$, resulting in the final dataset $P_2$.

### 4.2.7 Analysis

In our analysis, we used the sentiments in relation to tweet activity (number of tweets), topics of the tweets (hashtags), organisms the tweets were talking about (predicted), and themes identified from previous studies on CRISPR mentioned earlier (through regular expressions) to gain different kinds of insights. Wherever we used sentiments for numerical calculations, we used +1 for positive, 0 for neutral, and −1 for negative sentiment. Further, we extrapolated the numbers for 2019 where applicable for better comparison since we only had data until

May 31, 2019. The different parts of the analysis are explained in more detail in the following paragraphs.

The first part of the analysis was concerned with the development of the sentiment in relation to the number of tweets over time. The detection of a temporary deviation from the general sentiment was of particular interest. While we included all tweets for the analysis of activity, we excluded tweets with neutral sentiment for the analysis of sentiment to make deviations more visible. We aggregated activity and sentiments on a daily basis. For the sentiments, however, the sentiment value of a specific day was determined by taking the mean value of all positive and negative sentiments within a sliding 7-day window centered around that day (±3 days). Further, we tested whether the yearly means based on the positive and negative tweet sentiments were significantly different from each other with the Welch's $t$-test (Ruxton 2006; Alhabash et al. 2018) using scipy's statistics module (Jones, Oliphant, Peterson, et al. 2001). We then used scipy's module for peak detection (ibid.) to detect events of interest, using a relative prominence cut-off of 0.2. In order to identify potential sources for the change in sentiment, we manually identified major events that relate to CRISPR.

In the second part, we used the predictions of the model $M_O$ and the sentiments to compare the development of the sentiment for different organisms. We calculated the mean sentiments over a month and excluded all months that did not have at least 100 tweets for the respective organism. Further, we used the same test as we did for the yearly means to compare the organism class means based on the individual tweet sentiments (positive, negative, and neutral).

Third, we analyzed hashtags as a proxy for the topics a user was talking about in his or her tweet. The hashtag #CRISPR was excluded from the analysis since CRISPR was the overarching topic all tweets had in common. We counted the occurrences of every hashtag per year. We used the exact hashtags and did not group similar hashtags. For example, the hashtags #crisprbaby and #crisprbabies were treated as different hashtags. We did this due to the difficulty of automatically matching similar hashtags, since they can be a composition of multiple words that made strategies like stemming not straightforward. For each hashtag and year, we then calculated the mean sentiment and selected the 15 most common hashtags for each year for further analysis. We then manually compared how these top 15 topics per year increased and decreased in popularity throughout the years, as well as how the sentiments for these topics changed.

In the fourth and last part of our analysis, we based our analysis on the earlier conducted studies. We conducted a literature search in scientific databases according to a predefined search strategy (see Multimedia Appendix B). The search was conducted in the fall of 2017. We reviewed the resulting studies and identified the reasons why people had a positive or negative attitude towards CRISPR and issues that concerned them. In the process, we summarized

these reasons and concerns for each study and compiled a list with a short description for each of them. Since there was thematic overlap across the studies, we inductively determined the themes of these summaries and compiled a regular expression representing each theme based on the summary text. Additionally, we added themes and corresponding regular expressions based on publications and events that occurred between the fall of 2017 and the summer of 2019. The regular expressions then allowed us to automatically check for matches on the entire Twitter dataset as a proxy for the presence of the themes that occurred in the studies. See Multimedia Appendix B.2 for the themes and regular expressions.

## 4.3 Results

### 4.3.1 Overview

Our analysis includes over 1,300,000 tweets (dataset $P_1$, $n$ = 1,311,544) over the time period from January 1, 2013 until May 31, 2019. The predicted sentiments of the tweets were predominantly positive (685,578/1,311,544; 52.3%) or neutral (528,196/1,311,544; 40.3%). Only a minor fraction was predicted as negative (97,770/1,311,544; 7.5%). In the following sections, we report our results focusing on different aspects.

### 4.3.2 Temporal Development

Figure 4.2 shows a temporal analysis of the predicted sentiments in relation to key historical events surrounding CRISPR. A sentiment of zero indicates an equal portion of positive and negative tweets, and the values 1 and −1 indicate a signal with only positive or negative tweets, respectively. Figure 4.2A shows the sentiments between July 2015 and June 2019. The time period before July 2015 was excluded, as activity was too low for a high-resolution sentiment signal. The sentiment remained mostly positive, with an average of 85% positive tweets and only 15% negative tweets. Especially over the initial time period until March 2017, the sentiment shows little variation. After that, the sentiment reveals a series of sharp negative spikes, on multiple occasions dropping below zero. Over the observed time period, the sentiment shows a slight negative trend (slope of $-0.061\,y^{-1}$, standard error $0.005\,y^{-1}$), as indicated by the linear trend line in orange. The differences between the yearly means of the tweet sentiments were all significant ($P < 0.001$; see Multimedia Appendix B.3 for all means, standard deviations, and test statistics).

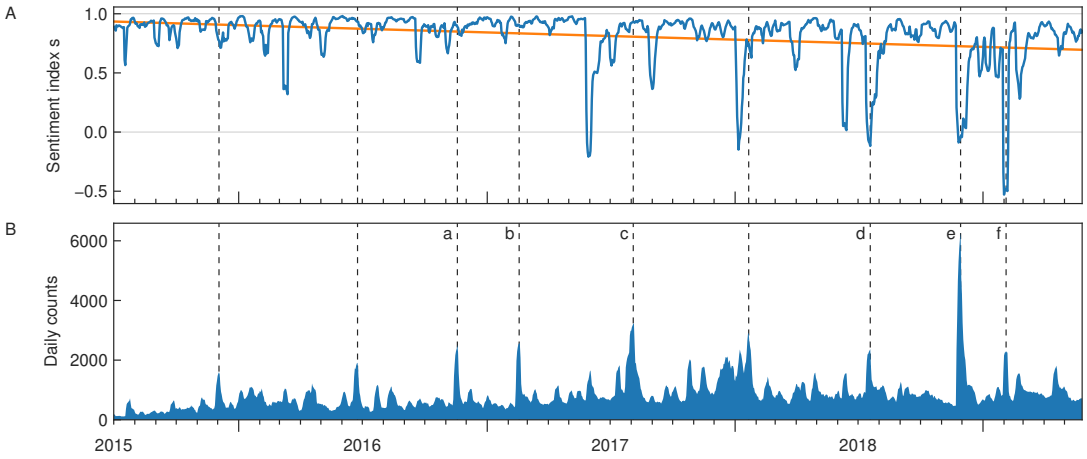



Figure 4.2: **A.** Predicted sentiment towards CRISPR between July 2015 and June 2019. The blue curve denotes the sentiment $s$, which is calculated as the mean of the weighted counts of positive and negative tweets over a centered rolling window of 7 days. The orange curve denotes a linear fit of the sentiment $s$. **B.** Daily counts of all analyzed tweets. The blue area shows the daily sum of positive, negative, and neutral tweets as the mean within a 7-day centered rolling window. All peaks above a relative prominence of 0.2 are marked with dashed lines; a–f denote peaks that coincide with certain events.

We then compared the sentiment curve to the observed activity surrounding CRISPR in the same time span, as shown in Figure 4.2B. Shown are the mean daily counts of the sample $P_1$ over a sliding window of 7 days. Activity varied considerably, with an average baseline of about 1000 tweets per day and peaks of up to roughly 6000 tweets per day.

We detected 9 peaks of interest. They are marked with dashed lines in Figure 4.2. When comparing peaks of high activity to the sentiment, it can be seen that peaks of high activity before mid-2018 did not result in a negative sentiment response. Peaks of strong negative sentiment started to appear in 2017 but it was not accompanied by the same level of activity until after 2018.

In a second step, major news events were manually mapped to coinciding peaks (for a full list, see Multimedia Appendix B.4). A subset of these peaks was marked with letters a–f in Figure 4.2B for illustrative purposes. In all cases, the most retweeted tweet within days of the peak was linking a news article describing the event. The events include the first use of CRISPR in humans by a group of Chinese scientists in November 2016 (peak a) and the US Patent Office deciding in favor of the Broad Institute (peak b). Both of these events did not lead to a significant change in sentiment. Peak c coincides with the publication of a study that reported the correction of a mutation in human embryos (Ma et al. 2017), causing widespread media attention and, as before, did not cause a drop in sentiment. However, in July 2018, a study by the Wellcome Sanger Institute (Kosicki, Tomberg, and Bradley 2018) warned about serious side effects, such as cancer, that CRISPR could have when used in humans (peak d). This peak led to a clear negative response in the sentiment index and marks the first negative peak with high media attention. When researcher He Jiankui revealed creating the world's first genetically edited babies in November 2018 (Regalado 2018) (peak e), the highest activity was recorded. Although He's revelation caused a strong negative signal, the strongest negative sentiment was recorded shortly after, in February 2019 (peak f). This event coincides with the re-emergence of a news story from August 2017 when biohackers managed to encode a malware program into a strand of DNA (Greenberg 2017).

### 4.3.3 Organisms

In order to improve our understanding of the sentiment signal, the data were predicted with respect to which organism each tweet was about (see the Methods section). We predicted the organism of the tweets in the dataset $P_1$ ($n$ = 1,311,544) resulting in the classes animals (7.6%), bacteria (2.4%), embryos (4.3%), humans (30.3%), plants (4.9%), and unspecified (50.6%). It is noteworthy that more than half of all tweets do not specifically refer to an organism in the context of CRISPR. After unspecified, the class humans is the second largest group, followed with some margin by animals (eg, mice for animal testing), plants, and embryos. The classes humans and embryos combined account for a little more than one-third of all tweets. Tweets

specifically mentioning CRISPR in the context of bacteria were rather rare.

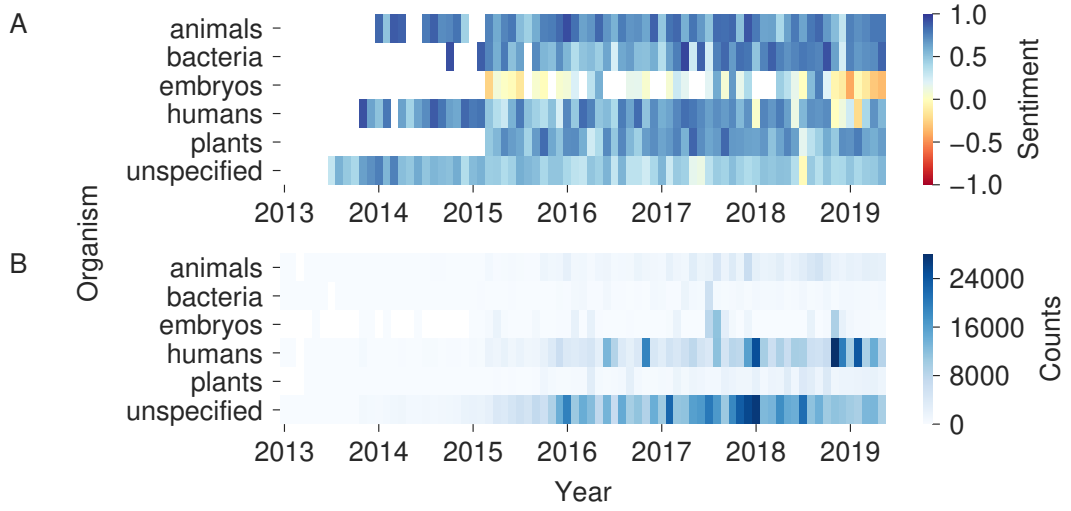



Figure 4.3: **A.** Heatmap of monthly sentiments by predicted organism. The sentiments were calculated as the mean of the weighted counts by sentiment (the weights included -1, 0, and 1 for negative, neutral, and positive tweets, respectively) for each month and organism class. Blue and red colors indicate positive and negative sentiment values, respectively. The sentiments of heatmap cells with < 100 tweets of that month and organism are transparent. **B.** Monthly counts by predicted organism.

Figure 4.3A shows the monthly sentiment for each organism class, which are based on the monthly counts shown in Figure 4.3B (all monthly means and standard deviations can be found in Multimedia Appendix B.5). Of all classes, embryos exhibited the most negative-leaning sentiment (mean sentiment 0.14 over all monthly means) and was also the class with the strongest variations between months (SD 0.27). Further, a relatively high sentiment was measured for the classes animals (mean 0.70, SD 0.14), bacteria (mean 0.65, SD 0.18), and plants (mean 0.61, SD 0.14), followed by the class humans (mean 0.58, SD 0.23), which showed a dip in the sentiment in the months following November 2018. The class unspecified had a slightly lower sentiment (mean 0.45, SD 0.13) compared with the other classes. In addition to this monthly breakdown, the differences between the organism class means based on the individual tweets were all significant ($P < 0.001$), except for the difference between the class means of bacteria and plants with a 3.8% probability of occurring by chance ($P = 0.038$; see Multimedia Appendix B.3 for all test statistics).

### 4.3.4 Hashtags

The most frequently used hashtags of every year revealed the topics of highest interest and how they evolved over time (see Figure 4.4). Naturally, the occurrences of individual hashtags increased over the years along with the total number of tweets. Certain very common hashtags, such as #dna, #science, #biotech, or #geneediting and #genomeediting, appeared as top hashtags in multiple years. When relating the hashtags with the sentiment of the text they appeared in, we can see that most of these common hashtags were used in the context of a positive or very positive sentiment. The 3 hashtags with the most positive sentiments and that were used at least 100 times were #cancer (mean sentiment 0.85, SD 0.36) in 2015, #hiv (mean 0.90, SD 0.34) in 2016, and #researchhighlight (mean 1.00, SD 0.06) in 2019. It is also notable that #science was among the 5 most common hashtags in every year except for 2013 and was consistently related to a positive sentiment, with means between 0.52 (in 2018) and 0.74 (in 2013).

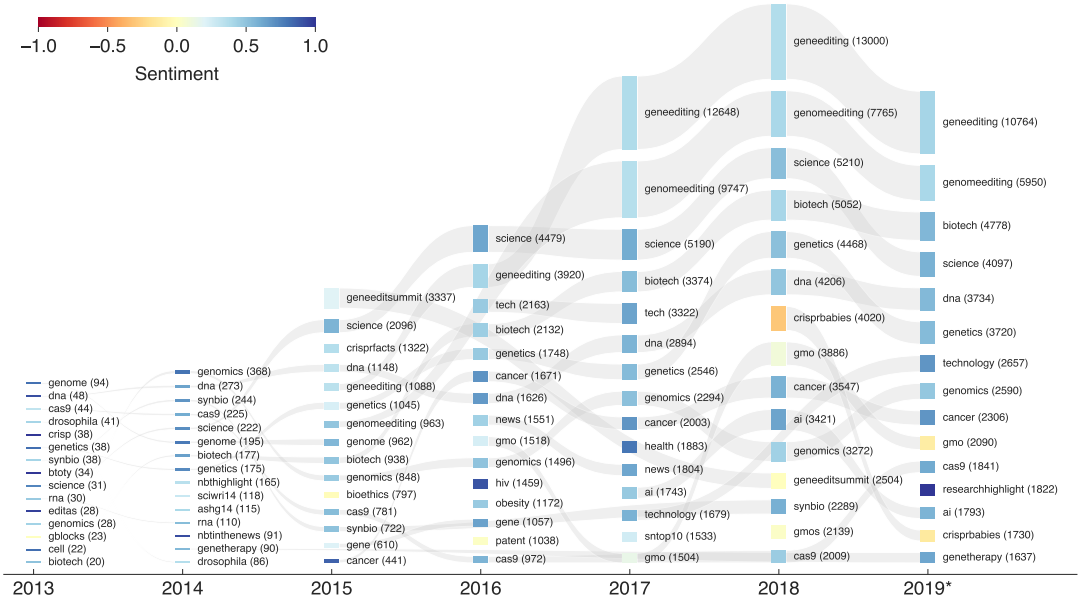



Figure 4.4: Visualization of the sentiment associated with the most frequently used hashtags every year. For every year, the 15 hashtags with the highest counts for that year are included (the hashtag #crispr was excluded). The hashtags are sorted by yearly counts (indicated by the bar height), where the hashtag with the highest count is at the top. The color represents the average sentiment for the respective hashtag, with blue representing a very positive sentiment and red representing a very negative sentiment. If a hashtag is listed in multiple years, the occurrences are linked with a gray band. The number of tweets with the hashtag is indicated in parentheses next to the respective hashtag. For the year 2019, the counts were extrapolated from the months before June to the full year.

Only a few hashtags were related to negative sentiments. The most prominent one was #crisprbabies, with mean sentiments of –0.30 (SD 0.65) in 2018 and –0.13 (SD 0.63) in 2019, followed by #gmo (mean –0.11, SD 0.76) in 2019, #bioethics (mean –0.02, SD 0.45) in 2015, and #geneeditsummit (mean –0.01, SD 0.46) in 2018. It is worth noting that the hashtag #geneeditsummit only appeared in 2015 and 2018 and that its associated sentiment dropped from 0.20 to –0.01. The hashtag refers to the two summits on human genome editing, which were held in Washington D.C. in 2015 and in Hong Kong in November 2018, coinciding with the first gene editing of viable human embryos. Similarly, the hashtag #gmo became slightly more negative in 2018, with a mean sentiment of 0.09 compared to 2016 (mean 0.24) and 2017 (mean 0.14) and even dropped to –0.11 in 2019. The hashtag #bioethics only appeared in 2015 and was associated with a relatively low sentiment of –0.02. This may highlight the various ethical concerns raised during the 2015 Human Gene Editing summit. See Multimedia Appendix B.6 for the full list of the counts, sentiments, and standard deviations of the most used hashtags by year.

### 4.3.5 Themes

In comparison to the hashtags, the themes derived from previous studies can relate the Twitter discussion to known themes of interest to the public (see the Methods section for a description of the analysis). The 6 themes that were matched most are presented in Figure 4.5 and grouped by positive, neutral, and negative sentiments. The themes include genome (with a total count of 526,612 [extrapolated for 2019]), baby (68,269), disease (64,181), embryo (49,084), treatment (35,865), and mutation (34,884). Unsurprisingly, the theme “genome” was matched most frequently, occurring in 34% of the tweets.

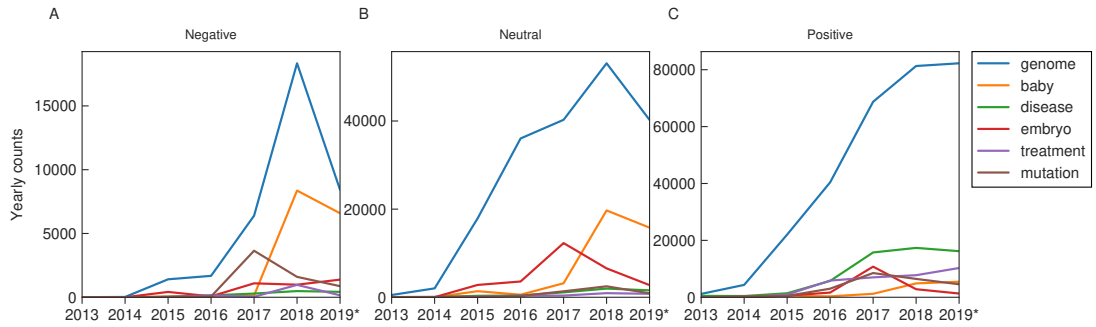



Figure 4.5: Yearly occurrences of themes. Multiple themes with distinct regex patterns were matched to the text of tweets, and the 6 most frequent themes were selected. Panels **A**, **B**, and **C** show the yearly counts of themes when grouped by negative, neutral, and positive sentiment, respectively. For the year 2019, the counts were extrapolated from the months before June to a full year.

The reported themes show distinct occurrence patterns depending on sentiment, yielding an aggregated picture of the discussion surrounding CRISPR throughout the years. Spikes are evident in certain years (see Multimedia Appendix B.7 for the counts per year of the top 6 themes), and the most significant change in occurrences happened for the theme “baby”, which increased substantially from 2017 to 2018, likely associated with the “CRISPR babies” scandal in November 2018. While a spike could be observed for all 3 sentiments, the increase was far more pronounced in the neutral and negative classes (see Figure 4.5). The theme “mutation” shows a negative peak in 2017, when risks about potential side effects of CRISPR surfaced. Relative to other themes, the themes “disease” and “treatment” were major themes in a discussion associated with a positive sentiment.

## 4.4 Discussion

### 4.4.1 Principal findings

We have generated the first high-resolution temporal signal for sentiments towards CRISPR on Twitter, spanning a duration of more than 6 years. Our results suggest that, overall, the CRISPR technology was discussed in a positive light, which aligns well with a previous study that considered the coverage of CRISPR in the press (Marcon et al. 2019). However, more recently, the sentiment reveals a series of strong negative dips, pointing to a more critical view. The frequency and magnitude of these dips have increased since 2017, which is underlined by the overall declining sentiment. It is noteworthy that the dips usually coincide with high activity, suggesting that many people are only exposed to the topic of CRISPR when it is presented in an unfavorable way.

Further, we could tie the most prominent peaks in tweeting activity to real world events. The last 3 peaks, which coincide with the release of possibly concerning news (side effects, CRISPR babies, malware), also align with strong dips in the tweet sentiment. Together, this indicates that there is at least a partial connection between tweets and the discourse off Twitter and that the sentiment changes are not only the result of a self-contained discussion on the social media platform. Even more so, the peak detection potentially allows the timely identification of significant incidents that can shape public discourse and opinion.

As shown in the breakdown of sentiment by organism, the negative sentiment was stronger in the embryo and human classes but stayed mostly positive towards other organisms. The data therefore suggest that the many ethical issues related to human germline editing are reflected in the tweets. However, criticism may not be targeted at the use of CRISPR in humans per se: Hashtags such as #hiv or #genetherapy were connected to very positive sentiments, which suggests a positive attitude towards developing CRISPR for use in medical treatment. This aspect is further strengthened when considering the sentiment of themes

such as “treatment” or “disease”. These observations are in line with several surveys in which participants demonstrated strong support of CRISPR for use in medical treatment but were critical regarding modifications of human germline cells (Weisberg, Badgio, and Chatterjee 2017; McCaughey, Sanfilippo, et al. 2016; Gaskell et al. 2017; Hendriks et al. 2018; McCaughey, Budden, et al. 2019).

The dataset that includes continuous observations over a long period of time allows for conclusions to be drawn about the public perception of CRISPR both on short and long time scales. For example, when the article on biohacking re-emerged in 2019 (peak f), shortly after the discussions around CRISPR babies, it was discussed in significantly more negative terms than at the time of its publication in 2017. Therefore, the intermediate developments seem to have had a negative influence on the perception of the event. This is in line with the overall negative trend. The presence and absence of themes observed in the data hint at the influence that key events might have on the discussion. While the theme “mutation” was discussed intensely in 2017, its occurrence in tweets dropped in the following year, 2018, in which “baby” became the most occurring theme except for “genome”.

Our results support the use of Twitter and similar platforms for the study of public discourse. Discussion about a subject matter can be investigated in real-time, in depth at the level of individual statements, and on the basis of existing data. The insights gained through such studies can bring new issues to light, indicate which topics need extra attention with respect to ethical considerations and policy making, and allow a quicker response to technological advancements. In addition, the presented method offers a novel approach to promote public engagement, especially in the areas of biotechnologies and health care, as argued by the Nuffield Council on Bioethics (Bioethics 2012).

### 4.4.2 Limitations

Although the predicted sentiment index seems to overlap well with survey results, it cannot be directly used as a substitute for an opinion poll. Polling allows for the collection of answers to specific questions of interest instead of inferring them from public statements. Furthermore, the Twitter community is not necessarily representative of the whole population of a country. However, sentiment analysis avoids the disadvantages of traditional methods such as response bias and provides more detailed insights through access to granular data of online discussions.

We cannot exclude the possibility that the gradual decrease over time was influenced or caused by a general shift in the sentiment of the scientific Twitter community. Our analysis relies only on Twitter, and we did not validate the findings on another social media platform. Also, we cannot directly tie the sentiment in tweets to the conversation off Twitter. Nonetheless, our results show that there is a connection between tweets, findings in earlier studies, and

real-world events and that insights can be gained from this type of analysis on Twitter that are not accessible through other methods.

Further, we acknowledge that most people's opinions might not fit into the positive, neutral, and negative classes presented in this study. We therefore tried to counteract this problem by categorizing the data not only by sentiment but also by relevance and organism, allowing for a better understanding of the measured sentiment. Furthermore, we recognize the challenging nature of deducing someone's true opinion based on a short message alone and the fact that it is only possible within a statistical margin of error. This error is slightly larger for the negative class, as the F1 score of this class was relatively low compared to the other classes due to a strong label imbalance. We believe, however, that our method is nevertheless suitable to capture certain trends on a larger scale.

### 4.4.3 Conclusions and Future Direction

We demonstrated that the sentiment analysis of tweets provides a high-resolution picture of the ongoing debate on CRISPR, allowing us to study the evolution of the discourse while extending the capacity of traditional methods. Further, the presence of the same themes that have been identified in existing studies confirms the validity of our signal with respect to content. The existence of events that match the activity peaks also indicates the sensitivity of the signal towards off-Twitter incidents. Therefore, our approach offers an additional method to surveys and that can be deployed to get richer information, a larger sample size, and higher temporal resolution.

Future work can go beyond the deduction of sentiments and shed more light on the nature of discussions and arguments raised and how they influence each other, giving a better idea of the reasoning behind people's opinions. Furthermore, specific topics, such as the discussion surrounding a potential moratorium of CRISPR, may be analyzed in more detail and provide actionable outcomes.

Since the presented analysis can automatically process a large amount of data in almost real-time, it extends the traditional toolset of empirical methods for discourse analysis. It may therefore help analyze public opinion and support policy and decision making.

### Data and Code Availability

The data, machine learning models used, and source code for this analysis can be found in our public repository: `https://gitlab.ethz.ch/digitalbioethics/crispr-sentiment-analysis`.

### Acknowledgments

MM and MSch designed the experiment, performed the analysis, and wrote the paper in equal parts. MS and EV initiated the work, guided the experimental design, and made corrections to the paper. MS wrote the abstract. We thank Agata Ferretti for the support during the initial literature review and Ellen Lapper for proofreading.

### Conflicts of Interest

None declared.

## References

Rosenberg, Steven A et al. (1990). "Gene transfer into humans—immunotherapy of patients with advanced melanoma, using tumor-infiltrating lymphocytes modified by retroviral gene transduction". In: *New England Journal of Medicine* 323.9, pp. 570–578.

Comai, L et al. (1985). "Expression in plants of a mutant aroA gene from Salmonella typhimurium confers tolerance to glyphosate". In: *Nature* 317.6039, pp. 741–744.

Johnson, Irving S (1983). "Human insulin from recombinant DNA technology". In: *Science* 219.4585, pp. 632–637.

Ye, Xudong et al. (2000). "Engineering the provitamin A ($\beta$-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm". In: *Science* 287.5451, pp. 303–305.

Thomas, Kirk R, Kim R Folger, and Mario R Capecchi (1986). "High frequency targeting of genes to specific sites in the mammalian genome". In: *Cell* 44.3, pp. 419–428.

Choulika, Andre et al. (1995). "Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae." In: *Molecular and cellular biology* 15.4, pp. 1968–1973.

Jasin, Maria (1996). "Genetic manipulation of genomes with rare-cutting endonucleases". In: *Trends in Genetics* 12.6, pp. 224–228.

Porteus, Matthew and David Baltimore (2003). "Chimeric nucleases stimulate gene targeting in human cells.(Brevia)". In: *Science* 300.5620, pp. 763–764.

Barrangou, Rodolphe et al. (2007). "CRISPR provides acquired resistance against viruses in prokaryotes". In: *Science* 315.5819, pp. 1709–1712.

Jinek, Martin et al. (2012). "A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity". In: *science* 337.6096, pp. 816–821.

Cong, Le et al. (2013). "Multiplex genome engineering using CRISPR/Cas systems". In: *Science* 339.6121, pp. 819–823.

Mali, Prashant et al. (2013). "RNA-guided human genome engineering via Cas9". In: *Science* 339.6121, pp. 823–826.

Liang, Puping et al. (2015). "CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes". In: *Protein & cell* 6.5, pp. 363–372.

Caplan, Arthur L et al. (2015). "No time to waste—the ethical challenges created by CRISPR: CRISPR/Cas, being an efficient, simple, and cheap technology to edit the genome of any organism, raises many ethical and regulatory issues beyond the use to manipulate human germ line cells". In: *EMBO reports* 16.11, pp. 1421–1426.

Bosley, Katrine S et al. (2015). "CRISPR germline engineering—the community speaks". In: *Nature biotechnology* 33.5, pp. 478–486.

Regalado, Antonio (2018). "Chinese scientists are creating CRISPR babies". In: *MIT Technology Review*. URL: https://www.technologyreview.com/s/612458/exclusive-chinese-scientists-are-creating-crispr-babies/.

Cyranoski, David and Heidi Ledford (2018). "Genome-edited baby claim provokes international outcry". In: *Nature* 563.7733, pp. 607–608.

Normile, Dennis (2018). *Shock greets claim of CRISPR-edited babies*.

Baltimore, David et al. (2015). "A prudent path forward for genomic engineering and germline gene modification". In: *Science*, aab1028. DOI: 10.1126/science.aab1028.

Lander, Eric S (2015). "Brave new genome". In: *New England Journal of Medicine* 373.1, pp. 5–8. DOI: 10.1056/NEJMp1506446.

Lander, Eric S et al. (2019). *Adopt a moratorium on heritable genome editing*. DOI: 10.1038/d41586-019-00726-5.

Legge Jr, Jerome S and Robert F Durant (2010). "Public opinion, risk assessment, and biotechnology: Lessons from attitudes toward genetically modified foods in the European Union". In: *Review of Policy Research* 27.1, pp. 59–76. DOI: 10.1111/j.1541-1338.2009.00427.x.

Paola Ferretti, Maria (2007). "Why public participation in risk regulation? The case of authorizing GMO products in the European Union". In: *Science as Culture* 16.4, pp. 377–395. DOI: 10.1080/09505430701706723.

Marris, Claire (2001). "Public views on GMOs: deconstructing the myths: Stakeholders in the GMO debate often describe public opinion as irrational. But do they really understand the public?" In: *EMBO reports* 2.7, pp. 545–548. DOI: 10.1093/embo-reports/kve142.

Geller, Gail, Barbara A Bernhardt, and Neil A Holtzman (2002). "The media and public reaction to genetic research". In: *JAMA* 287.6, pp. 773–773. DOI: 10.1001/jama.287.6.773-JMS0213-3-1.

Travis, John (2015). "Inside the summit on human gene editing: a reporter's notebook". In: *Science* 10. DOI: 10.1126/science.aad7532.

National Academies of Sciences, Engineering, and Medicine and others (2017). *Human genome editing: science, ethics, and governance*. National Academies Press. DOI: 10.17226/24623.

Weisberg, Steven M, Daniel Badgio, and Anjan Chatterjee (2017). "A CRISPR New World: Attitudes in the Public toward Innovations in Human Genetic Modification". In: *Frontiers in public health* 5, p. 117. DOI: 10.3389/fpubh.2017.00117.

McCaughey, Tristan, Paul G Sanfilippo, et al. (2016). "A global social media survey of attitudes to human genome editing". In: *Cell stem cell* 18.5, pp. 569–572. DOI: 10.1016/j.stem.2016.04.011.

Gaskell, George et al. (2017). "Public views on gene editing and its uses". In: *Nature biotechnology* 35.11, p. 1021. DOI: 10.1038/nbt.3958.

Hendriks, S et al. (2018). "Reasons for being in favour of or against genome modification: a survey of the Dutch general public". In: *Human Reproduction Open* 2018.3, hoy008. DOI: 10.1093/hropen/hoy008.

McCaughey, Tristan, David M Budden, et al. (2019). "A need for better understanding is the major determinant for public perceptions of human gene editing". In: *Human gene therapy* 30.1, pp. 36–43. DOI: 10.1089/hum.2018.033.

Marcon, Alessandro et al. (2019). "CRISPR in the North American popular press". In: *Genetics in Medicine*, p. 1. DOI: 10.1038/s41436-019-0482-5.

Liu, Bing (2012). "Sentiment analysis and opinion mining". In: *Synthesis lectures on human language technologies* 5.1, pp. 1–167.

Ravi, Kumar and Vadlamani Ravi (2015). "A survey on opinion mining and sentiment analysis: tasks, approaches and applications". In: *Knowledge-Based Systems* 89, pp. 14–46.

Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore (2011). "Twitter sentiment analysis: The good the bad and the omg!" In: *Fifth International AAAI conference on weblogs and social media*. Citeseer.

Severyn, Aliaksei and Alessandro Moschitti (2015). "Twitter sentiment analysis with deep convolutional neural networks". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 959–962.

Müller, Martin M and Marcel Salathé (2019). "Crowdbreaks: Tracking health trends using public social media data and crowdsourcing". In: *Frontiers in public health* 7, p. 81.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30, pp. 5998–6008.

Sun, Chi, Luyao Huang, and Xipeng Qiu (2019). "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence". In: *arXiv preprint arXiv:1903.09588*.

Basile, Valerio et al. (2019). "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter". In: *13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, pp. 54–63.

Zampieri, Marcos et al. (2019). "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)". In: *arXiv preprint arXiv:1903.08983*.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

McKinney, Wes et al. (2010). "Data structures for statistical computing in python". In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. Austin, TX, pp. 51–56.

Fleiss, Joseph L and Jacob Cohen (1973). "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability". In: *Educational and psychological measurement* 33.3, pp. 613–619. DOI: 10.1177/001316447303300309.

Pagliardini, Matteo, Prakhar Gupta, and Martin Jaggi (2017). "Unsupervised learning of sentence embeddings using compositional n-gram features". In: *arXiv preprint arXiv:1703.02507*. DOI: 10.18653/v1/n18-1049.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297. DOI: 10.1007/BF00994018.

Joulin, Armand et al. (2016). "Bag of tricks for efficient text classification". In: *arXiv preprint arXiv:1607.01759*.

Wolf, Thomas et al. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing". In: *ArXiv*, arXiv–1910.

Paszke, Adam et al. (2019). "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in neural information processing systems*, pp. 8026–8037.

Ruxton, Graeme D (2006). "The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test". In: *Behavioral Ecology* 17.4, pp. 688–690.

Alhabash, Saleem et al. (2018). "140 characters of intoxication: Exploring the prevalence of alcohol-related tweets and predicting their virality". In: *Sage open* 8.4, p. 2158244018803137.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001). *SciPy: Open source scientific tools for Python*. URL: https://www.scipy.org/.

Ma, Hong et al. (2017). "Correction of a pathogenic gene mutation in human embryos". In: *Nature* 548.7668, p. 413. DOI: 10.1038/nature23305.

Kosicki, Michael, Kärt Tomberg, and Allan Bradley (2018). "Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements". In: *Nature biotechnology* 36.8, p. 765. DOI: 10.1038/nbt.4192.

Greenberg, Andy (2017). "Biohackers encoded malware in a strand of DNA". In: *Wired*. URL: https://www.wired.com/story/malware-dna-hack.

Bioethics, Nuffield Council on, ed. (2012). *Emerging biotechnologies: technology, choice and the public good*. ISBN:978-1-904384-27-4. London: Nuffield Council on Bioethics. ISBN: 978-1-904384-27-4.

# 5 Experts and authorities receive disproportionate attention on Twitter during the COVID-19 crisis

**Published as:**
Gligorić K., Horta Ribeiro M., Müller M., Altunina O., Peyrard M., Salathé M., Colavizza G., and West R. (2020). Experts and authorities receive disproportionate attention on Twitter during the COVID-19 crisis *arXiv preprint* arxiv:2008.08364

**Abstract**

Timely access to accurate information is crucial during the COVID-19 pandemic. Prompted by key stakeholders' cautioning against an "infodemic", we study information sharing on Twitter from January through May 2020. We observe an overall surge in the volume of general as well as COVID-19-related tweets around peak lockdown in March/April 2020. With respect to engagement (retweets and likes), accounts related to healthcare, science, government and politics received by far the largest boosts, whereas accounts related to religion and sports saw a relative decrease in engagement. While the threat of an "infodemic" remains, our results show that social media also provide a platform for experts and public authorities to be widely heard during a global crisis.

## 5.1 Introduction

Social media, in particular Twitter (David Andre Broniatowski, Paul, and Dredze 2014), plays a central role in emergency response and has proven to be effective for monitoring ongoing crises (Reuter and Kaufhold 2018; Wagner et al. 2018; Shin et al. 2016; Salathé et al. 2013). In addition, Twitter has also proven to be essential for overcoming public health crises by facilitating access to trustworthy information, especially when the coordinated effort of entire populations was required (H. W. Park, S. Park, and Chong 2020). At the same time, the widespread adoption of social media has been linked to the propagation of low-quality information, mis-information, and disinformation (D. M. Lazer et al. 2018; Swire-Thompson and D. Lazer 2020), with some studies concluding that such information goes viral more easily, and has broader reach, than trustworthy information on social media (Vosoughi, Roy, and Aral 2018; Del Vicario et al. 2016). These findings are particularly pertinent to the COVID-19 crisis, which is unfolding during a time of unprecedented Internet penetration, drawing enormous attention on both traditional and social media (H. W. Park, S. Park, and Chong 2020; Depoux et al. 2020; Chen, Lerman, and Ferrara 2020; Yang, Torres-Lugo, and Menczer 2020; Li et al. 2020; Alshaabi et al. 2020; Vicari and Murru 2020). This fact has led the World Health Organization to declare a state of "infodemic" (WHO 2020a), urging that "people must have access to accurate information to protect themselves and others" (WHO 2020b).

When considering global health crises in the past, experts and public institutions are considered to be trusted information sources (Dutta-Bergman 2003). Their visibility might, however, be diminished by the spread of low-quality or false information in social media (Reuter and Kaufhold 2018). This phenomenon has been studied to some degree in the context of the 2015-2016 Zika outbreak, where misinformation and pseudo-scientific claims surged in parallel with growing media attention (Dredze, David A Broniatowski, and Hilyard 2016; Hossain et al. 2020) and, as a study on Facebook data revealed, received the most public engagement (Sharma et al. 2017). Nevertheless, the role of experts was deemed crucial for correcting misinformation (Vraga and Bode 2017), and key international experts on Twitter were able to facilitate the flow of accurate and vital information (Hagen et al. 2018).

The role of experts during the ongoing COVID-19 pandemic is still unclear. Initial work has mostly focused on the role of key government officials. A survey-based study found that government officials, such as Dr. Anthony Fauci, a lead member of the White House Coronavirus task force, have been most effective at advocating for social distancing measures, when compared to other influential figures (Abu-Akel, Spitz, and West 2020). A content analysis of viral COVID-19-related tweets by world leaders revealed that the majority of tweets were informative and only a minority were political (Rufai and Bunce 2020). A number of studies paint a dire picture of abundant misinformation online (H. W. Park, S. Park, and Chong 2020; Cinelli et al. 2020; Ahmed et al. 2020; Ferrara 2020; Kouzy et al. 2020), raising valid

concerns over whether opinions by scientific experts are being heard, especially as the crisis continues. A preliminary study suggested that, while false information was tweeted more than science-backed information, the latter was shared more via retweets (Pulido et al. 2020). Therefore, it is of paramount importance to accurately quantify who was and is speaking and who was and is successful in being heard. A study based on a selection of Twitter users that could be matched to US voter registration records finds that between January and June 2020 journalists, media outlets, and political accounts have consistently received high attention, whereas epidemiologists and public health professionals only made up for a small fraction of the most retweeted users, possibly due to a smaller average follower base (Gallagher et al. 2020). It is noteworthy, however, that this study only considered the unique situation in the US, which is currently in the midst of a heated election race.

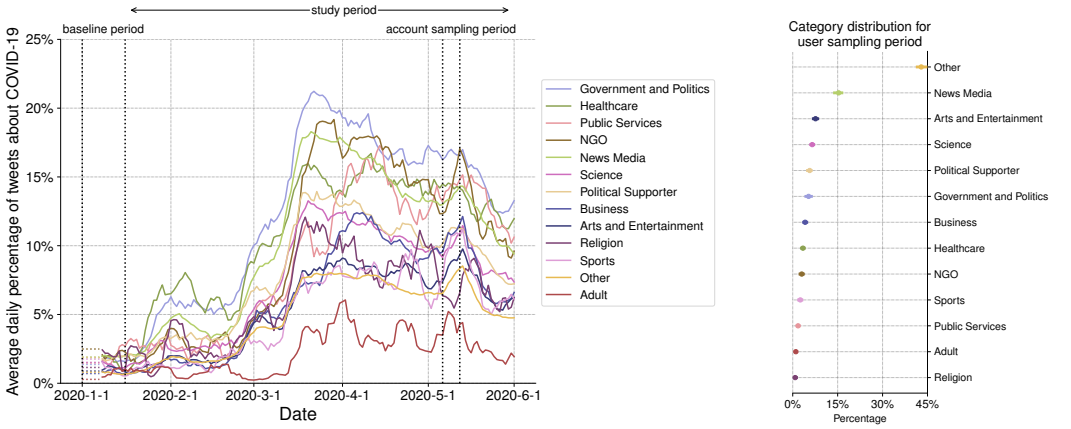
The same work identifies three important shortcomings in the current literature, which have made it difficult to fully address the role of experts during the COVID-19 crisis: a) Most work is conducted on an incomplete samples of tweets, which limits the results' generality; b) current studies are usually US-centric, although it is known that the debate surrounding COVID-19 was and is geographically very diverse (Hernández-García and Giménez-Júlvez 2020); b) a solid methodology for identifying "expert" accounts is still missing, a crucial component in order to properly answer the question of the role of experts on Twitter.

In order to address these outstanding limitations, we leverage the *complete stream of COVID-19-related tweets*, made available to the authors by Twitter for academic research, and perform a *large-scale human annotation of user accounts* to accurately quantify how much attention experts are receiving during the pandemic.

## 5.2 Results

The overall goal of the present research is to map the Twitter landscape during the COVID-19 pandemic from an account-centric angle: who is speaking and who is being heard? To better understand the plurality of voices taking part in the public debate on Twitter, we developed a custom taxonomy of categories of user accounts (see legend of Fig. 5.1A for a list of categories; see methods for details on how the taxonomy was derived). We employed Twitter's complete COVID-19 streaming endpoint, to which access was granted starting 6 May 2020. The stream includes all tweets containing one of 590 multilingual keywords related to COVID-19. The population we study consists of all user accounts that posted COVID-19-related content that has received a non-negligible amount of attention. Fig. 5.1A summarizes the specific study design that was implemented. Based on the first full week of the COVID-19 stream (6–12 May 2020; "account sampling period" in Fig. 5.1A), we constructed a sample of 14,200 Twitter accounts that each had posted at least one COVID-19-related tweet with at least 10 retweets or likes (henceforth, "engagements"), and annotated each account in the sample with its category

using crowdsourcing (see methods). The sample was constructed to be representative of the overall population. The distribution over account categories is plotted in Fig. 5.1B. We then queried Twitter's application programming interface (API) to collect all tweets —— regardless of whether they contained a COVID-19 keyword —— for the 14,200 sampled accounts during the 5-month period from 1 January to 31 May 2020. The first 2 weeks were used as a "baseline period" to calibrate accounts' behavior, which was tracked during the following 4.5 months ("study period" in Fig. 5.1A).




(a) (b)

Figure 5.1: **Study design.** (A) We study Twitter accounts that posted at least one COVID-19-related tweet that received at least 10 retweets + likes during the week of 6–12 May 2020 (account sampling period, shaded gray). We create a sample of these accounts, categorize them into 13 categories (cf. legend), and collect their entire Twitter timelines from 1 January to 31 May 2020. The first 14 days serve as a baseline period (shaded green), and the remaining 4.5 months, as the study period. Inverse probability weighting (see methods) is used throughout all analyses to make the sample representative. Lines in (A) represent the percentage of tweets related to COVID-19 per category for the sampled accounts (7-day moving averages; giving every account equal weight). Starting March 2020, a substantial fraction of tweets refers to the pandemic. (B) Distribution over categories among all accounts in the account sampling period, estimated from the manually categorized, representative sample.

First, in Fig. 5.1A, we track the fraction of tweets containing a COVID-19 keyword, macro-averaged over all accounts per category. Whereas all categories posted very small fractions (0 %–2.5 %) of COVID-19-related content during the baseline period in early January 2020, the topic became more prevalent in late January and peaked between mid March and early April, when up to 1 in 5 tweets contained a COVID-19 keyword for some categories, with the highest peaks observed for categories of direct relevance for the pandemic: Government & Politics (peak 21%), NGOs (19%), News Media (18%), Public Services (17%), and Healthcare (17%). Less directly relevant categories also referred to COVID-19 in considerable fractions of their posts, e.g., Religion (12%), Sports (11%), and Arts & Entertainment (10%). This first result highlights the deep impact the COVID-19 pandemic has had on the Twitter ecosystem.

Next, we investigate whether the studied accounts have changed their overall tweeting frequency during the COVID-19 pandemic. This analysis considers all tweets posted by the studied accounts, regardless of whether they contain a COVID-19 keyword or not. We calibrated an account's tweet volume during the baseline period and computed, for each subsequent week, the percentage change over the baseline. The results, visualized as blue curves in Fig. 5.2, show that tweet volume increased considerably for all categories, compared to the pre-pandemic baseline. The most notable cases are Religion, which peaked at +207%, and Healthcare, at +175%. Even the least affected categories showed a strong increase, with News Media peaking at +63%, and Arts & Entertainment, at +73%.

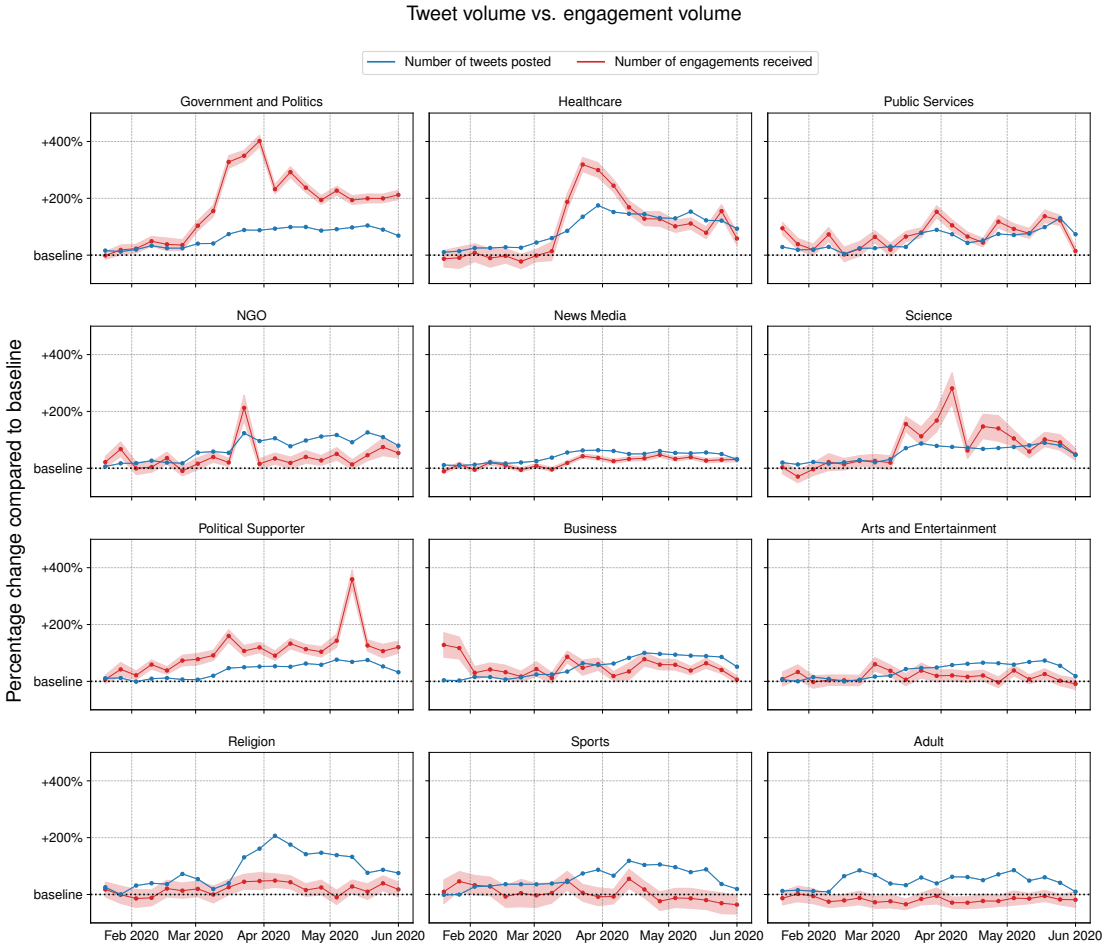



Figure 5.2: **Tweet volume vs. engagement volume.** Weekly percentage increase over the early-January baseline (cf. Fig. 5.1A) with respect to the number of tweets posted (blue) and engagements (retweets + likes) received (red) (macro-averages over accounts; estimated from the representative sample of accounts, cf. Fig. 5.1A; with 95% confidence intervals). In all categories, tweet volumes (blue) rise far above baseline, particularly starting mid March 2020, when lockdowns were imposed worldwide. Engagement (red) behaves more heterogeneously, rising more for experts and authority categories (i.e. health, science, government, politics). Red lying above [below] blue corresponds to a rate of engagements-per-tweet that is higher [lower] than at baseline. That is, Government & Politics and Political Supporters see lasting, whereas Health and Science see transient, boosts in engagements-per-tweet.

In order to determine to what extent the increase in tweeting is associated with an increase in being noticed, we also measured the engagement (retweets + likes) received by each account, again calibrated against the pre-pandemic baseline. The results, visualized as red curves in Fig. 5.2, show that engagement volume behaved more heterogeneously than tweet volume. Some categories saw substantial increases in engagement —– much larger than the respective increases in tweet volume (red above blue in Fig. 5.2). In particular, for Government & Politics, the increase in engagement peaked at +402%, whereas the increase in tweet volume peaked at only +102%. Similar effects were observed for Healthcare (+319% vs. +175%), Science (+281% vs. +89%), and Political Supporters (+359% vs. +76%). Accounts in these categories thus became, on average, more "effective" at tweeting, with a higher number of engagements per tweet than at baseline. Conversely, for other categories, engagement per tweet decreased with the pandemic (red below blue in Fig. 5.2). Most notably, Religion saw little increase in engagement (peak +49%), despite having increased its tweet volume most out of all the categories (peak +207%). Similar effects were observed for Sports (+55% vs. 119%) and Adult content (+2% vs. +86%). It is noteworthy that, among the "effective" categories, two distinct patterns emerge: on the one hand, for Healthcare and Science, the blue curve in Fig. 5.2 converges to nearly the same value as the red curve (i.e., engagement per tweet reverts to the level of the baseline period), whereas, on the other hand, for Government & Politics and Political Supporters, the red curve remains consistently above the blue curve (i.e., engagement per tweet stays above the level of the baseline period). To summarize, Healthcare and Science saw transient, whilst Government & Politics and Political Supporters saw persistent, boosts in engagement that far exceeded the respective boosts in tweet volumes. On the contrary, Religion, Sports, and Adult content accounts saw a decrease in engagement, despite the fact that they, too, tweeted more.

To directly compare categories to each other, we computed two global rankings of accounts (both computed 1–4 June 2020, when account timelines were collected), one with respect to engagement counts, the other with respect to follower counts. Average ranks (normalized such that 1 and 0 correspond to top and bottom, respectively) are plotted for all categories in Fig. 5.3. We will discuss the follower-count ranking (x-axis) later, and for now focus on the engagement ranking (y-axis). Average engagement ranks were significantly ($p < 0.05$, two-sided KS tests) higher for tweets from the study period (end points of arrows) than for tweets from the baseline period (starting points of arrows) for Healthcare, Science, Government & Politics, Political Supporters, Public Services, and News Media, whereas the effect was reversed for Religion, Sports, Adult content, and Business. While these results echo the findings from Fig. 5.2, they also add nuance: as all accounts participated in the rank computations, Fig. 5.3 may be considered a "zero-sum game", in the sense that one account's increase must be offset by another account's decrease. Viewed in this light, Fig. 5.3 suggests that Healthcare, Science, Government & Politics, etc., have gained attention relative to Religion, Sports, and Adult content.

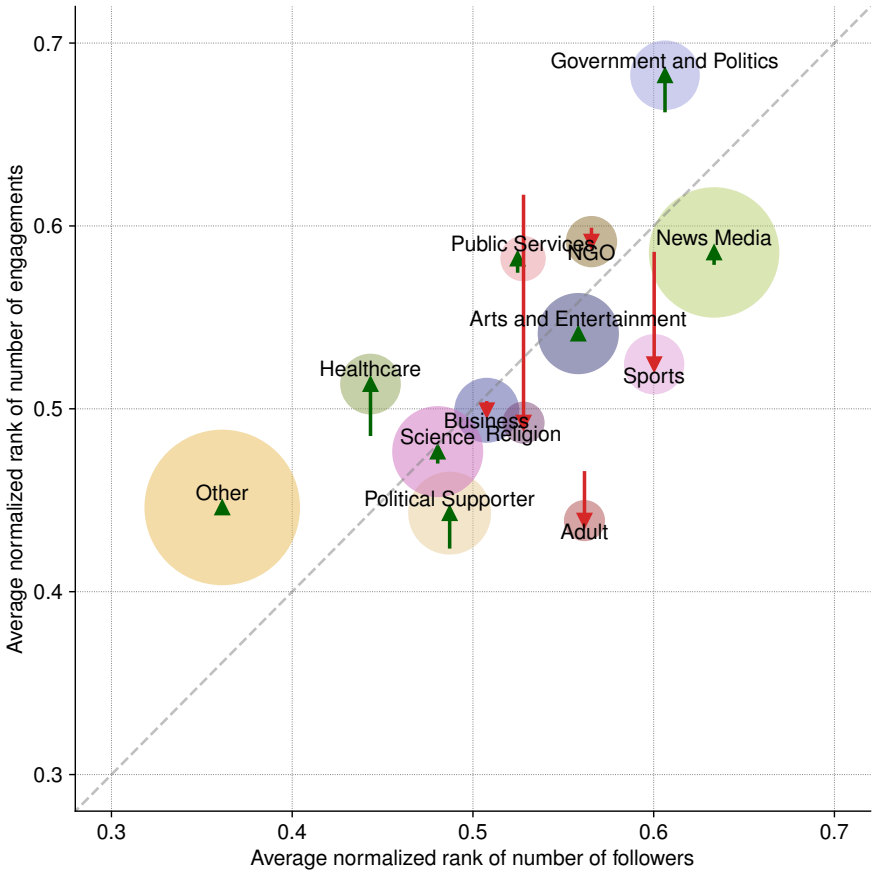



Figure 5.3: **Rank-based comparison of account categories.** The y-axis shows normalized ranks with respect to the number of engagements (retweets + likes) received for tweets posted during the baseline period (arrow starting points) and for tweets posted during the study period (arrow end points), averaged over the accounts in the respective category. The x-axis shows ranks with respect to follower counts (as observed after the end of the study period, 1–4 June 2020). Ranks were normalized such that 1 and 0 correspond to top and bottom, respectively. Disk radius is proportional to the number of tweets posted by the category in the study period. Categories linked to experts and authorities (i.e. health, science, government, politics, news) have risen (upward arrows), whereas Religion, Sports, and Adult have fallen (downward arrows). Healthcare, Government & Politics, Public Services, and NGOs are particularly much engaged-with, relative to their follower counts (position above diagonal).

Follower counts on Twitter vary widely across accounts (Cha et al. 2010). The intuitive expectation that a larger follower count is associated with more engagement is overall confirmed by Fig. 5.3, with a category-level Spearman rank correlation of 0.71 ($p = 0.0067$, $t(11) = 3.33$) in the baseline period, and 0.62 ($p = 0.024$, $t(11) = 2.62$) in the study period. Some important exceptions, however, emerge: Healthcare accounts on average rank lowest with respect to follower count during the study period (12 out of 12 when ignoring the "Other" category), but rank in the upper half (6 out of 12) with respect to engagement. The opposite effect is observed for Sports, Arts & Entertainment, and Adult content, which are in the top half with respect to follower count, but in the bottom half with respect to engagement. These findings suggest that the increased attention to categories that are most directly important in the fight against the pandemic is not merely a consequence of the size of their follower base.

## 5.3 Discussion

A large body of work has focused on the rampant misinformation present on social media during the COVID-19 crisis. A particular focus has been given to Twitter, as it is widely used by public health officials to reach out to the public and inform citizens in rapidly evolving crisis situations. In this work, we have taken a comprehensive approach to quantifying the global attention given to experts on Twitter during the first five month of the COVID-19 pandemic.

We have shown that Twitter accounts associated with scientific experts and public authorities are boosted during the pandemic. While accounts in all categories on average increase their tweet volume, accounts related to Science, Healthcare, Government and Politics receive the largest boosts in engagement. We also found that the ways in which accounts belonging to experts and authorities are boosted seem to differ. As the crisis broadened from a health crisis to a societal crisis, accounts related to healthcare and science received progressively less attention, whereas attention to governments and politicians remained high. This finding might point to the increasing relevance of politics and the economic consequences of the pandemic.

On the one hand, our work confirms previous literature on the topic, which has shown that Twitter users are amplifying relevant content during crisis situations (Reuter and Kaufhold 2018; Wagner et al. 2018; Shin et al. 2016). On the other hand, however, preliminary work in the context of COVID-19 suggested that media outlets were more strongly amplified than scientists or health experts on Twitter (Gallagher et al. 2020), which we do not confirm. There could be several reasons for this disagreement. Most of all, our work gives a global picture, whereas previous work focused on the US situation, which might indeed be very different from the rest of the world. This is also confirmed by the large differences observed across languages (cf. Figure C.5). Furthermore, the above-cited analysis (ibid.) only considers users with real names in their profile, who could be matched to US voter registration records.

A caveat of our analysis is that it is based on self-declared account descriptions, thus we were not able to link a sizable fraction to a clear category. Certain groups of users might be less inclined to publicly mention their professional status. Furthermore, our analysis does not take into consideration the actual content of the messages. This is an important direction for future work, since the quality of specific messages and the alignment with the scientific consensus can vary within categories.

In conclusion, our work offers a more general overview on the current online debate on COVID-19, by providing a complete and global picture of attention patterns on Twitter. While we do not downplay the issues surrounding misinformation, our main result is encouraging since we show that, ultimately, Twitter users are paying disproportionate attention to experts and authorities during the COVID-19 crisis.

## 5.4 Methods

### Description of the data

Fig. 5.4 presents a diagram with all the original and derived data sources used. We employ Twitter's complete COVID-19 streaming endpoint,[I] which was made available to researchers upon request (Dataset **A**). The endpoint includes all tweets containing one of several multilingual keywords – curated by Twitter – related to COVID-19, as well as all retweets and replies to those tweets.[II] We focus on COVID-19 tweets posted during the week from May 6 to May 12, 2020, written in ten major languages: English, Japanese, Spanish, Portuguese, French, German, Italian, Arabic, Indonesian and Hindi. The language of a tweet is detected by Twitter and obtained directly from the tweet object. We limit our analysis to accounts that tweeted at least one popular COVID tweet during the week of sampling (i.e. a tweet that received at least 10 retweets). We then perform sampling and annotation according to our taxonomy, to produce Dataset **B**: the annotated sample. Next, we get the timelines for all such accounts in the collected sample, collecting all the tweets they posted in 2020, and we study all of their tweets posted between Jan 1 and May 31, 2020 (Dataset **C**). To do so, we employ Twitter's API[III] (for accounts with less than 3200 tweets between 01/01/2020 and 31/05/2020), and Twint[IV], a crawler that uses a Web UI for scraping (for accounts with more than 3200 tweets in this time frame). Additionally, we leverage the annotated sample to train a machine learning classifier which is used to expand the labels by classifying the remaining accounts in the entire week

[I] Announced at: `https://blog.twitter.com/developer/en_us/topics/tools/2020/covid19_public_conversation_data.html`

[II] The complete list of keywords is available at: `https://developer.twitter.com/en/docs/labs/covid19-stream/overview`

[III] `https://developer.twitter.com/en/docs/tweets/timelines/overview`

[IV] `https://github.com/twintproject/twint`

for the COVID-19 stream, to produce Dataset **D**. Overall, we start our analysis from 467.36k tweets that received at least 10 retweets, posted during the seven-day account sampling period in May, by 196.95k unique accounts (Dataset **A**). After sampling (Dataset **B**) and enriching the timelines, dataset **C** consists of 11.47M tweets (736.73k out which contain a COVID-19 keyword, using the list of COVID-19 keywords curated by Twitter). In our analyses, we calibrate an account's tweet volume and engagement during the baseline period and compute, for each subsequent week, the percentage change over the baseline. To account for the possibility that some days of the week (Monday, Tuesday, etc.) might generally see higher tweet volumes, calibration is done by the day of the week, for the day-level analyses.

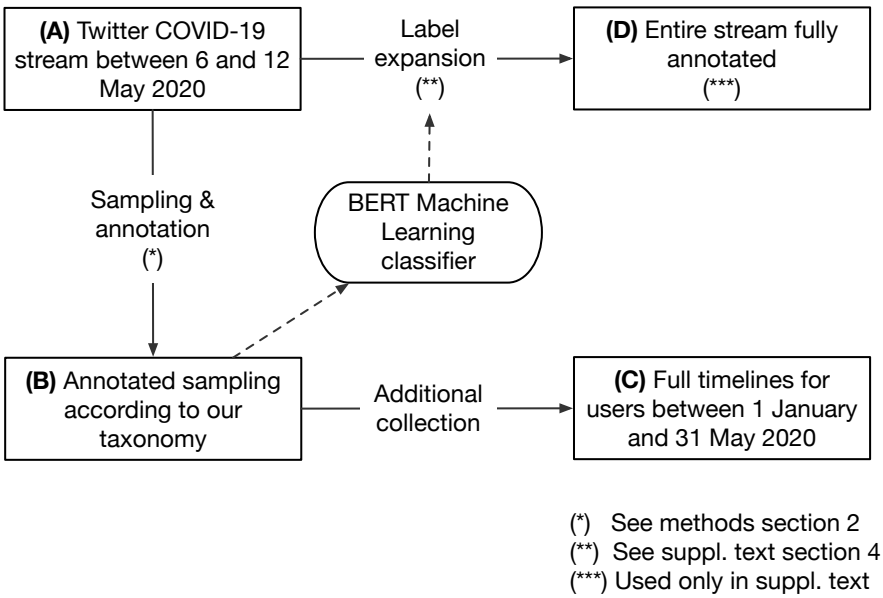



Figure 5.4: Diagram with the original and derived datasets we used. **(A)** Our original data source consists of all tweets from the COVID-19 stream between the 6th and the 12th of May 2020. **(B)** We then sample a fraction of these accounts and annotate them according to the taxonomy we developed. **(C)** For the annotated accounts, we additionally collect their entire timelines between the 1st of January to the 31st of May 2020. **(D)** Lastly, we leverage the annotated sample to train a machine learning classifier which is used to classify the remaining accounts in the entire week for the COVID-19 stream.

## Annotation methodology

To better understand the attention patterns on Twitter amidst the COVID-19 crisis, we develop a taxonomy of account categories and then proceed to annotate 14,200 accounts using Amazon Mechanical Turk. We devise our taxonomy based on techniques from grounded theory, building a robust categorization scheme of Twitter accounts who participate in COVID-19 discussions. Our methodology encompasses three steps: 1) Account sampling; 2) Iterative development of the taxonomy; 3) Crowdsourced annotation.

### Account sampling

For both iterative development of taxonomy and crowdsourced annotation, we first select a subsample of the accounts who posted at least one tweet with 10 retweets or more about COVID-19 between May 6 and the May 12 and who tweeted in one of the 10 most popular languages in the sample: English, Japanese, Spanish, Portuguese, Italian, Arabic, German and French, Hindi and Indonesian (Table C.3).

1. First, we restrict ourselves to studying only those accounts which posted at least one popular tweet in the 7 days. A tweet is popular if it has received at least 10 retweets. This requirement ensures that sampled accounts received a non-negligible amount of attention. Such accounts comprise 1.96% of all accounts, 1.73% of all tweets, and 84.05% of all retweets, in the COVID-19 stream during the account sampling period.
2. Second, for each language (Table C.3), we calculate quintiles for the number of followers and number of retweets. By doing so, for each language, we have split accounts into 25 "buckets" where each bucket corresponds to a different combination of quintiles for the number of followers and of retweets.
3. Third, we sampled the same number of accounts from each bucket. We sample accounts across languages proportional to the log of the number of tweets in that language, so that accounts tweeting in bigger languages are not over-represented.
4. Lastly, we translated all account metadata from accounts that were not tweeting in English into English using Google's translation API.

Overall, tweets that got at least 10 retweets obtain 84.05% of all retweets on COVID-19 tweets, so in this way, we capture the majority of the engagement COVID-19 tweets receive in total.

### Iterative development of the taxonomy

Next, we explain the steps taken to develop the taxonomy.

1. **Building the initial taxonomy.** Before inspecting the data, the authors discussed broad relevant categories of individuals and entities likely to play a significant role in the COVID-19 online debate. It was determined that categories have to either represent concrete occupations (researcher, medical doctor, and similar) disparately affected or in other ways essential in the context of the pandemic; or, groups of individuals or institutions that shape public discourse. Also, categories had to be significantly represented in the data. However, this was only considered at the end of each iteration, when considering which labels to incorporate to the taxonomy.

2. **Initial inspection.** Three researchers (all authors of the paper) independently explored three different random samples of account descriptions in English, consisting of a hundred accounts each. This was done to build a common understanding of the type of descriptions prevalent in the data. We defer explaining how the samples were generated to section 5.4. For each account, researchers assessed the information about how the account presents itself: the description of the account, Twitter handle, and name. Researchers carefully analyzed the account descriptions considering the categories and wrote notes about the applicability of categories. After that, researchers shared their observations, discussed the initial categories, and adapted them.

3. **Iterative Coding.** Iterative coding was done as follows. In each iteration, three researchers (all authors of this paper) annotated the same set of 100 accounts, with the possibility of expanding the category set. Each account was to be assigned any number of categories, which were determined based on accounts' self-declaration on Twitter (we did not inspect any other information beyond the description, the account name, and the screen name). At the end of each round, researchers individually discussed all disagreements and the overall appropriateness of the categories. Then, they made changes to the categories when necessary, adding new categories or tweaking the definitions of existing categories. Before starting the iterative coding, the researchers agreed on the criteria for stopping the iterations. All of the following three criteria had to be satisfied: 1) Average pairwise Fleiss Kappa agreement is greater than 0.6; 2) Researchers agree that the categories are not ambiguous; 3) The difference in the prevalence of "Other" between two subsequent iterations is smaller than 5%. We repeated this annotation process three times before satisfying all three criteria, the rounds yielded inter-annotator agreements of 0.6, 0.65, and 0.67, respectively. We depict the final taxonomy in Tables C.1 and C.2. Notice that during the analyses in the paper, we collapsed some of the categories together as some were rather sparse, and as their joint interpretation was useful.

Orthogonal to categories, annotators were also asked to identify for each account, whether the account belonged to an individual or an institution. For this annotation, in the iterative coding stage, inter-annotator agreement scores were of 0.63, 0.89, and 0.83, respectively.

### Crowdsourced annotation

We detail the crowdsourcing annotation process, where we annotate 14,200 sampled accounts. This amounts to 7.2% of all accounts adhering to our restrictions, a total of 14,200 accounts out of 196,948. The human intelligence task (HIT) design is shown in Figure C.1. Crowdsourced workers were paid 0.50 USD per HIT, and each HIT consisted of a batch of 10 different account annotation tasks. According to our estimates, it took 2–3 minutes to complete a single HIT, which made the compensation for the task substantially above US federal minimum wage of 7.25$/h. Annotators had to select the categories from small boxes, each of which contained a description of the category, as well as a couple of explained examples (an account bio, and the reason why it would fit in a given category). To study the feasibility of the annotation through crowdsourcing we ran a pilot where crowd-workers had to annotate the same tasks as the researchers did in their last iteration (when the categories were already set). We found that the results were satisfactory, majority vote label of crowd workers agreeing with the majority vote category of researchers 82% of the time. For the type of account (individual vs. institution) the agreement was of 91%. Once the feasibility was established, we proceeded to annotate the accounts collecting 3 independent annotations per account. For accounts for which there was no clear agreement on the category (i.e., there is no single most frequent annotation of type or category attributed by multiple workers), we collected annotation by an additional fourth annotator. In total, we annotated 14,200 accounts belonging to 10 languages. We report the inter-annotator agreement for each language in Table C.4. For each account, we determine its dominant category as the most frequent annotation marked by at least two workers. If there are multiple most frequent annotations assigned by multiple workers, we break the tie randomly to choose one dominant (4.65% of accounts). If there is no agreement, i.e., there is no most frequent category annotation given by at least two workers, we don't assign a dominant category annotation (7.26% of accounts). Finally, we limit our analysis to accounts tweeting in English, Japanese, Spanish, Portuguese, Italian, Arabic, German and French, and discard Hindi and Indonesian, as we spotted lower inter-annotator agreement compared to the other languages (less than 0.2), likely due to poorer automated translation quality.

### Inverse Probability Weighting

In all the conducted analyses, we had to extrapolate the distribution of categories we observed in the sampled data to all the accounts. Recall that we divided all the tweets into 25 buckets and sampled, for each language, the same amount of accounts for each bucket. However, the buckets did not have the same amount of accounts each, and thus it may be that we over-represented some of the buckets and under-represented others. To address this issue, we perform an Inverse Probability Weighting scheme where we calculate the probability of being

sampled, $ps$, at each bucket $k$ as:

$$ps_k = \frac{\#sampled_k}{\#accounts_k} \tag{5.1}$$

and use the inverse value, that is $ps_k^{-1}$ as the weight for all accounts in that bucket. Intuitively, this means that if we proportionally sampled twice from one of the buckets, these accounts will receive half the weight. Let $1_{\{cat,acc\}}$ be an indicator variable that indicates, for a given account and a given category, whether most annotators thought the account belonged to the category. To calculate the probability of a given category for a given language, we simply calculate, for all accounts of that language, the average of the indicator variable $1_{\{cat,acc\}}$ weighted according to the bucket the account was in. To obtain a confidence interval, we bootstrap this calculation 1000 times. That is, we generate a random sample for each language obtaining $k$ accounts from each bucket (thus simulating the original sampling procedure) and then calculate the category distribution. We repeat it 1000 times to obtain 95% confidence intervals. This procedure is used to obtain representative weights for Figures 5.1, 5.2, and 5.3. We use the same methodology to provide supplementary view on the category and type prevalence across languages in Fig. C.2.

## Acknowledgements

The Digital Epidemiology Lab (M.M. and M.S.) received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874735, "Versatile emerging infectious disease observatory - forecasting, nowcasting and tracking in a changing world (VEO)", and was supported with access to Cloud TPUs from Google's TensorFlow Research Cloud and Google Cloud credits in the context of COVID-19-related research. Data was made available by Twitter for academic research in the context of COVID-19.

## Author contributions

K.G. and M.H.R. produced the annotated dataset and led the analysis; M.M. and O.A. performed the analysis, extracted COVID-19 stream data and created the machine-learning models. M.P. assisted in the development of the user taxonomy. G.C., M.S., and R.W. formulated research goals. All authors developed and designed the methodology and the conceptual design of the project, and all authors participated in writing the manuscript and provided critical review.

## Competing interests

The authors declare no competing interests.

## References

Broniatowski, David Andre, Michael J Paul, and Mark Dredze (2014). "Twitter: big data opportunities". In: *Inform* 49, p. 255.

Reuter, Christian and Marc-André Kaufhold (2018). "Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics". In: *Journal of Contingencies and Crisis Management* 26.1, pp. 41–57.

Wagner, Moritz et al. (2018). "The added value of online user-generated content in traditional methods for influenza surveillance". In: *Scientific reports* 8.1, pp. 1–9.

Shin, Soo-Yong et al. (2016). "High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea". In: *Scientific reports* 6, p. 32920.

Salathé, Marcel et al. (2013). "Influenza A (H7N9) and the importance of digital epidemiology". In: *The New England journal of medicine* 369.5, p. 401.

Park, Han Woo, Sejung Park, and Miyoung Chong (2020). "Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea". In: *Journal of Medical Internet Research* 22.5, e18897.

Lazer, David MJ et al. (2018). "The science of fake news". In: *Science* 359.6380, pp. 1094–1096.

Swire-Thompson, Briony and David Lazer (2020). "Public health and online misinformation: challenges and recommendations". In: *Annual Review of Public Health* 41, pp. 433–451.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). "The spread of true and false news online". In: *Science* 359.6380, pp. 1146–1151.

Del Vicario, Michela et al. (2016). "The spreading of misinformation online". In: *Proceedings of the National Academy of Sciences* 113.3, pp. 554–559.

Depoux, Anneliese et al. (2020). *The pandemic of social media panic travels faster than the COVID-19 outbreak*.

Chen, Emily, Kristina Lerman, and Emilio Ferrara (2020). "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set". In: *JMIR Public Health and Surveillance* 6.2, e19273.

Yang, Kai-Cheng, Christopher Torres-Lugo, and Filippo Menczer (2020). *Prevalence of low-credibility information on twitter during the covid-19 outbreak*.

Li, Lifang et al. (2020). "Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo". In: *IEEE Transactions on Computational Social Systems* 7.2, pp. 556–562.

Alshaabi, Thayer et al. (2020). "How the world's collective attention is being paid to a pandemic: COVID-19 related 1-gram time series for 24 languages on Twitter". In: *arXiv preprint arXiv:2003.12614*.

Vicari, Stefania and Maria Francesca Murru (2020). "One Platform, a Thousand Worlds: On Twitter Irony in the Early Response to the COVID-19 Pandemic in Italy". In: *Social Media+ Society* 6.3, p. 2056305120948254.

WHO (2020a). *Novel Coronavirus (2019-nCoV): situation report, 13*.

– (2020b). *WHO media briefing, February 8, 2020*. URL: https://www.who.int/dg/speeches/detail/director-general-s-remarks-at-the-media-briefing-on-2019-novel-coronavirus---8-february-2020 (visited on 09/01/2020).

Dutta-Bergman, Mohan (2003). "Trusted online sources of health information: differences in demographics, health beliefs, and health-information orientation". In: *Journal of medical Internet research* 5.3, e21.

Dredze, Mark, David A Broniatowski, and Karen M Hilyard (2016). "Zika vaccine misconceptions: A social media analysis". In: *Vaccine* 34.30, p. 3441.

Hossain, Tamanna et al. (2020). "Detecting COVID-19 Misinformation on Social Media". In:

Sharma, Megha et al. (2017). "Zika virus pandemic-analysis of Facebook as a social media health information platform". In: *American journal of infection control* 45.3, pp. 301–302.

Vraga, Emily K and Leticia Bode (2017). "Using expert sources to correct health misinformation in social media". In: *Science Communication* 39.5, pp. 621–645.

Hagen, Loni et al. (2018). "Crisis communications in the age of social media: A network analysis of Zika-related tweets". In: *Social Science Computer Review* 36.5, pp. 523–541.

Abu-Akel, Ahmad, Andreas Spitz, and Robert West (2020). "The Fauci effect: Public Health messaging during the COVID-19 pandemic". In:

Rufai, Sohaib R and Catey Bunce (2020). "World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis". In: *Journal of Public Health*.

Cinelli, Matteo et al. (2020). "The covid-19 social media infodemic". In: *arXiv preprint arXiv:2003.05004*.

Ahmed, Wasim et al. (2020). "COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data". In: *Journal of Medical Internet Research* 22.5, e19458.

Ferrara, Emilio (2020). "# covid-19 on twitter: Bots, conspiracies, and social media activism". In: *arXiv preprint arXiv:2004.09531*.

Kouzy, Ramez et al. (2020). "Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter". In: *Cureus* 12.3.

Pulido, Cristina M et al. (2020). *COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information*.

Gallagher, Ryan J et al. (2020). "Sustained Online Amplification of COVID-19 Elites in the United States". In: *arXiv preprint arXiv:2009.07255*.

Hernández-García, Ignacio and Teresa Giménez-Júlvez (2020). "Assessment of health information about COVID-19 prevention on the internet: infodemiological study". In: *JMIR public health and surveillance* 6.2, e18717.

Cha, Meeyoung et al. (2010). "Measuring user influence in twitter: The million follower fallacy." In: *Icwsm* 10.10-17, p. 30.

# 6 International expert communities on Twitter become more isolated during the COVID-19 pandemic

**Published as:**
Durazzi F., Müller M., Salathé M., Remondini D. (2020). International expert communities on Twitter become more isolated during the COVID-19 pandemic *arXiv preprint* arxiv:2011.06845

**Abstract**

COVID-19 represents the most severe global crisis to date whose public conversation can be studied in real time. To do so, we use a data set of over 350 million tweets and retweets posted by over 26 million English speaking Twitter users from January 13 to June 7, 2020. In characterizing the complex retweet network, we identify several stable communities, and are able to link them to scientific expert groups, national elites, and political actors. We find that scientific expert communities received a disproportionate amount of attention early on during the pandemic, and were leading the discussion at the time. However, as the pandemic unfolded, the attention shifted towards both national elites and political actors, paralleled by the introduction of country-specific containment measures and the growing politicization of the debate. Scientific experts remained present in the discussion, but experienced less reach and a higher degree of segregation and isolation. Overall, the emerging communities are characterized by an increased self-amplification and polarization. This makes it generally harder for information from international health organizations or authorities to reach a broad audience. These results may have implications for information dissemination in future global crises.

## 6.1 Introduction

Twitter has been widely used as a tool for emergency response in previous crises and disasters and a large body of work has focused on optimizing communication in order to adjust or nudge human behaviour under such conditions (Chen et al. 2008; Li and Rao 2010; Martínez-Rojas, Pardo-Ferreira, and Rubio-Romero 2018; Salathé et al. 2013). Research finds that during a time of crisis, when information is scarce and heavily sought after, social media enables the critical flow of information due to its collaborative nature (Graham, Avery, and Park 2015; Lee Hughes and Palen 2009). However, the underlying network and community structure is likely to have a significant influence over which information users are exposed to (Conover, Ratkiewicz, and Francisco 2011; M. E. Newman 2006). In particular, retweet interactions have been shown to reflect real-life community structure, such as political parties (Cherepnalkoski and Mozetič 2016), and to characterize the information sharing dynamics of different clusters of users (Bovet and Makse 2019). Likewise, communities of users sharing misleading news tend to be more connected and clustered (Pierri, Piccardi, and Ceri 2020).

The structure of such information networks has been assessed in various disease outbreak scenarios, most prominently in the context of the Zika virus outbreak 2015–2016 in the US (Hagen et al. 2018). Among other findings the work suggests that key international experts acted as "boundary spanners", who were able to spread information quickly through the network. Even though it has been observed that false information travels faster in social networks (Mendoza, Poblete, and Castillo 2010; Vosoughi, Roy, and Aral 2018), the work finds that Twitter users have made efforts to diffuse reliable information by such key experts. Related work demonstrates how the association between the geographical location of users dictates topics of conversations, thereby matching the temporal and geographical spread of the Zika outbreak (Stefanidis et al. 2017).

Previously listed work allows for anecdotal evidence of crisis-induced behaviour on social media, but compared to the COVID-19 crisis the events were 1) shorter in time scale 2) more localized and 3) orders of magnitude smaller in terms of analyzed data. As the virus spread across the world within only a few weeks, Twitter became the primary news source among expert groups and medical personnel. Reliable sources, such as peer-reviewed literature or other officially vetted channels, were simply too slow to be useful during this time (Rosenberg, Syed, and Rezaie 2020). However, this initially positive role of Twitter and other social media has now been overshadowed by the spread of numerous conspiracy theories and other low-quality mis- and disinformation about the pandemic, peaking in what the WHO now considers an "infodemic" (*WHO Director General Tedros Adhanom Ghebreyesus at the Munich Security Conference on February 15, 2020* 2020; Cinelli et al. 2020). Preliminary work focused on this issue and used social network analysis in order to determine drivers of the conspiracy theory which claims a link between 5G and COVID-19 (Ahmed et al. 2020). Although misinformation

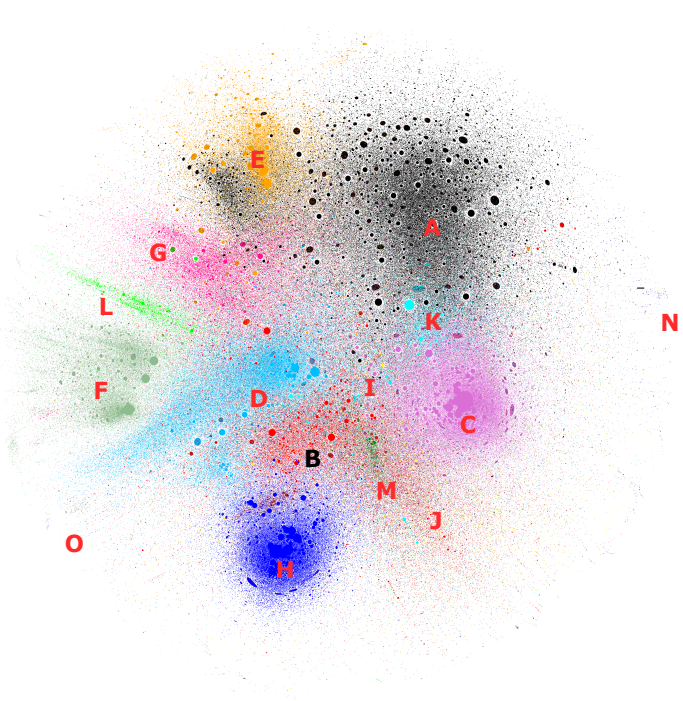
is a major concern, more recent work suggests that experts and authorities are being heard and have received disproportionate attention (Gligoric et al. 2020), but that key specialists may experience low reachability (Mourad et al. 2020). Moreover, other work finds that key medical professionals and scientific experts may experience lower "sustained amplification", meaning that the attention given to this group has not been constant, and overall lower, compared to media outlets or key political figures (Gallagher et al. 2020). Our work attempts to clarify the somewhat ambiguous premises about the role of experts during the pandemic.

With the help of a comprehensive dataset of more than 350 million tweets, we identify the key communities of English speaking Twitter users involved in the COVID-19 debate, starting from early January to June 2020. We then provide a detailed analysis of the evolution of this massive communication network and characterize the interaction dynamics regulating the sharing of information both within and between the communities.

## 6.2 Results

### 6.2.1 Aggregated network

In the giant component of the directed network (see Material and methods 6.4.4), the out-degree distribution (i.e. number of retweets received by each user) follows a very skewed distribution, typical of many real-world networks (Figure D.1). Users with an out-degree higher than 1500 represent the 0.1% of the users in the network, but their tweets have been retweeted >200M times (77.0% of all retweets). The community detection algorithm reveals thousands of communities, spanning from millions down to duplets, with size decreasing very sharply (Figure D.2). By aggregating over multiple repetitions of the algorithm, we identify 15 communities (labeled with letters from A to O, in decreasing order of size) with more than 105 users, encompassing 97.9% of all users in the giant component.




| S | C | Name | Dominant category | Size |
|---|---|---|---|---|
| | | I | Sports | 2.1% |
| | | J | Science | 2.0% |
| | | D | Science | 9.4% |
| | | M | Science | 0.8% |
| | | E | Arts & Entertainment | 7.2% |
| | | N | Adult content | 0.6% |
| | | A | Arts & Entertainment | 33.3% |
| | | O | Business | 0.6% |
| | | B | Science | 10.6% |
| | | G | Science | 6.4% |
| | | F | Science | 6.9% |
| | | L | Science | 1.1% |
| | | H | Political Supporter | 5.4% |
| | | C | Political Supporter | 10.0% |
| | | K | Arts & Entertainment | 1.9% |

Super-community

National elite
Other
International expert
Political

Figure 6.1: Retweet network of a randomly sampled connected component of 1M users, colored by community. Node size is proportional to node out-degree. In the table, the column "S" designates the color of the super-communities used throughout this work, "C" lists the community color in the network layout and "Name" the respective community name. The "Dominant category" column specifies the most abundant user category in the community (excluding "Other"). "Size" denotes the ratio of users in the community with respect to the total number of users in the network.

Figure 6.1a shows the network of 1M users (corresponding to 4.3% of all nodes) randomly sampled but forming a connected component. Even if the layout algorithm has no information about the communities we identified, it recovers a similar stratification of the nodes. We observe that communities A and B fill a great portion of the network, without being as densely connected as the other communities. Some communities (C, H, F, L) are more peripheral and exhibit a neat modular structure, which implies a high degree of internal connectivity (see also Fig D.3).

### 6.2.2 Characterization of communities

In order to characterize the 15 largest communities, we 1) infer the presumed location of the users and 2) use the results of a Machine Learning classifier, which was trained to predict the user category based on a user's description (see Materials and methods 6.4.2). Self-reported location information was available for 54% of nodes in the network, while a clear category could be assigned to 25% of considered users (Table D.1). Since the diversity in nationality of the communities is of interest, we calculate the entropy of the distribution of countries for each community (similarly to the alpha diversity index in ecological communities): a high entropy indicates an almost uniform distribution of users across all countries, whereas a low entropy implies an uneven distribution, possibly skewed toward a specific country. Some communities are strongly associated with user location, in particular communities D (UK), E (Philippines and Southeast Asia), F (India), G (African countries), J (Canada), L (Pakistan) and M (Australia). The US represents more than a half of the users from communities C, H, I, and K, while communities A, B, N and O are more heterogeneous in terms of location, containing a large fraction of tweets located outside of the US.

In some communities, the most retweeted users have a strong association to political and cultural topics. In order to better understand these mixed communities, we investigate the predicted user categories for each community (see Materials and methods 6.4.2 and Figure D.4). Out of the 13 categories, the category "Other" is the majority category for all considered communities, being assigned to 78% of all users. Science, being the second largest category (at 4.7% of all users), was represented at 9% or higher in communities B, D, G, J and M. Other communities, such as H or C, have a higher fraction of users who identify with a political movement.

In this study, we are particularly interested in the specific user categories "Science", "Healthcare", "Media", "Politics & Government", "Public Services", and "Political Supporter". The percentage of users in these six categories and the entropy measure for internationality were used for hierarchical clustering at the community level (see Figure 6.2). With the help of the emerging dendrogram, we recognize four super-communities, which we name as follows:

- International expert: communities exhibiting an increased number of users of category "Science" and a high international diversity.
- National elite: communities exhibiting an increased number of users belonging to official categories (i.e. "Science", "Healthcare", "Media", "Politics & Government", and "Public Services") and a low international diversity.
- Political: communities exhibiting an increased number of users associated with political activism (category "Political Supporter") or involved in politics ("Politics & Government").
- Other: communities which are not linked to any of the categories of interest. This includes artists, entrepreneurs, and non-governmental activists.

The naming of super-communities is further validated by the manual inspection of the top users in the respective communities. In particular, only two communities (B and G) are assigned to "international expert", showing a clearly distinguishable pattern. These communities consist of a large fraction of users who are presumably working in scientific or health-related fields. Community B's top users include well-known news agencies, as well as the WHO, making this community led by official media and scientific information spreaders. Community G is similar, but it has more users from African countries, in particular Nigeria and South Africa. National elites, i.e. communities I, J, D and M, are communities with a high proportion of official categories, but linked to specific countries. Among the political super-community, communities F and L have the highest proportions of users and institutions involved in politics, while communities C, H and K are driven by US-specific political debates. Upon visual inspection of a sample of accounts, it emerges that community C and K are more often associated with the Democratic party, and H with users from the Republican party. All other communities, including community A (comprising 32% of all users), show characteristics which were not deemed relevant for this analysis.

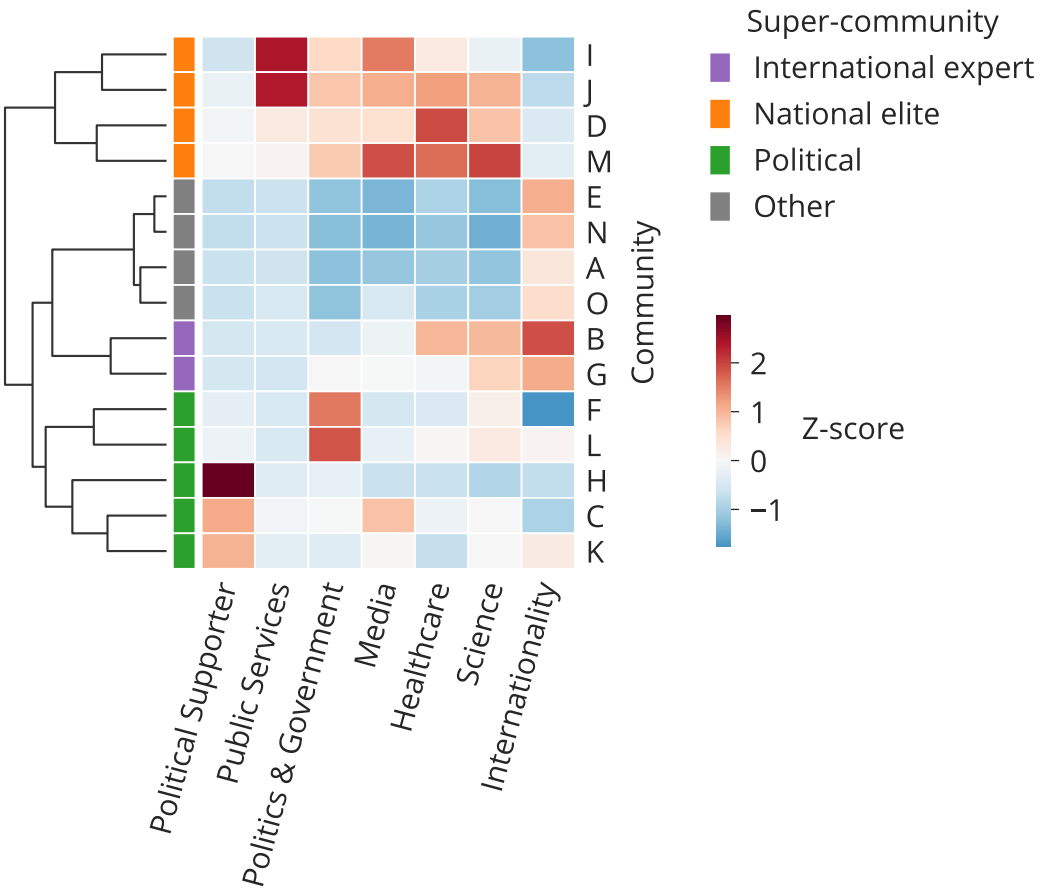



Figure 6.2: Clustering of communities into super-communities. The heatmap shows the Z-scores (i.e. standardized values) for seven chosen features. The four super-communities denote the emerging clusters.

### 6.2.3 Network dynamics

In order to trace the evolution of the network throughout the pandemic, we reconstructed several networks by splitting the Twitter data into non-overlapping windows of 1 week. We collapse the 1-week retweet networks into networks with four nodes, corresponding to the four previously assigned super-communities (see Figure 6.3a) and thereby compute the so-called mixing matrix (M. Newman 2010). In these networks, we draw a link from super-community $i$ to $j$, with a weight $w_{ij}$ equal to the number of times super-community $j$ retweeted $i$ in a given week.

We assign each node a size attribute $N_i$, representing the number of users of super-community $i$ retweeting or being retweeted in the given week. Figure 6.3b shows the temporal evolution of this attribute. Two principal observations can be made: first, the value of $N$ varies over time in correspondence with the phases of the epidemic. We identify an initial peak in total number of users in the beginning of February (peak $a$), and a second one at the end of March (peak $b$). These two peaks presumably correspond to the first diffusion of news about COVID-19 in China and, later, to their diffusion worldwide (Figure D.5). In the time between peak $a$ and $b$ the number of users in the COVID-19 debate has doubled, followed by a slow decay between April and June. Second, most of the change in the number of users stems from the "Other" super-community. This implies that the three super-communities of interest remained present throughout the entire observation period at a relatively static level.

In Figure 6.3c, we show the average attention per user of super-community $i$, defined as:

$$A_u^i = \frac{\sum_j w_{ij}}{N_i} \tag{6.1}$$

i.e. the weighted out-degree of super-community $i$, normalized by its size. The international expert super-community faces an increase in average attention per user in January and stabilizes at a higher level compared to other super-communities until the beginning of March. After a narrow peak in March, the political super-community plateaus in April at roughly three times the attention level of national elites and international experts.

We then split the total attention, i.e. the sum of all weighted edges $W$, for each super-community $i$ into an internal and external component for every weekly network:

$$a_i^{ext} = \frac{\sum_{j \neq i} w_{ij}}{W} \tag{6.2}$$

$$a_i^{int} = \frac{w_{ii}}{W} \tag{6.3}$$

$$a_i^{ext} + a_i^{int} = 1 \tag{6.4}$$

The external component $a_i^{ext}$ represents the attention given to super-community $i$ from the other super-communities, while the internal component $a_i^{int}$ quantifies self-amplification. Figure 6.3d shows that the external attention component is decreasing overall, indicating a decrease in attention between super-communities. This is particularly true for the international expert super-community, which received broad attention in the very beginning of the pandemic, peaking again in mid-February, and then decaying in a monotonic way until the end of our sampling.

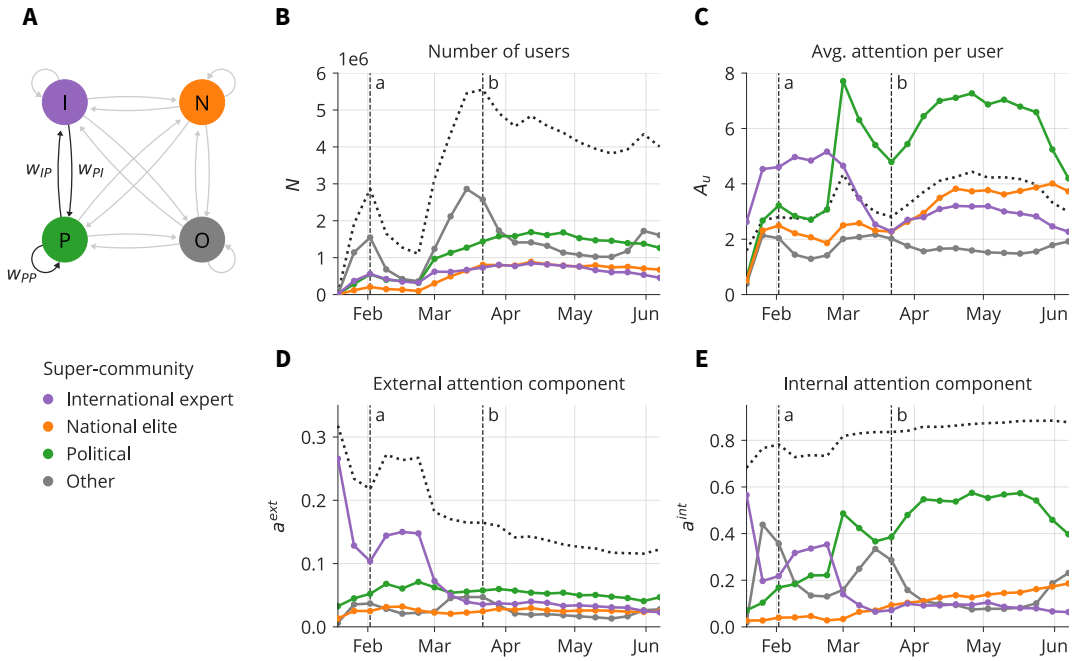



Figure 6.3: Evolution of weekly aggregated networks by super-community, with dotted lines corresponding to the statistics across all users. **(A)** Diagram representing the networks collapsed to the super-community level. Edge direction represents the flow of information via retweets, i.e. from retweeted to retweeting super-community. **(B)** Size of the super-communities in terms of number of users. **(C)** Average attention per user. **(D)** External component of the attention toward super-communities. **(E)** Internal component of the attention toward super-communities. Indicated as *a* and *b* are the first and second peak in terms of network size, as shown in Figure 6.3b.

Figure 6.3e shows that the internal attention component is increasing overall, highlighting an increased self-amplification within the network. We observe that this increase is mostly driven by the political communities and to a lesser degree by the national elites. The international expert super-community, on the other hand, decreased internal sharing of content after March. Overall, we note that the dynamics partially mirror Figure 6.3c, since the internal attention component makes up most of the total attention given to the super-communities.

### 6.2.4 Sustained attention towards top users

So far, in our analysis the observed dynamics of super-communities is a result of the average behavior across all the users in the Twittersphere, while in reality most of the dynamics are driven by a relatively small set of users who receive disproportionate attention. Further, we have focused on the number of retweets (node out-degree) as a canonical measure of attention, but the retweet count of a single viral tweet might exaggerate the user's real impact in the overall debate.

In this section, we address these caveats. Here, we only consider the top 1000 users for each super-community in terms of retweets received, i.e. 4000 users receiving 55.0% of all retweets. Furthermore, we adopt an alternative measure to characterize the attention given to users, namely a retweet $h$-index, as previously introduced by (Gallagher et al. 2020). Originally proposed in the context of academic citations (Hirsch 2005), the $h$-index in this case reflects that the user has received at least $h$ retweets on $h$ of their original tweets.

Figure 6.4a and 6.4b compare the rank in terms of retweets received $r_{rt}$ and $h$-index $r_h$ both on the user and the super-community level, respectively. A user placed in the top-left in the figure plane suggests that few of their tweets received punctual attention at high virality, whereas the bottom-right suggests sustained or long-lasting attention at low virality.

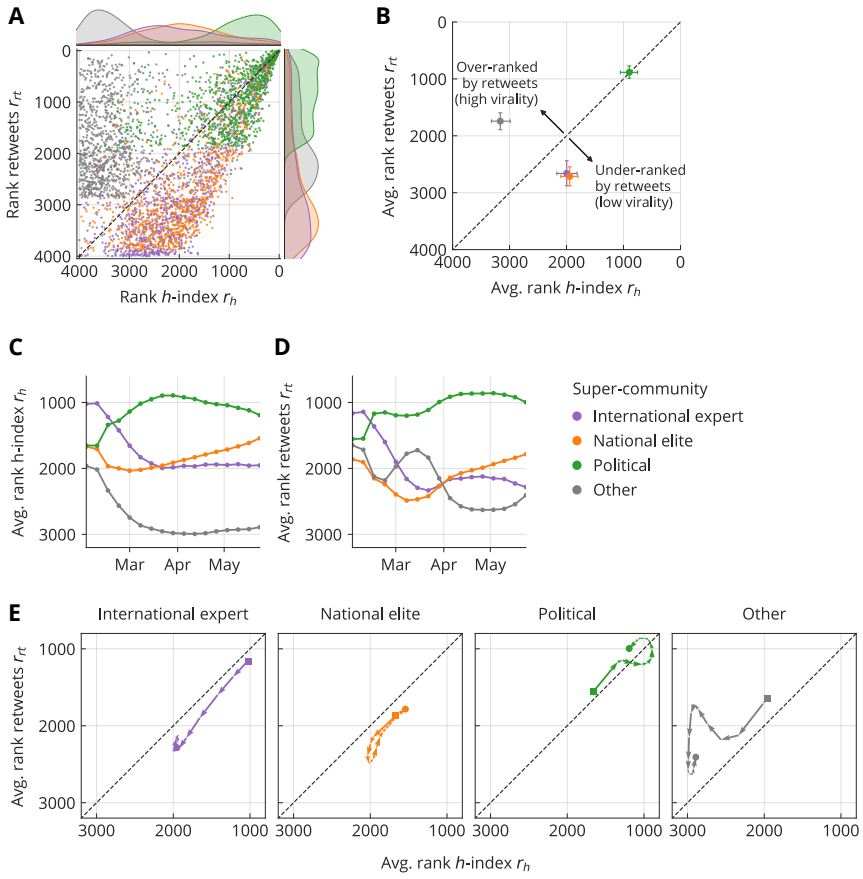



Figure 6.4: Static and temporal communication patterns of the top 1000 users of each super-community. **(A)** Comparison between the rank in terms of $h$-index and retweets by user, as well as respective marginal distributions. **(B)** The average rank by super-community with bootstrapped 95% confidence intervals. **(C)** Weekly $h$-index rank computed within a rolling time window of 1 month. **(D)** Weekly rank in retweets computed within a rolling time window of 1 month. **(E)** Vector plot of retweet and $h$-index ranks by super-community. Each arrow denotes the change in ranks within one week. The first and last week are marked with a square and circle, respectively.

Generally, users belonging to the political super-community are ranked highest both in terms of retweets and $h$-index, receiving most of the attention both on a punctual and an extended time scale. National elites and international experts behave very similarly: they rank medium to high in terms of $h$-index, but low in the number of retweets received (low virality). Lastly, the "Other" category ranks generally lowest in $h$-index and intermediate in retweets, thus is characterized by a higher virality in terms of attention.

In order to understand the temporal dimension of these results, we formulate the previous metrics in a time-dependent fashion on the same set of top users. We consider all tweets and retweets posted within a rolling time window of 1 month width and a 1 week step. We then compute the rank by $h$-index (Figure 6.4c) and retweets (Figure 6.4d) averaged by super-community. Additionally, we show the resulting data as a vector plot (Figure 6.4e). We find that international experts scored very highly in both metrics initially and then experienced a drop in ranks. Similarly, national elites faced an initial decline in both metrics but then increased in ranks above the level of international expert after April. The temporal view of this data therefore adds nuance to the picture obtained in Figure 6.4b.

We conclude that although national elites and international experts share a significant overlap in the static view (cf. Figure 6.4b), they reveal distinct temporal dynamics (cf. Figure 6.4c–e). This result reflects how the international expert community faced a decline in attention, while the discussion has been moving onto more local grounds, with the political and national debate gaining momentum.

## 6.3 Discussion

In this work we use a complex network approach to answer some relevant questions about the role of experts in relation to the COVID-19 pandemic in the English-speaking Twittersphere. Using a community detection algorithm, we identify 15 user communities, exhibiting a stable structure throughout the pandemic. We are able to group these communities into four main super-communities related to the prevalent user categories and the degree of internationality (namely international expert, national elite, political, and other) and assess their interaction patterns over time.

In the Twitter landscape of COVID-19 we identify a single major group of scientific experts with a highly international distribution of users, and multiple country-specific communities which appear to engage more in the respective national debates. Additionally, we find several large country-specific communities which are mostly characterized by political activism, thus highlighting the substantial politicization in the discussion surrounding the pandemic.

Our results emphasize the role of the international super-community of scientific experts

in the beginning of the, at the time, largely unknown pandemic. This super-community received disproportionate attention and had broad reach across many cultural and political communities, as reflected by the high total volume of retweets received, in particular from non-expert communities. As demonstrated by the high internal attention component, international experts also shared content frequently among themselves, possibly in an attempt to rapidly share scientific insights about the novel virus.

In a second critical moment in March the number of COVID-19 cases exploded in almost all parts of the world, leading to massive media attention, which is well reflected in the increase in community sizes, as well as in the total number of tweets. It is noteworthy that this added attention has not been allocated to international experts but rather to political leaders, reflecting the loss of influence of the former in the evolution of the debate.

As the pandemic unfolded in April, and while many English speaking countries underwent a strict lockdown, we observe a growing politicization of the debate, reflected by the fact that content by the political communities is now shared most. Meanwhile, the analysis reveals the picture of an increasingly segregated international expert super-community. Furthermore, compared to January and February, and in contrast to all other groups, the international expert super-community also reduced interactions among themselves, even though their size remained constant.

The analysis of the sustained amplification patterns of highly retweeted users mostly confirms our previous analysis: compared to political leaders, top users in the international expert super-community received intermediate levels of sustained attention but their content lacked virality. Additionally, the results show the importance of the national elites in influencing the discussion after the end of March: national elites' top users show a positive trend in both punctual and sustained attention received, to the point of surpassing the international expert super-community.

Our work allows for the resolution of some of the discrepancies in recent literature on the role of scientific experts in the COVID-19 pandemic by adding a temporal dimension to the picture. We are able to confirm that scientific experts were heard early on in the pandemic, as found by (Gligoric et al. 2020). This also fits in well with previous literature, which confirms that Twitter users amplify information from trusted sources in crisis situations (Reuter and Kaufhold 2018; Wagner et al. 2018; Shin et al. 2016). Further, we detect and trace these groups on the interaction network level and find that the reachability and attention of scientific experts, as received both from within the group and from outside, has declined over time, as indicated by earlier work (Mourad et al. 2020). We can partially confirm previous work (Gallagher et al. 2020) that suggested low sustained attention to top medical experts. However, by giving a temporal dimension to our analysis, we find that scientific experts ranked highest both in sustained and punctual attention in the very beginning of the pandemic and only later became increasingly

isolated in terms of attention. We believe this is an important result which underlines the role of scientific experts as possible boundary spanners during the early phase of a pandemic.

Our claims are based on a comparatively large dataset encompassing a total of 354M tweets by 26M users. This dataset can be considered comprehensive (see Materials and methods 6.4.1). However, a limitation of this work is that a substantial part of the network consists of the super-community labelled as "Other", encompassing around 50% of all users. It is difficult to make general statements about the true nature of this group, as it includes a very diverse set of users, with most of them reporting unspecific profile information. As our work is based on the self-reported expert status of users, future research is required to properly understand their true nature.

Our work leads to two main conclusions: 1) Under the unique circumstance of an emerging virus causing a global pandemic, the Twitter platform allowed thousands of international experts to quickly and efficiently exchange information, while also being amplified by non-expert communities. 2) As the pandemic developed, Twitter users directed more of their attention towards the national debates, overall leading to more segregated communities.

As the world faces a range of societal and economical issues related to the COVID-19 pandemic, there is a growing need for a coherent communication strategy by trusted sources in order to combat misinformation. In light of our results, it is challenging to envision a strategy which is globally applicable. However, our work informs the development of temporally and locally adaptive communication strategies, which may involve the inclusion of key influential figures in order to reach the niches of the increasingly segregated network of the COVID-19 debate on Twitter.

## 6.4 Materials and methods

### 6.4.1 Data collection

Twitter data was collected through the Twitter API, specifically through the filter streaming endpoint, using the Crowdbreaks platform (Müller and Salathé 2019). The data used in this work consists of a total of 353,993,900 tweets (thereof 267,026,740 retweets) posted by 26,262,332 users in a 146 day observation period, i.e. from January 13 to June 7, 2020. These tweets have been identified by Twitter to be in English language and match one or more of the keywords "wuhan", "ncov", "coronavirus", "covid" and "sars-cov-2". Note that keywords have changed over time, as the new names for virus were introduced (for details refer to section D.1 and Table D.2). The data is complete with respect to these keywords, except during a period between mid-March to mid-April when volume exceeded the 1% threshold imposed by Twitter and was subsampled by an (unknown) degree.

### 6.4.2 User categorization

In order to be able to interpret the identified network communities, accounts were categorized by occupational role and account type. This categorization was conducted using a Machine Learning classifier which was trained to determine the category of an account solely based on the user description (user bio). The classifiers were first published in the context of related work on the attention given to experts during COVID-19 (Gligoric et al. 2020). In this work, we use the published English language BERT model (`bert-english-pt`) in order to determine the category of each account in our dataset. In the aforementioned study (ibid.), the categories have been determined in an iterative coding process to best categorize users into 13 categories, which are: Adult content, Arts & Entertainment, Business, Healthcare, Media, Non-governmental organization (NGO), Political Supporter, Government & Politics, Public Services, Religion, Science, Sports, and Other. For further details on the coding process or the training of the machine learning classifiers, please refer to the referenced study (ibid.).

### 6.4.3 Geo-localization

Tweet objects contain both structured and unstructured forms of geographical information. In this work, we employed a procedure to geo-localize tweets on the country level using the Python library local-geocode ([I], please refer to section D.2 for detailed explanations). A user's country location was determined from the majority of the user's geo-localized tweets. Geolocation could be inferred for 75% of tweets belonging to 54% of the considered users.

### 6.4.4 Network analysis

We study the full directed retweet network, consisting of all retweets collected during the entire 147 day observation period (267M retweets). The nodes of the network represent users who have at least once retweeted or have been retweeted by another user. An edge was established from user A to user B if B retweeted A at least once during the whole period of data collection. Therefore, the edge direction indicates the flow of information. We assigned a weight to this edge equal to the number of times user B has retweeted user A. The reconstructed (weighted and directed) network has 22.9M nodes and 177M unique edges. In order to study the communities in this network, we consider only the largest connected component of the network, consisting of 22.5M nodes and 176M unique edges. The discarded components only consisted of isolated nodes or duplets, making up 1.57% of the nodes and 0.13% of the edges in the original network. We ran Louvain's community detection algorithm (Blondel et al. 2008), as implemented in Python's package Networkit (Staudt and Meyerhenke 2016). The algorithm attempts to detect clusters of nodes by recursively maximizing the network's

[I] `https://github.com/mar-muel/local-geocode`

modularity. Standard modularity was adopted as a scoring function, meaning that intra-cluster edges were counted with the same weight as inter-clusters edges. For the community detection task, we considered the network edges as undirected in order to reduce the computational burden. Due to the stochasticity of the clustering algorithm, we ran 50 trials and assigned each node to the community it was most frequently associated with. The identification of largest communities was found to be stable both among repeated runs of the algorithm (Figure D.6) as well as when comparing monthly time windows of the dataset (Figure D.7, please refer to section D.3 for further details). Thus, the results of the community detection can be considered as fairly robust. The coordinates of the network layout in Figure 6.1 were processed by Gephi software (Bastian, Heymann, and Jacomy 2009), using the ForceAtlas2 algorithm (Jacomy et al. 2014) with gravity set to 0.05 with the "stronger gravity" option enabled.

**Data availability.** All data and code can be found on our public GitHub repository `https://github.com/FraDurazzi/twitter-network-covid19`. The full Twitter dataset used in this work is available on Zenodo (Müller et al. 2020).

**Author contributions.** M.M. collected the data. F.D. and M.M. analyzed the data. F.D., M.M., D.R. and M.S. designed the study and wrote the paper.

**Competing interests.** The authors declare no competing interests.

**Acknowledgments.** We thank Marion Koopmans for her valuable comments and ideas that helped to design the study. This project was funded through the European Union's Horizon 2020 research and innovation programme under grant agreement No. 874735 "Versatile emerging infectious disease observatory - forecasting, nowcasting and tracking in a changing world (VEO)".

## References

Chen, Rui et al. (2008). "Coordination in emergency response management". In: *Communications of the ACM* 51.5, pp. 66–73. ISSN: 00010782. DOI: 10.1145/1342327.1342340.

Li, Jessica and H.R. Rao (2010). "Twitter as a Rapid Response News Service: An Exploration in the Context of the 2008 China Earthquake". In: *The Electronic Journal of Information Systems in Developing Countries* 42.1, pp. 1–22. ISSN: 1681-4835. DOI: 10.1002/j.1681-4835.2010.tb00300.x.

Martínez-Rojas, María, María del Carmen Pardo-Ferreira, and Juan Carlos Rubio-Romero (2018). "Twitter as a tool for the management and analysis of emergency situations: A systematic literature review". In: *International Journal of Information Management* 43, pp. 196–208. ISSN: 02684012. DOI: 10.1016/j.ijinfomgt.2018.07.008.

Salathé, Marcel et al. (2013). "Influenza A (H7N9) and the Importance of Digital Epidemiology". In: *New England Journal of Medicine* 369.5, pp. 401–404. ISSN: 0028-4793. DOI: 10.1056/nejmp1307752.

Graham, Melissa W., Elizabeth J. Avery, and Sejin Park (2015). "The role of social media in local government crisis communications". In: *Public Relations Review* 41.3, pp. 386–394. ISSN: 03638111. DOI: 10.1016/j.pubrev.2015.02.001.

Lee Hughes, Amanda and Leysia Palen (2009). "Twitter adoption and use in mass convergence and emergency events". In: *International Journal of Emergency Management* 6.3/4. ISSN: 17415071. DOI: 10.1504/IJEM.2009.031564.

Conover, M, J Ratkiewicz, and M Francisco (2011). "Political polarization on twitter". In: *Icwsm*. ISSN: 15205126. DOI: 10.1021/ja202932e.

Newman, M. E.J. (2006). "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences of the United States of America*. DOI: 10.1073/pnas.0601602103. arXiv: 0602124 [physics].

Cherepnalkoski, Darko and Igor Mozetič (2016). "Retweet networks of the European Parliament: evaluation of the community structure". In: *Applied Network Science* 1. ISSN: 23648228. DOI: 10.1007/s41109-016-0001-4.

Bovet, Alexandre and Hernán A. Makse (2019). "Influence of fake news in Twitter during the 2016 US presidential election". In: *Nature Communications* 10. ISSN: 20411723. DOI: 10.1038/s41467-018-07761-2. arXiv: 1803.08491.

Pierri, Francesco, Carlo Piccardi, and Stefano Ceri (2020). "Topology comparison of Twitter diffusion networks effectively reveals misleading information". In: *Scientific Reports* 10. ISSN: 20452322. DOI: 10.1038/s41598-020-58166-5. arXiv: 1905.03043.

Hagen, Loni et al. (2018). "Crisis Communications in the Age of Social Media: A Network Analysis of Zika-Related Tweets". In: *Social Science Computer Review* 36.5, pp. 523–541. ISSN: 15528286. DOI: 10.1177/0894439317721985.

Mendoza, Marcelo, Barbara Poblete, and Carlos Castillo (2010). "Twitter under crisis: Can we trust what we RT?" In: *SOMA 2010 - Proceedings of the 1st Workshop on Social Media Analytics*. ISBN: 9781450302173. DOI: 10.1145/1964858.1964869.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018). "The spread of true and false news online". In: *Science* 359.6380, pp. 1146–1151.

Stefanidis, Anthony et al. (2017). "Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts". In: *JMIR Public Health and Surveillance* 3.2. ISSN: 2369-2960. DOI: 10.2196/publichealth.6925.

Rosenberg, Hans, Shahbaz Syed, and Salim Rezaie (July 2020). "The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic". In: *Canadian Journal of Emergency Medicine* 22.4, pp. 418–421. ISSN: 14818043. DOI: 10.1017/cem.2020.361. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7170811/.

*WHO Director General Tedros Adhanom Ghebreyesus at the Munich Security Conference on February 15, 2020* (2020). URL: https://www.who.int/dg/speeches/detail/munich-security-conference.

Cinelli, Matteo et al. (Mar. 2020). "The COVID-19 Social Media Infodemic". In: *Scientific Reports* 10. DOI: 10.1038/s41598-020-73510-5. arXiv: 2003.05004. URL: http://arxiv.org/abs/2003.05004.

Ahmed, Wasim et al. (2020). "COVID-19 and the 5G conspiracy theory: Social network analysis of twitter data". In: *Journal of Medical Internet Research* 22.5. ISSN: 14388871. DOI: 10.2196/19458.

Gligoric, Kristina et al. (2020). "Experts and authorities receive disproportionate attention on Twitter during the COVID-19 crisis". In: *arXiv preprint arXiv:2008.08364v1*. arXiv: 2008.08364v1. URL: https://github.com/digitalepidemiologylab/experts-covid19-twitter.

Mourad, A. et al. (2020). "Critical Impact of Social Networks Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-Based Study and Research Directions". In: *IEEE Transactions on Network and Service Management*. DOI: 10.1109/tnsm.2020.3031034. arXiv: 2005.08820.

Gallagher, Ryan J et al. (2020). "Sustained Online Amplification of COVID-19 Elites in the United States". In: *arXiv preprint arXiv: 2009.07255v1*. arXiv: 2009.07255v1. URL: https://developer.twitter.com/en/docs/labs/covid19-.

Newman, Mark (2010). *Networks: An introduction*. Oxford: OUP Oxford. ISBN: 9780199206650. DOI: 10.1093/acprof:oso/9780199206650.001.0001.

Hirsch, J. E. (2005). "An index to quantify an individual's scientific research output". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.46, pp. 16569–16572. ISSN: 00278424. DOI: 10.1073/pnas.0507655102. arXiv: 0508025 [physics].

Reuter, Christian and Marc André Kaufhold (2018). "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis Informatics". In: *Journal of Contingencies and Crisis Management* 26.1, pp. 41–57. ISSN: 14685973. DOI: 10.1111/1468-5973.12196.

Wagner, Moritz et al. (2018). "The added value of online user-generated content in traditional methods for influenza surveillance". In: *Scientific Reports* 8. ISSN: 20452322. DOI: 10.1038/s41598-018-32029-6.

Shin, Soo Yong et al. (2016). "High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea". In: *Scientific Reports* 6. ISSN: 20452322. DOI: 10.1038/srep32920.

Müller, Martin M. and Marcel Salathé (2019). "Crowdbreaks: Tracking health trends using public social media data and crowdsourcing". In: *Frontiers in Public Health* 7. ISSN: 22962565. DOI: 10.3389/fpubh.2019.00081. arXiv: 1805.05491.

Blondel, Vincent D. et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment*. ISSN: 17425468. DOI: 10.1088/1742-5468/2008/10/P10008. arXiv: 0803.0476.

Staudt, Christian L. and Henning Meyerhenke (2016). "Engineering Parallel Algorithms for Community Detection in Massive Networks". In: *IEEE Transactions on Parallel and Distributed Systems* 27.1, pp. 171–184. ISSN: 10459219. DOI: 10.1109/TPDS.2015.2390633. arXiv: 1304.4453.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy (2009). "Gephi: An Open Source Software for Exploring and Manipulating Networks". In: *Third International AAAI Conference on Weblogs and Social Media*. ISBN: 978-1-57735-421-5. DOI: 10.1136/qshc.2004.010033.

Jacomy, Mathieu et al. (2014). "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software". In: *PLoS ONE* 9.6. ISSN: 19326203. DOI: 10.1371/journal.pone.0098679.

Müller, Martin et al. (Nov. 2020). *COVID-19 Twitter data, keyword stream 2020-01-13 to 2020-06-06*. Version 1. Zenodo. DOI: 10.5281/zenodo.4267033. URL: https://doi.org/10.5281/zenodo.4267033.

# 7 COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter

**Published as:**
Müller M., Salathé M., Kummervold P. (2020). COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter *arXiv preprint* arXiv:2005.07503

**Abstract**

In this work, we release COVID-Twitter-BERT (CT-BERT), a transformer-based model, pretrained on a large corpus of Twitter messages on the topic of COVID-19. Our model shows a 10–30% marginal improvement compared to its base model, BERT-LARGE, on five different classification datasets. The largest improvements are on the target domain. Pretrained transformer models, such as CT-BERT, are trained on a specific target domain and can be used for a wide variety of natural language processing tasks, including classification, question-answering and chatbots. CT-BERT is optimised to be used on COVID-19 content, in particular from social media.

## 7.1 Introduction

Twitter has been a valuable source of news and a public medium for expression during the COVID-19 pandemic. However, manually classifying, filtering and summarising the large amount of information available on COVID-19 on Twitter is impossible and has also been a challenging task to solve with tools from the field of machine learning and natural language processing (NLP). To improve our understanding of Twitter messages related to COVID-19 content as well as the analysis of this content, we have therefore developed a model called COVID-Twitter-BERT (CT-BERT)[I].

Transformer-based models have changed the landscape of NLP. Models such as BERT, RoBERTa and ALBERT are all based on the same principle – training bi-directional transformer models on huge unlabelled text corpuses (Vaswani et al. 2017; Devlin et al. 2018; Liu et al. 2019; Lan et al. 2019). This process is done using methods such as mask language modelling (MLM), next sentence prediction (NSP) and sentence order prediction (SOP). Different models vary slightly in how these methods are applied, but in general, all training is done in a fully unsupervised manner. This process generates a general language model that is then used as input for a supervised finetuning for specific language processing tasks, such as classification, question-answering models, and chatbots.

Our model is based on the BERT-LARGE (English, uncased, whole word masking) model. BERT-LARGE is trained mainly on raw text data from Wikipedia (3.5B words) and a free book corpus (0.8B words) (Devlin et al. 2018). Whilst this is an impressive amount of text, it still contains little information about any specific subdomain. To improve performance in subdomains, we have seen numerous transformer-based models trained on specialised corpuses. Some of the most popular ones are BIOBERT (J. Lee et al. 2020) and SCIBERT (Beltagy, Cohan, and Lo 2019). These models are trained using the exact same unsupervised training techniques as the main models (MLM/NSP/SOP). They can be trained from scratch, but this requires a very large corpus, so a more common approach is to start with the trained weights from a general model. In this study, this process is called domain-specific pretraining. When trained, such models can be used as replacements for general language models and be trained for downstream tasks.

## 7.2 Method

The CT-BERT model is trained on a corpus of 160M tweets about the coronavirus collected through the Crowdbreaks platform (Müller and Salathé 2019) during the period from January 12 to April 16, 2020. Crowdbreaks uses the Twitter filter stream API to listen to a set of COVID-

[I] `https://github.com/digitalepidemiologylab/covid-twitter-bert`

19-related keywords[II] in the English language. Prior to training, the original corpus was cleaned for retweet tags. Each tweet was pseudonymised by replacing all Twitter usernames with a common text token. A similar procedure was performed on all URLs to web pages. We also replaced all unicode emoticons with textual ASCII representations (e.g. `:smile:` for a smiley) using the Python emoji library[III]. In the end, all retweets, duplicates and close duplicates were removed from the dataset, resulting in a final corpus of 22.5M tweets that comprise a total of 0.6B words. The domain-specific pretraining dataset therefore consists of 1/7th the size of what is used for training the main base model. Tweets were treated as individual documents and segmented into sentences using the spaCy library (Honnibal and Montani 2017).

All input sequences to the BERT models are converted to a set of tokens from a 30,000-word vocabulary. As all Twitter messages are limited to 280 characters, this allows us to reduce the sequence length to 96 tokens, thereby increasing the training batch sizes to 1024 examples. We use a dupe factor of 10 on the dataset, resulting in 285M training examples and 2.5M validation examples. A constant learning rate of 2e-5, as recommended on the official BERT GitHub[IV] when doing domain-specific pretraining.

Loss and accuracy was calculated through the pretraining procedure. For every 100,000 training steps, we therefore save a checkpoint and finetune this towards a variety of downstream classification tasks. Distributed training was performed using Tensorflow 2.2 on a TPU v3-8 (128GB of RAM) for 120 h.

### 7.2.1 Evaluation

To assess the performance of our model on downstream classification tasks, we selected five independent training sets. Three of them are publicly available datasets, and two are from internal projects not yet published. All datasets consist of Twitter-related data.

#### COVID-19 Category (CC)

This dataset is a subsample of the data used for training CT-BERT, specifically for the period between January 12 and February 24, 2020. Annotators on Amazon Turk (MTurk) were asked to categorise a given tweet text into either being a personal narrative (33.3%) or news (66.7%). The annotation was performed using the Crowdbreaks platform (Müller and Salathé 2019).

[II] wuhan, ncov, coronavirus, covid, sars-cov-2
[III] `https://pypi.org/project/emoji/`
[IV] `https://github.com/google-research/bert`

#### Vaccine Sentiment (VS)

This dataset contains a collection of measles- and vaccination-related US-geolocated tweets collected between March 2, 2011 and October 9, 2016. The dataset was first used by Pananos et al. (Pananos et al. 2017), but a modified version from Müller et al. (Müller and Salathé 2019) was used here. The dataset contains three classes: positive (towards vaccinations) (51.9%), negative (7.1%) and neutral/others (41.0%). The neutral category was used for tweets which are either irrelevant or ambiguous. Annotation was performed on MTurk.

#### Maternal Vaccine Stance (MVS)

The dataset is from a so far unpublished project related to the stance towards the use of maternal vaccines. Experts in the field annotated the data into four categories: neutral (41.0%), discouraging (25.3%), promotional (43.9%) and ambiguous (14.3%). Each tweet was annotated threefold, and disagreement amongst the experts was resolved in each case by using a common scoring criterion.

#### Twitter Sentiment SemEval (SE)

This is an open dataset from SemEval-2016 Task 4: Sentiment Analysis in Twitter (Nakov et al. 2019). In particular, we used the dataset for subtask A, a dataset annotated fivefold into three categories: negative (15.7%), neutral (45.9%) and positive (38.4%). We make a small adjustment to this dataset by fully anonymising links and usernames.

#### Stanford Sentiment Treebank 2 (SST-2)

SST-2 is a public dataset consisting of binary sentiment labels, negative (44.3%) and positive (55.7%), within sentences (Socher et al. 2013). Sentences were extracted from a dataset of movie reviews (Pang and L. Lee 2005) and did not originate from Twitter, making SST-2 our only non-Twitter dataset.

The dataset split size is predefined for the SST-2 and SE datasets. For the SST-2 dataset, the test dataset is not released. For the other datasets, we aimed at a split of around 50%-30% between the training and development sets, leaving a test set of 20% which was not used in this work. Our intention was not to optimise the finetuned models but to thoroughly evaluate the performance of the domain-specific CT-BERT-model. We experimented with different numbers of epochs for each training dataset for BERT-LARGE (i.e. checkpoint 0 of CT-BERT) and selected the optimal one. We then used this number in subsequent experiments on the respective dataset. We ended with three epochs for SST-2, CC and SE, five epochs for VC and 10 epochs for MVC, all with a learning rate of 2e-05. The number of epochs was dependent on

both the size and balance of the categories. Larger and unbalanced sets require more epochs.

| Dataset | Classes | Train | Dev | Labels |
|---|---|---|---|---|
| COVID-19 Category (CC) | 2 | 3094 | 1031 | Personal \| News |
| Vaccine Sentiment (VC) | 3 | 5000 | 3000 | N \| Neutral \| Positive |
| Maternal Vaccine Stance (MVS) | 4 | 1361 | 817 | Disc \| A \| N \| Promotional |
| Stanford Sentiment Treebank 2 (SST-2) | 2 | 67,349 | 872 | Negative \| Positive |
| Twitter Sentiment SemEval (SE) | 3 | 6000 | 817 | Neg \| Neutral \| Positive |

Table 7.1: Overview of the evaluation datasets. All five evaluation datasets are multi-class datasets with sometimes strong label imbalance, visualised by the proportional bar width in the label column. N and Neg stand for negative; Disc and A stand for discouraging and ambiguous, respectively.

## 7.3 Results

### 7.3.1 Domain-sepcific pretraining

Figure 7.1 shows the progress of pretraining CT-BERT at intervals of 25k training steps and the evaluation of 1k steps on a held-out validation dataset. All metrics considered improve throughout the training process. The improvement on the MLM loss task is most notable and yields a final value of 1.48. The NSP task improves only marginally, as it already performs very well initially. Training was stopped at 500,000, an equivalent of 512M training examples, which we consider as our final model. This corresponds to roughly 1.8 training epochs. All metrics for the MLM and NLM tasks improve steadily throughout training. However, using loss/metrics for these tasks to evaluate the correct time to stop training is difficult.

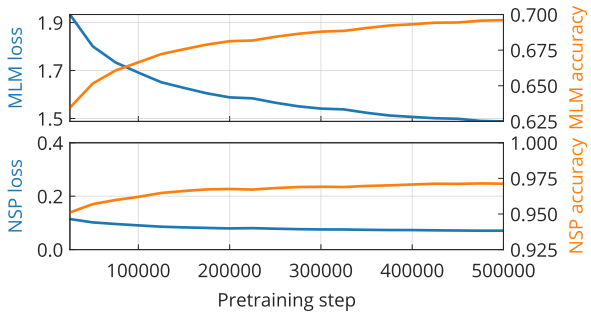



Figure 7.1: Evaluation metrics for the domain-specific pretraining of CT-BERT. Shown are the loss and accuracy of masked language modelling (MLM) and next sentence prediction (NSP) tasks.

### 7.3.2 Evaluation on classification datasets

To assess the performance of our model properly, we compared the mean F1 score of CT-BERT with that of BERT-LARGE on five different classification datasets. We adapted the number of training epochs for each dataset according to its size in order to have a similar number of training steps for each dataset. Our final model shows higher performance on all datasets (a mean F1 score of 0.833) compared with BERT-LARGE (a mean F1 score of 0.802). As the initial performance varies widely across datasets, we compute the relative improvement in marginal performance (ΔMP) for each dataset. ΔMP is calculated as follows:

$$\Delta\text{MP} = \frac{F_{1,\,\text{BERT-LARGE}} - F_{1,\,\text{CT-BERT}}}{1 - F_{1,\,\text{BERT-LARGE}}}$$

From this metric, we can observe the largest improvement of our model on the COVID-19-specific dataset (CC), with a ΔMP value of 25.88%. The marginal improvement is also high on the Twitter datasets related to vaccine sentiment (MVS). Our model likewise shows some improvements on the SST-2 and SemEval datasets, but to a smaller extent.

| Dataset | BERT-LARGE | CT-BERT | ΔMP |
|---|---|---|---|
| COVID-19 Category (CC) | 0.931 | 0.949 | 25.88% |
| Vaccine Sentiment (VC) | 0.824 | 0.869 | 25.27% |
| Maternal Vaccine Stance (MVS) | 0.696 | 0.748 | 17.07% |
| Stanford Sentiment Treebank 2 (SST-2) | 0.937 | 0.944 | 10.67% |
| Twitter Sentiment SemEval (SE) | 0.620 | 0.654 | 8.97% |
| Average | 0.802 | 0.833 | 17.57% |

Table 7.2: Comparison of the final model performance with BERT-LARGE. CT-BERT shows improvements on all datasets. The marginal improvement is the highest on the COVID-19-related dataset (CC) and lowest on the SST-2 and SemEval datasets.

### 7.3.3 Evaluation on intermediary pretraining checkpoints

So far, we have seen improvements in the final CT-BERT model on all evaluated datasets. To understand whether the observed decrease in loss during pretraining linearly translates into performance on downstream classification tasks, we evaluated CT-BERT on five intermediary versions (checkpoints) of the model and on the zero checkpoint, which corresponds to the original BERT-LARGE model. At each intermediary checkpoint, 10 repeated training runs (finetunings) for each of the five datasets were performed, and the mean F1 score was recorded. Figure 7.2 shows the marginal performance increase (ΔMP) at specific pretraining steps. Our experiments show that downstream performance increases fast up to step 200k in the pretraining and only demonstrates marginal improvement afterwards. The loss curve, on the other hand, shows a gradual increase even after step 200k. We also note that for the COVID-19-related dataset, most of the marginal improvement occurred after 100k pretraining steps. SST-2, the only non-Twitter dataset, improves much more slowly and reaches its final performance only after 200k pretraining steps.

Amongst runs on the same model and dataset, some degree of variance in performance was observed. This variance is mostly driven by runs with a particularly low performance. We observe that the variance is dataset dependent, but it does not increase throughout different pretraining checkpoints and is comparable to the variance observed on BERT-LARGE (pretraining step zero). The most stable training seems to be on the SemEval training set, and the least stable one is on SST-2, but most of this difference is within the error margins.

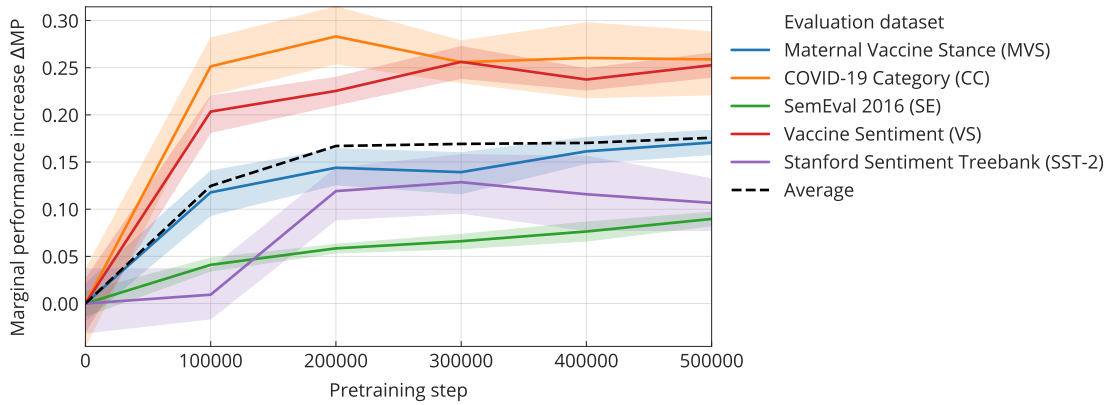



Figure 7.2: Marginal performance increase in the F1 score (ΔMP) on finetuning on various classification tasks at increasing steps of pretraining. Zero on the x-axis corresponds to the base model, which is BERT-LARGE in this case. Our model improves on all evaluated datasets, with the biggest relative improvement being in the COVID-19 category dataset. The bands show the standard error of the mean (SEM) out of 10 repeats.

## 7.4 Discussion

The most accurate way to evaluate the performance of a domain-specific model is to apply it on specific downstream tasks. CT-BERT is evaluated on five different Twitter-based datasets. Compared to BERT-LARGE, it improves significantly on all datasets. However, the improvement is largest in datasets related to health, particularly in datasets related to COVID-19. We therefore expect CT-BERT to perform similarly well on other classification problems on COVID-19-related data sources, but particularly on text derived from social media platforms.

Whilst it is expected that the benefit of using CT-BERT instead of BERT-LARGE is greatest when working with Twitter COVID-19 text, it is reasonable to expect some performance gains even when working with general Twitter messages (SemEval dataset) or with a non-Twitter dataset (SST-2).

Our results show that the MLM and NSP metrics during the pretraining align to some degree with downstream performance on classification tasks. However, compared with COVID-19 or health-related content, out-of-domain text might require longer pretraining to achieve a similar performance boost.

Whilst we have observed an improvement in performance on classification tasks, we did not test our model on other natural language understanding tasks. Furthermore, at the time of this paper's writing, we only had access to one COVID-19-related dataset. The general performance of our model might be improved further by considering pretraining under different hyperparameters, particularly modifications to the learning rate schedules, training batch sizes and optimisers. Future work might include evaluation on other datasets and the inclusion of more recent training data.

The best way to evaluate pretrained transformer models is to finetune them on downstream tasks. Finetuning a classifier on a pre-trained model is considered computationally cheap. The training time is usually done in an hour or two on a GPU. Using this method for evaluation is more expensive, as it requires evaluating multiple checkpoints to monitor improvement and on several varied datasets to show robustness. As finetuning results vary between each run, each experiment must be performed multiple times when the goal is to study the pretrained model. In this case, we repeated the training for six checkpoints, 10 runs for each checkpoint on all the five datasets. A total of 300 evaluation runs were performed. The computational cost for evaluation is therefore on par with the pretraining. Large and reliable training and validation sets make this task easier, as the number of repetitions can be reduced.

All the tests are done on categorisation tasks, as this task is easier in terms of both data access and evaluation. However, transformer-based models can be used for a wide range of tasks, such as named entity recognition and question answering. It is expected that CT-BERT can

also be used for these kinds of tasks within our target domain.

Our primary goal in this work was to obtain stable results on the finetuning in order to evaluate the pre-trained model, not to necessarily optimise the finetuning. The number of finetuning epochs and the learning rate, for instance, are optimised for BERT-LARGE, not for CT-BERT. This means that there is still great room for optimisation on the downstream task.

## Data Availability

The model, code and public datasets are available in our GitHub repository: `https://github.com/digitalepidemiologylab/covid-twitter-bert`.

## Funding

PK received funding from the European Commission for the call H2020-MSCA-IF-2017 and the funding scheme MSCA-IF-EF-ST for the VACMA project (grant agreement ID: 797876).

MM and MS received funding through the Versatile Emerging infectious disease Observatory grant as a part of the European Commission's Horizon 2020 framework programme (grant agreement ID: 874735).

The research was supported with Cloud TPUs from Google's TensorFlow Research Cloud and Google Cloud credits in the context of COVID-19-related research.

## Conflicts of Interest

The authors have no conflicts of interest to declare.

## References

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30, pp. 5998–6008.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Lan, Zhenzhong et al. (2019). "Albert: A lite bert for self-supervised learning of language representations". In: *arXiv preprint arXiv:1909.11942*.

Lee, Jinhyuk et al. (2020). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36.4, pp. 1234–1240.

Beltagy, Iz, Arman Cohan, and Kyle Lo (2019). "Scibert: Pretrained contextualized embeddings for scientific text". In: *arXiv preprint arXiv:1903.10676*.

Müller, Martin M and Marcel Salathé (2019). "Crowdbreaks: Tracking health trends using public social media data and crowdsourcing". In: *Frontiers in public health* 7, p. 81.

Honnibal, Matthew and Ines Montani (2017). "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* 7.1.

Pananos, A Demetri et al. (2017). "Critical dynamics in population vaccinating behavior". In: *Proceedings of the National Academy of Sciences* 114.52, pp. 13762–13767.

Nakov, Preslav et al. (2019). "SemEval-2016 task 4: Sentiment analysis in Twitter". In: *arXiv preprint arXiv:1912.01973*.

Socher, Richard et al. (2013). "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Pang, Bo and Lillian Lee (2005). "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales". In: *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, pp. 115–124.

# 8 Discussion

I have presented three main parts in this thesis: (i) the Crowdbreaks platform as a way to overcome concept drift (chapters 2–3) (ii) several applications of this platform (chapters 4–6) and (iii) a domain-specific BERT model for analysing Twitter content during COVID-19 (chapter 7). I will first summarize the principal findings of each chapter (section 8.1). Afterwards, I will discuss the results both from a methodological perspective by discussing the Crowdbreaks platform (section 8.2) and from a broader public health perspective (section 8.3). I will end the discussion with open challenges and an outlook for the field in general.

## 8.1 Principal findings

In chapter 2, the Crowdbreaks platform is described from a technical point of view and the use case of vaccine sentiment tracking is presented. Crowdbreaks can be seen as an active learning framework which is leveraging a crowdsourced annotation approach. In contrast to traditional research workflows, projects on Crowdbreaks are running over extended periods of time, which allows to trace opinions or health trends over multiple years while keeping algorithms up-to-date. The work outlines some of the logical building blocks in order to achieve this goal. The justification for building such a platform is posing a "chicken or egg"-type of problem: Since in order to have real-world evidence of phenomena like concept drift and develop strategies against it, we first have to build a platform to collect data, run annotations, train algorithms and see how they work on newly collected data.

Chapter 3 provides the result of such analysis in the context of vaccine sentiment. We find that indeed concept drift occurs in our dataset. The topical shift induced by the world-wide COVID-19 pandemic have had severe impacts on model performance. We also find that this impact would've been strong enough to miss a declining sentiment trend and therefore would've led to wrong conclusions in downstream analysis. In this work we also attempt to

build a better understanding of the nature of the observed drift. We explain differences in initial performance with semantic ambiguity and find correlations between concept drift and the dissimilarity of corpus vectors.

Chapter 4 focuses on a historical analysis of opinions surrounding the novel gene editing technology CRISPR/Cas9 on Twitter. By training an algorithm to detect stance towards this novel technology we are able to trace a long-term historic sentiment trend back to 2013, when the first application of CRISPR/Cas9 was announced. Our results overlap with surveys which conclude that CRISPR/Cas9 was and is predominantly viewed in a positive light. However, we find indications that public opinion may have declined, due to a series of recent scandals and that the general public is predominantly exposed to the topic of CRISPR when it is discussed in a negative light. We are able to further explain the sentiment trends by categorizing the data by organism and theme and are able to confirm existing survey results on these topics.

In chapters 5 and 6 we show that the COVID-19 pandemic may have fundamentally changed who we listen to or how we interact on social media. Chapter 5 lays the foundation for this work by introducing a taxonomy of occupational categories of users based on their profile description. By leveraging the complete COVID-19 Twitter stream in all languages and a large-scale crowdsourcing effort, we are able to generate a temporal picture of who paid attention to whom in the first 5 months of the pandemic. We find that accounts related to Science, Healthcare and Politics & Government received the largest boost in engagement early on in the pandemic. We show that, on a global level, Twitter users turned predominantly to experts in a time of crisis. In the context of what has been called a social media "infodemic", this result may come as a surprise.

In chapter 6 we consider a large dataset of 354M English tweets about COVID-19 and construct a retweet network. We run a community detection algorithm on this network and are able to characterize 15 distinct user communities. By leveraging the classifier developed in chapter 5 and the internationality of users we are able to characterize the communities. We then group the communities into 4 super-communities (namely international experts, national elite, political, and other) and quantify both internal and external attention patterns between these 4 super-communities. Our results emphasize the role of international experts in the beginning of the crisis, confirming the results from the previous chapter on a network level. However, as the crisis continued in April, our analysis reveals that the international expert community became increasingly segregated and most of the external attention has been allocated to the political super-community. This shift is reflecting a growing politicization of the debate, and is indicative of the fact that the pandemic has moved to more local grounds. Our work may therefore lay the basis for the development of a coherent communication strategy by local and global actors.

Lastly, in chapter 7, we introduce COVID-Twitter-BERT, a domain-specific language model

which may be used in various downstream NLP tasks on COVID-19-related Twitter data. The work describes the first version of this model, which was trained on a complete dataset of English COVID-19-related tweets up to April 16, 2020. An improved version of the model with training data up until July 5, 2020 has since been released on the official GitHub repository[I]. Based on the evaluated datasets the model achieves significant improvement over the BERT-Large base model. The model has since been used by the research community for various NLP tasks on Twitter data, such as filtering for informative content (Nguyen et al. 2020), identification of tweets which are worth fact-checking (Shaar et al. 2020; Alkhalifa et al. 2020), and misinformation detection (Hossain et al. 2020).

## 8.2 A digital framework for social media studies

Throughout the previously listed work we used and developed the Crowdbreaks platform to collect, analyze and interpret this data. In this section I will elaborate on the ways we used the platform to overcome common technical challenges in these projects. Furthermore, I will outline which of these challenges remain and how we might want to address them in future work.

### 8.2.1 Real-time data collection

The previously listed results are based on the analysis of more than half a billion tweets or roughly 2.6TB of compressed data. Given the trends in social media adoption, it is safe to assume that the quantities of data will only increase and the next generation of researchers will likely be faced with datasets many factors larger. It is already obvious from these numbers that data collection cannot be run from an ordinary laptop anymore. Therefore the Crowdbreaks platform heavily depends on cloud infrastructure. Crowdbreaks addresses the challenges that appear when collecting data in real-time, such as the buffering of spikes, scaling the system based on different load, system monitoring, and crash recovery.

### 8.2.2 Collection of annotation data

As outlined in the introduction, the collection of annotation data is of key importance as it lays the ground truth data for any signal that is inferred by a future classifier. Crowdbreaks represents the annotation procedure as a question tree, in which multiple questions are asked about a single tweet. This allows for the explorations of more complex or nuanced categorizations and allows for more efficient annotations. Questions related to possible ambiguities or uncertainties in the annotation task may allow to generate better ground truth

[I] `https://github.com/digitalepidemiologylab/covid-twitter-bert`

datasets.

The process of annotation is often seen as a one-time effort, usually involving the crowdsourcing of the task by recruiting a large number of so-called "crowdworkers" on platforms such as Amazon Turk (MTurk). On Crowdbreaks, this view is challenged by running a continuous annotation process and also allowing anyone in the public to participate in this annotation process. The fact that continuous annotation is needed as a way to monitor and possibly overcome concept drift has been shown in chapter 3. However, the question whether public users could provide enough annotation data to overcome concept drift could not be addressed in this work and at current time the majority of annotations are still collected through MTurk.

The work in chapter 3 suggests that, for the vaccine sentiment task, 300 newly annotated tweets per month would likely be sufficient to overcome mild drift for a simple FastText model. We have also found that concept drift affects classes differently, and that anti-vaccine content drifts faster. Such insights could possibly be integrated into the query strategy for label selection in order to overcome concept drift with fewer annotations. Overall, these results point to the direction that concept drift could be overcome already with limited engagement by the public.

Similar to other citizen science projects, future work will be needed to answer how the quality of public annotations compares to annotations collected through MTurk. Furthermore, improvements in incentivizing user engagement are required, possibly by making the platform more interactive and educational.

### 8.2.3 Training of models

Automatic re-training without human intervention remains to be a challenging technical problem. This is because machine learning is still a craft which relies on careful human tweaking. However, manual tweaking of models clearly becomes infeasible when hundreds of models are run in parallel and need to be updated on a regular basis.

Since the first release of the Crowdbreaks platform as described in chapter 2, the landscape of natural language processing (NLP) has already drastically changed. Most of the advancements in the field have been a consequence of making use of large volumes of raw text (meaning non-annotated, unstructured text) in unsupervised pre-training. We have seen in chapter 7 that by further domain-specific pre-training (DSP) of BERT models, performance increases of 10–30% can be expected.

Multiple strategies for leveraging DSP in the context of Crowdbreaks can be envisioned and need to be explored. A possible strategy is the pre-training of a general Twitter model which can then be further adapted to a specific project-level sub-domain. Future work will have to

show whether the DSP technique could also be used to overcome concept drift by continuous small steps of adaptation to newly collected data. Important engineering challenges remain also in deploying such large models and using them for real-time inference.

### 8.2.4 Automation

A researcher who enters the field and starts analyzing social media data is faced with the typical workflow ranging from collection of data, annotation, training a machine learning model, and prediction of this data. The technical obstacles along this path are usually underestimated by researchers. Addressing these issues from scratch leaves less time for the interpretation of results. In the worst case the research activities will take longer than the traditional research workflow, at which point an important advantage of digital methods is negated. Simply, ignoring the issues can have negative impacts on research quality and in general reflect badly on the field as a whole.

Many of the discussed aspects of Crowdbreaks are therefore challenges of automation and standardization of research workflows. These tasks may sound mundane to some researchers, but are often key to research quality and proper interpretation of results. By standardizing complex research workflows Crowdbreaks is able to address the issues of reproducibility in the field.

## 8.3 From social media signals to public health decision making

In this work we have measured three different signals from social media data. In chapters 2 and 3 we have analyzed vaccine sentiment, chapter 4 investigated sentiment towards CRISPR, and in chapter 5 and 6 we have looked at attention patterns on social media. Such signals reflect how users behave online and, perhaps more importantly, what information they are exposed to. It is important to keep in mind that such signals, if measured correctly, have validity in themselves and are therefore already useful to public health.

But even if the measured signals are useful to public health, how can we translate these signals into actionable advice on which we can base public health decisions? This question points to how signals on social media relate to indicators measured outside of that system and how they can be validated.

### 8.3.1 Validation of signals

Vaccine sentiment is different from the signals analyzed in other chapters in that it may be compared to a measurable public health indicator, i.e. vaccine uptake. A temporal and geo-

graphical link between vaccine sentiment on Twitter and vaccine uptake could be established in previous studies (Salathé and Khandelwal 2011; Huang et al. 2017; Bello-Orgaz, Hernandez-Castro, and Camacho 2017). Although not all studies find a clear correlation (Brooks 2014) and future work on the topic is still required, it is fair to say that, overall, there is good evidence that vaccine sentiment trends correlate with vaccine uptake. At the time of writing, it is too early to answer whether the decline in vaccine sentiment during the COVID-19 pandemic is a precursor for lower vaccination rates for SARS-CoV-2 or other viruses.

In chapter 4 we find agreement between measured sentiment trends towards CRISPR and a limited number of surveys on the topic. This result gives some validation that the measured sentiment could be used as a proxy for public opinion on the topic. It is important to keep in mind that, like social media data, surveys and polls have a number of biases. A disagreement with survey data would therefore not necessarily invalidate the CRISPR sentiment trend.

In chapters 5 and 6 we measure attention patterns towards different groups during the COVID-19 pandemic. In this case the signal does not act as a proxy for a health behavior but reflects the behavior itself, i.e. the act of paying attention online. Attention behaviors have very direct implications for public health officials' ability to communicate to the public. As a consequence, it is conceivable that attention to health officials also correlates with the public's adherence to control measures. However, further research is required to understand how attention phenomena link to health behaviors.

In conclusion, we find that the observed signals are of very different nature and may therefore be used very differently in the public health context. For certain signals such as vaccine sentiment a validation with ground truth data is certainly required. Other signals might not necessarily require outside validation in order to be actionable for public health.

Nevertheless, all signals would in principle benefit from validation on external indicators as they allow us to better understand and interpret trends. However, often such external indicators are published at a delay and the interpretation of social media trends will have to occur in absence of corroborating external indicators or ground truth data.

Trends predicted by machine learning models can be more trustworthy if we can be sure that they are not an artefact or a consequence of concept drift, as explored in chapter 3. Human annotation of a subset of newly collected data through a platform like Crowdbreaks might therefore serve as a form of internal validation for the predicted trends. This idea has also been suggested by (Tufekci 2014) who recommends "qualitative pull-outs" of data in order to validate the correctness of the analysis. Additionally, sentiment trends across multiple social media platforms as well as a comparison to other sources of behavioral data may be leveraged to further validate trends.

### 8.3.2 Interpreting shifts

The validation of these signals relies on being able to interpret shifts or changes in signals from social media. In all three previously discussed cases we have observed shifts: In chapter 3, we have found preliminary evidence of a declining sentiment towards vaccines during the COVID-19 pandemic. In chapter 4 we have analysed the impacts of scandals, such as the CRISPR babies, and have observed negative spikes in sentiment. In chapters 5 and 6, we were able to study shifts in attention patterns during COVID-19, first towards scientific experts and later towards political leaders.

In order to observe a shift or change a sufficiently long baseline observation period is required. Observation periods smaller than one year may be considered even problematic due to the possible influence of seasonality effects. In general, we can conclude that signals from social media become more trustworthy and interpretable the longer we observe them. This is also because traditional health data, which could be used for validation purposes, are published at a much slower pace compared to observations on social media. This further underlines the need for the long-term analysis of such trends through a platform like Crowdbreaks.

### 8.3.3 Sentiment signals

In the previous section we compared vaccine sentiment to vaccine uptake rates and evaluated whether CRISPR sentiment could be used as a proxy for public opinion. The sentiment signal in these studies is a simple mean between negative (-1), neutral (0), and positive (+1) tweets. A common concern is that such classes are too simplistic in nature. It is therefore important to critically consider whether sentiment is a suitable measure for representing opinions.

The word sentiment might be misleading since a tweet can be in favor of the topic, therefore should be labelled positive, but express a negative sentiment. For this reason instead of sentiment the term stance is often used in literature. However, one could argue that from the perspective of other users, the exposure to negative sentiment might lead to a negative connotation of the topic, even if the tweet was arguing in favor of the topic. The "error" we are making by wrongly assessing the user's true opinion might turn out to be less significant.

Studies on vaccination-related social media data show that such data can provide insight into a wide range of concerns, beliefs or misconceptions (Larson et al. 2013). Clearly, the categorization of such data into positive, neutral and negative classes is therefore a gross simplification of opinions and means we are not utilizing this data to its fullest potential. It is therefore important to note that sentiment is only a first-order approximation. However, due to being able to attach a numerical value to its labels, it allows for a simple interpretation and comparison to signals outside of social media.

## 8.4 Open challenges

Although the Crowdbreaks platform addresses many existing limitations in the field, there are plenty of challenges remaining. Going into the details of these limitations would be beyond the scope of this work but it is nevertheless important to mention them.

### 8.4.1 Biases

Although biases appear in almost any measurement system, it is fair to acknowledge that traditional systems have had a longer time to understand and correct for known biases. Being aware of all biases in social media studies can therefore help to interpret the observed signals. Addressing these issues is also linked to the previously discussed validation of methods.

#### Biases in data collection

The presented projects make use of keyword-based filtering. Data might either be wrongly collected or collected by mistake depending on the choice of keywords. The filtering process should therefore not only be reliant on keyword based filtering but also a filtering by relevance. In chapter 4 we have successfully used a relevance classifier to address for this bias.

#### Demographic biases

A frequently discussed bias is the bias in user demographics which may not overlap with the population of interest (Mislove et al. 2011). However, as discussed in the introduction, due to the further adoption of social media across age groups the demographic biases might be less problematic today. Still, certain demographic groups may be less willing to share information on a topic, therefore introduce bias.

#### Reporting biases

Fundamentally, we can only analyze what users report. Users may report differently on certain topics due to social stigma, which may lead to certain opinions not voiced in fear of public backlash (Tufekci 2014). If work is conducted on topics with severe social stigma it is important to address this bias.

#### Algorithmic biases

In this work we have made frequent use of machine learning algorithms to predict sentiment, relevance, and user categories. It is fair to assume that the prediction errors are non-randomly

distributed (Caliskan, Bryson, and Narayanan 2017). Although bias is a general problem in all domains of machine learning, recent NLP models have revealed significant racial discrimination (Sweeney 2013) and gender bias (Lu et al. 2020). Active research is conducted on how such models can be "debiased" (Bolukbasi et al. 2016).

#### System drift

A fundamental challenge with research using Twitter (or any third-party) data is the lack of knowledge about the data generation process. Research has mostly focused on the degree of randomness of the sampling process on Twitter's end (Morstatter, Pfeffer, Liu, and Carley 2013; Morstatter, Pfeffer, and Liu 2014). However, more concerning are sudden or slow changes to these systems, which could be difficult to detect or overcome.

### 8.4.2 Ethical challenges

Even though Twitter users agreed to terms of service that their data is publicly available when they signed up on the platform, they might (a) not be fully aware of this or (b) not consenting for their data to be used for research. The question of consent may also depend on the subject matter that is discussed (Hudson and Bruckman 2004). Surveys of Twitter users in the context of studying mental health has shown that, given the results were aggregated, users were overall positive towards research being done on their data (Mikal, Hurst, and Conway 2016). Nevertheless, the field is in an active debate on the questions of whether informed consent is required and what the users' expectation of privacy are (Vayena et al. 2015; Kostkova 2018). Due to these reasons many researchers decide to treat public Twitter data as private. As it stands today, the field is in need for further clarity in terms of ethical guidelines and best research practices (O'Connor 2013). The ultimate goal is to find a balance between the right to user privacy and the potential for such data to be used for public good (Ienca et al. 2018).

### 8.4.3 Limitations to data sharing

Although Twitter data is publicly available, the sharing of this data beyond the tweets' identifiers is not permitted. This mechanism allows users to have their content deleted, should they wish to do so. It does however negatively impact the reproducibility of research with Twitter data, since researchers will likely not be able to recreate the exact same dataset in the future. However, reproducibility of such analysis is technically still possible if research teams collect data using the same keyword list.

## 8.5 Outlook

The recent breakthroughs in the field of natural language processing as well as the global COVID-19 pandemic have been major topics of this work. The impacts of these two developments on the field of digital epidemiology cannot yet be fully grasped, however both have the potential to fundamentally change the field.

We are only at the beginning of fully exploiting the recent advances in NLP by incorporating them into systems like Crowdbreaks. An improvement in model performance should automatically lead to higher trustworthiness and reliability of results. Improved categorization and filtering will eventually translate into more accurate signals and further reduce our dependence on human-annotated data. Nevertheless, we are faced with new engineering challenges to deploy such large models and, due to the increased dependency on latent representations of text, we may also be faced with a new set of ethical challenges. Social media usage has further increased across age groups, becoming the central means of communication during worldwide lockdowns. Health behaviors such as wearing a mask, hygiene measures, or vaccinations are now actively debated online. COVID-19 therefore represents a unique opportunity to study a variety of online signals and evaluate them as possible inputs to mathematical models for disease dynamics.

The real benefits of the Crowdbreaks platform will play out over time, both in terms of long-term observations as well as in terms of acceleration of research and their adoption in the field of public health.

## References

Nguyen, Dat Quoc et al. (2020). "WNUT-2020 Task 2: Identification of Informative COVID-19 English Tweets". In: *arXiv preprint arXiv:2010.08232*.

Shaar, Shaden et al. (2020). "Overview of CheckThat! 2020 English: Automatic identification and verification of claims in social media". In: *Cappellato et al.[10]*.

Alkhalifa, Rabab et al. (2020). "QMUL-SDS at CheckThat! 2020: determining COVID-19 tweet check-worthiness using an enhanced CT-BERT with numeric expressions". In: *arXiv preprint arXiv:2008.13160*.

Hossain, Tamanna et al. (2020). "COVIDLIES: Detecting COVID-19 Misinformation on Social Media". In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Salathé, Marcel and Shashank Khandelwal (2011). "Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control". In: *PLoS Comput Biol* 7.10, e1002199.

Huang, Xiaolei et al. (2017). "Examining Patterns of Influenza Vaccination in Social Media." In: *AAAI Workshops*.

Bello-Orgaz, Gema, Julio Hernandez-Castro, and David Camacho (2017). "Detecting discussion communities on vaccination in twitter". In: *Future Generation Computer Systems* 66, pp. 125–136.

Brooks, Benjamin (2014). "Using Twitter data to identify geographic clustering of anti-vaccination sentiments". PhD thesis.

Tufekci, Zeynep (2014). "Big questions for social media big data: Representativeness, validity and other methodological pitfalls". In: *arXiv preprint arXiv:1403.7400*.

Larson, Heidi J et al. (2013). "Measuring vaccine confidence: analysis of data obtained by a media surveillance system used to analyse public concerns about vaccines". In: *The Lancet infectious diseases* 13.7, pp. 606–613.

Mislove, Alan et al. (2011). "Understanding the demographics of twitter users." In: *Icwsm* 11.5th, p. 25.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334, pp. 183–186.

Sweeney, Latanya (2013). "Discrimination in online ad delivery". In: *Queue* 11.3, pp. 10–29.

Lu, Kaiji et al. (2020). "Gender bias in neural natural language processing". In: *Logic, Language, and Security*. Springer, pp. 189–202.

Bolukbasi, Tolga et al. (2016). "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: *Advances in neural information processing systems* 29, pp. 4349–4357.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley (2013). "Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose". In: *arXiv preprint arXiv:1306.5204*.

Morstatter, Fred, Jürgen Pfeffer, and Huan Liu (2014). "When is it biased? Assessing the representativeness of twitter's streaming API". In: *Proceedings of the 23rd international conference on world wide web*, pp. 555–556.

Hudson, James M and Amy Bruckman (2004). ""Go away": participant objections to being studied and the ethics of chatroom research". In: *The Information Society* 20.2, pp. 127–139.

Mikal, Jude, Samantha Hurst, and Mike Conway (2016). "Ethical issues in using Twitter for population-level depression monitoring: a qualitative study". In: *BMC medical ethics* 17.1, p. 22.

Vayena, Effy et al. (2015). "Ethical challenges of big data in public health". In: *PLoS Comput Biol* 11.2, e1003904.

Kostkova, Patty (2018). "Disease surveillance data sharing for public health: the next ethical frontiers". In: *Life sciences, society and policy* 14.1, p. 16.

O'Connor, Dan (2013). "The apomediated world: regulating research when social media has changed research". In: *The Journal of Law, Medicine & Ethics* 41.2, pp. 470–483.

Ienca, Marcello et al. (2018). “Considerations for ethics review of big data health research: A scoping review”. In: *PloS one* 13.10, e0204937.

# A Supplementary Information: Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic

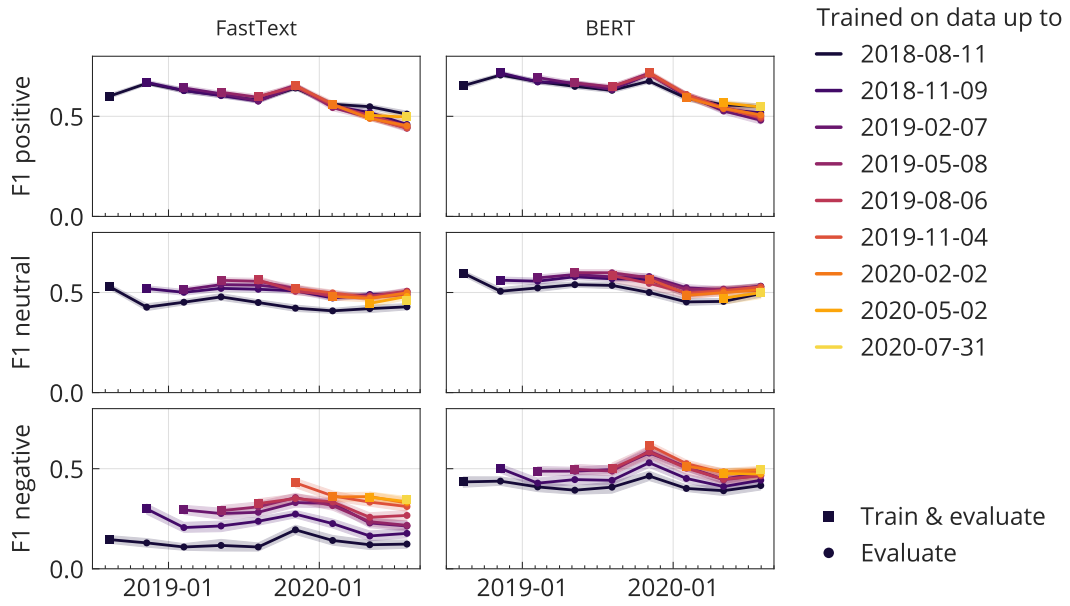



Figure A.1: Performance scores by class for FastText and BERT models. For an explanation of the Figure, please refer to Figure 3.2 in the main text. Unlike for the negative class, performance between FastText and BERT is comparable for the neutral and positive class. The "negative" class shows the strongest effects due to concept drift.

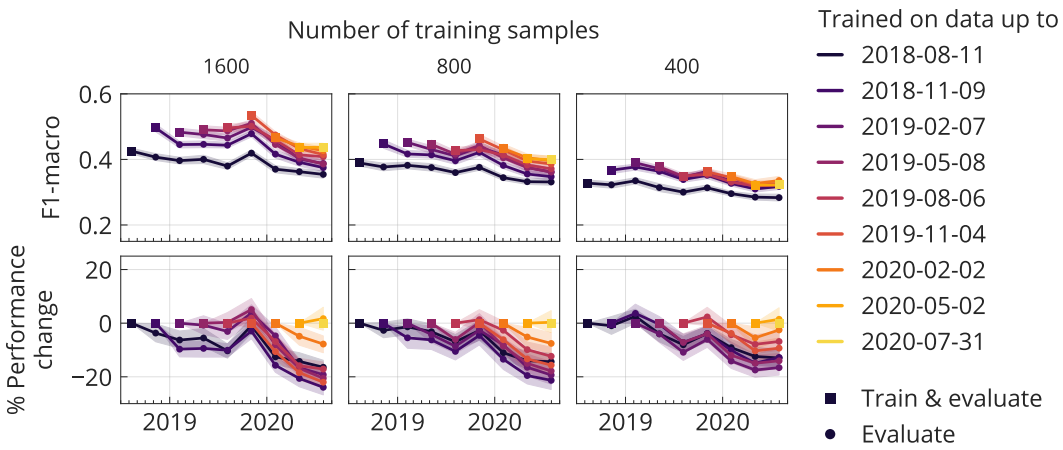



Figure A.2: Drift of FastText models depending on size of training data. The plots of the first column are identical to the FastText plots in Figure 3.2. For all experiments a training window length of 360 days was used. Initial performance is decreasing with a decreasing number of training samples. Overcoming concept drift is increasingly difficult, and is barely visible at 400 training samples.

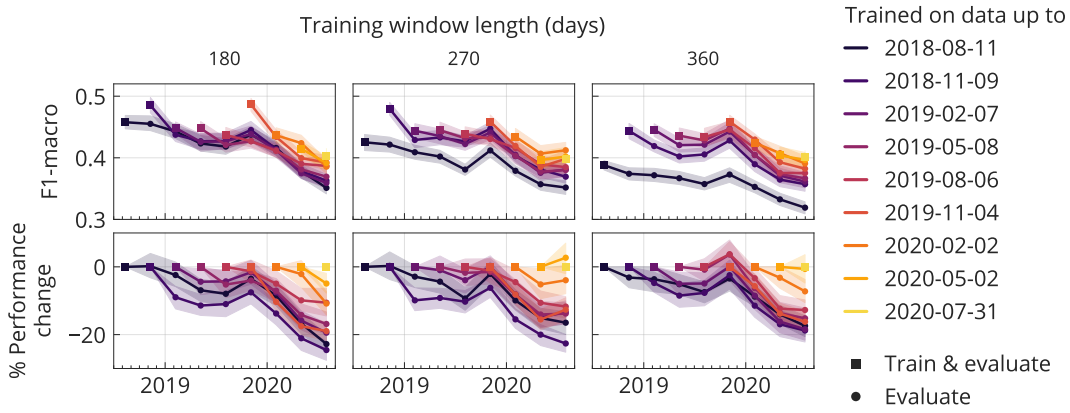



Figure A.3: Drift of FastText models depending on the length of the training window. Each model was trained on an equal number of 800 training examples, but distributed over 180, 270 or 360 days. A shorter training window is occasionally associated with slightly higher initial performance and slightly faster relative performance decrease on average.

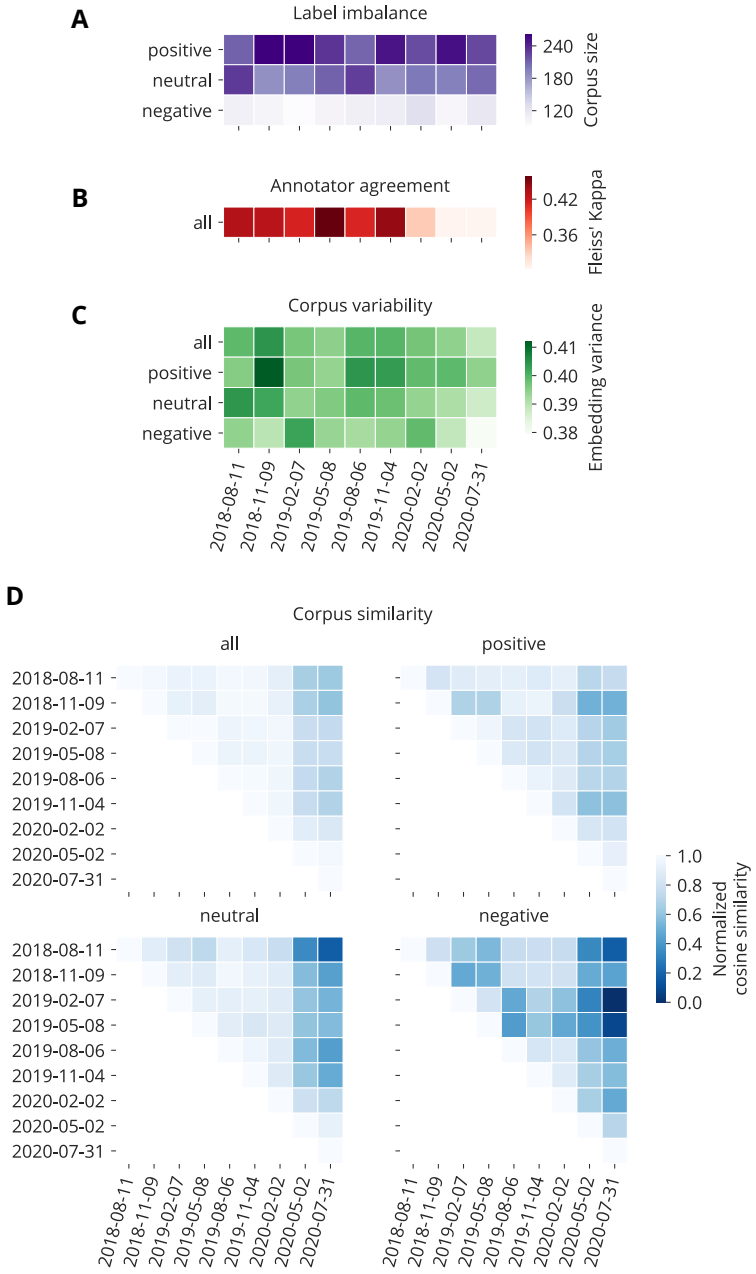



Figure A.4: This figure is equivalent to Figure 3.3 in the main text, except for the different datasets that were used. In Figure 3.3, we show the used training and evaluation set in the full time window. This figure shows the newly added training and evaluation data for each 90 day bin. For a detailed description of this figure please refer to Figure 3.3 in the main text.

# B Supplementary Information: Assessing Public Opinion on CRISPR/Cas9

| Year | Number of tweets |
|---|---|
| 2013 | 4818 |
| 2014 | 20,002 |
| 2015 | 131,211 |
| 2016 | 304,759 |
| 2017 | 437,931 |
| 2018 | 445,744 |
| 2019 | 163,579 (392,590) |
| **Total** | 1,508,044 |

Table B.1: Yearly counts. Number of tweets per year since January 1, 2013, until May 31, 2019. A steady increase in volume can be observed. In parentheses is the extrapolated number for 2019 (from the first five months).

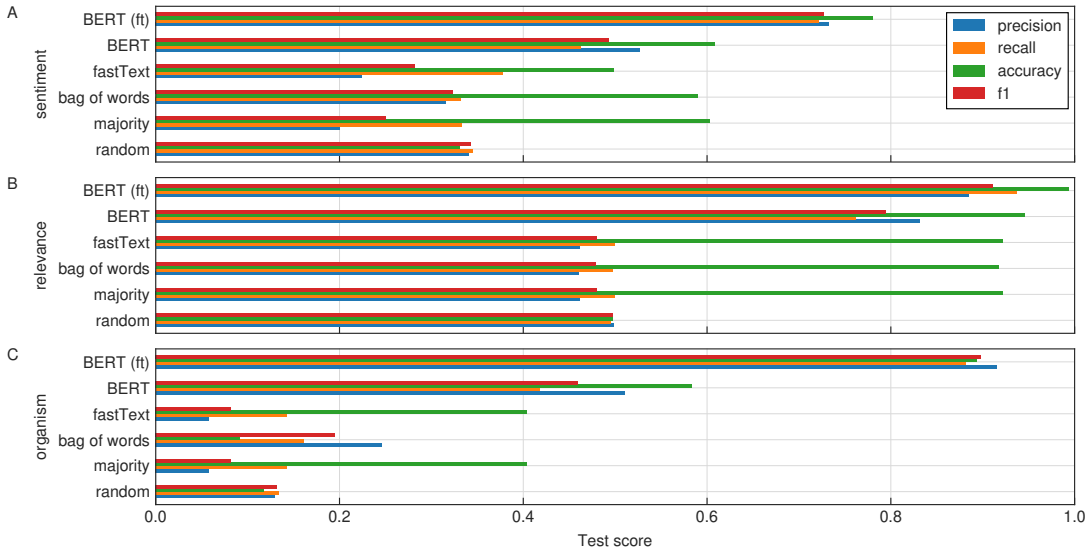



Figure B.1: Model performance. Classification scores for selected models. Subfigures A, B and C correspond to three different classifiers trained for sentiment, relevance and organism, respectively. The y-axis shows the best corresponding model for a specific model type after hyperparameter search was performed. The model types are random (pick a class at random), majority (always pick the most frequent class), bag of words, fastText, BERT and a fine-tuned version of BERT-large (denoted as BERT ft). The x-axis denotes the test performance scores of accuracy (green), and macro-averaged precision (blue), recall (orange) and F1 scores (red). The fine-tuned BERT model was the best performing model for all three classification problems irrespective of the metric used.

### Preliminary literature review search strategy and databases

Databases used: PubMed, Scopus, Web of science. Matching query in articles' title only:
`(( crispr OR gene-editing OR "genome editing" ) AND ( attitudes OR opinions OR perspectives OR believes OR reactions OR public ))`
103 publications were identified by the search (24 PubMed, 41 Scopus, 38 Web of Science). A total of 4 articles were included in the full-text analysis after duplicate removal and exclusion through abstract screening based on exclusion criteria:

- The article is not focussing on CRISPR
- The article is not referring to human subjects
- The article is not considering public opinions/attitudes
- The article is not an empirical study

Resulting documents:

- Blendon, R. J., Gorski, M. T., & Benson, J. M. (2016). The public and the gene-editing revolution. New England Journal of Medicine, 374(15), 1406-1411.
- McCaughey, T., Sanfilippo, P. G., Gooden, G. E., Budden, D. M., Fan, L., Fenwick, E., ... & Liang, H. H. (2016). A global social media survey of attitudes to human genome editing. Cell stem cell, 18(5), 569-572.
- Scheufele, D. A., Xenos, M. A., Howell, E. L., Rose, K. M., Brossard, D., & Hardy, B. W. (2017). US attitudes on human genome editing. Science, 357(6351), 553-554.
- Weisberg, S. M., Badgio, D., & Chatterjee, A. (2017). A CRISPR New World: Attitudes in the Public toward Innovations in Human Genetic Modification. Frontiers in Public Health, 5.

| Theme | Regular expression |
|---|---|
| disease | diseases? |
| health | restore |
| therapy | therapy\|therapeutic |
| germline | germline\|heritable\|stem[\s-]cell\|heritage |
| somatic | somatic |
| enhancement | enhanc(e\|ement\|ing) |
| improvement | improv(e\|ement\|ing) |
| treatment | treat(ment\|ing)? |
| reducing | (lower(ing)?\|reduc(e\|ing))\s.*risk |
| prevention | prevent(ion\|ing)? |
| risk | risks? |
| cure | cur(e\|ing) |
| progress | scientific progress |
| traits | traits? |
| abilities | abilit(y\|ies) |
| intelligence | intelligence |
| appearance | appearance |
| price | expensive |
| discovery | discovery?\|anticipat(e\|ion) |
| privacy | privacy |
| accuracy | accuracy |
| reliability | reliability |
| mutation | mutations? |
| eugenic | eugenic |
| trust | trust |
| children | child(ren)? |
| genome | genome\|genomics?\|genes?\|genetic |
| embryo | embryo(nic)? |
| baby | bab(y\|ies) |

Table B.2: Themes and regex patterns. Derived themes and corresponding regex patterns from preliminary literature review.

| Year | Sent | SD | Count | $P_{2013}$ | $P_{2014}$ | $P_{2015}$ | $P_{2016}$ | $P_{2017}$ | $P_{2018}$ | $P_{2019}$ | $T_{2013}$ | $T_{2014}$ | $T_{2015}$ | $T_{2016}$ | $T_{2017}$ | $T_{2018}$ | $T_{2019}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2013** | 0.96 | 0.29 | 2495 | – | < .001 | < .001 | < .001 | < .001 | < .001 | < .001 | – | 3.57 | −13.92 | −12.33 | −19.12 | −57.13 | −60.08 |
| **2014** | 0.98 | 0.20 | 10,249 | | – | < .001 | < .001 | < .001 | < .001 | < .001 | | – | −37.85 | −40.22 | −58.51 | −143.38 | −123.13 |
| **2015** | 0.87 | 0.49 | 59,840 | | | – | < .001 | < .001 | < .001 | < .001 | | | – | 5.16 | −11.89 | −100.74 | −90.12 |
| **2016** | 0.89 | 0.46 | 138,252 | | | | – | < .001 | < .001 | < .001 | | | | – | −23.31 | −133.50 | −106.77 |
| **2017** | 0.85 | 0.53 | 218,639 | | | | | – | < .001 | < .001 | | | | | – | −117.69 | −94.49 |
| **2018** | 0.62 | 0.78 | 253,307 | | | | | | – | < .001 | | | | | | – | −12.64 |
| **2019** | 0.58 | 0.81 | 100,566 | | | | | | | – | | | | | | | – |

| Organism | Sent | SD | Count | $P_{\text{animals}}$ | $P_{\text{bacteria}}$ | $P_{\text{embryos}}$ | $P_{\text{humans}}$ | $P_{\text{plants}}$ | $P_{\text{unspecified}}$ | $T_{\text{animals}}$ | $T_{\text{bacteria}}$ | $T_{\text{embryos}}$ | $T_{\text{humans}}$ | $T_{\text{plants}}$ | $T_{\text{unspecified}}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **animals** | 0.73 | 0.50 | 99,131 | – | < .001 | < .001 | < .001 | < .001 | < .001 | – | −41.7 | −175.18 | −145.33 | −52.39 | −182.95 |
| **bacteria** | 0.59 | 0.51 | 31,855 | | – | < .001 | < .001 | .038 | < .001 | | – | −99.11 | −46.78 | −2.07 | −61.81 |
| **embryos** | 0.22 | 0.58 | 55,908 | | | – | < .001 | < .001 | < .001 | | | – | 85.75 | 110.13 | 74.64 |
| **humans** | 0.45 | 0.70 | 396,878 | | | | – | < .001 | < .001 | | | | – | 54.34 | −29.72 |
| **plants** | 0.58 | 0.57 | 64,179 | | | | | – | < .001 | | | | | – | −74.09 |
| **unspecified** | 0.41 | 0.60 | 663,593 | | | | | | – | | | | | | – |

Table B.3: Mean sentiments and test statistics for years and organism classes. On the left, the table shows the mean sentiment (Sent) for each year and organism class based on the sentiments of the individual tweets. Further, the standard deviation (SD) and number (Count) of tweets for each group are reported. On the right, the $p$-values or the significance level if significant ($\alpha = 0.001$) and the $t$-values from Welch's $t$-test among the years and organism class means are shown. A value refers to the comparison between the classes given by its row and column labels. For example, the $p$-value for Welch's $t$-test for the difference between the mean of the bacteria class and the plants class is 0.038.

| # | Mark | Peak time | Event time | Event | Prominence |
|---|---|---|---|---|---|
| 1 | | 2015-12-03 | 2015-12-01 | First summit on human gene editing in Washington D.C. | 0.21 |
| 2 | | 2016-06-24 | 2016-06-22 | U.S. proposal for human trials passes safety reviews | 0.26 |
| 3 | a | 2016-11-18 | 2016-11-15 | First time use of CRISPR on humans in China | 0.34 |
| 4 | b | 2017-02-17 | 2017-02-15 | Broad Institute prevails in patent conflict | 0.33 |
| 5 | c | 2017-08-04 | 2017-08-02 | CRISPR successfully fixes a gene in viable human embryos | 0.44 |
| 6 | | 2018-01-21 | 2018-01-19 | Study on advances in CRISPR technology | 0.37 |
| 7 | d | 2018-07-19 | 2018-07-16 | Study shows the potential for side effects (e.g. deletions) of CRISPR | 0.29 |
| 8 | e | 2018-11-29 | 2018-11-26 | "CRISPR babies" scandal | 0.97 |
| 9 | f | 2019-02-04 | 2017-08-10 | Biohackers encode a malware program into DNA | 0.29 |

Table B.4: Identified events. Selected events with a peak prominence above 0.2. The marks correspond to the selected events in Figure 4.2 of the article. Peak times have been automatically detected as described in the methods section. The corresponding events have been inferred from visual inspection of the data.

| humans | | | | embryos | | | | animals | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | Month | Sent | SD | Year | Month | Sent | SD | Year | Month | Sent | SD |
| 2013 | 1 | – | – | 2013 | 1 | | | 2013 | 1 | – | – |
| | 2 | – | – | | 2 | | | | 2 | – | – |
| | 3 | | | | 3 | | | | 3 | | |
| | 4 | – | – | | 4 | | | | 4 | – | – |
| | 5 | – | – | | 5 | – | – | | 5 | – | – |
| | 6 | – | – | | 6 | | | | 6 | – | – |
| | 7 | – | – | | 7 | | | | 7 | – | – |
| | 8 | – | – | | 8 | | | | 8 | – | – |
| | 9 | – | – | | 9 | | | | 9 | – | – |
| | 10 | – | – | | 10 | | | | 10 | – | – |
| | 11 | 0.91 | 0.31 | | 11 | – | – | | 11 | – | – |
| | 12 | 0.63 | 0.48 | | 12 | – | – | | 12 | – | – |
| 2014 | 1 | 0.50 | 0.50 | 2014 | 1 | | | 2014 | 1 | 0.84 | 0.37 |
| | 2 | 0.79 | 0.41 | | 2 | – | – | | 2 | 0.50 | 0.50 |
| | 3 | – | – | | 3 | | | | 3 | 0.88 | 0.33 |
| | 4 | 0.64 | 0.49 | | 4 | | | | 4 | 0.86 | 0.38 |
| | 5 | 0.42 | 0.51 | | 5 | | | | 5 | – | – |
| | 6 | 0.62 | 0.50 | | 6 | – | – | | 6 | – | – |
| | 7 | 0.71 | 0.45 | | 7 | | | | 7 | 0.78 | 0.43 |
| | 8 | 0.88 | 0.33 | | 8 | | | | 8 | 0.87 | 0.35 |
| | 9 | 0.80 | 0.45 | | 9 | | | | 9 | 0.72 | 0.45 |
| | 10 | 0.64 | 0.48 | | 10 | | | | 10 | 0.73 | 0.44 |
| | 11 | 0.62 | 0.50 | | 11 | | | | 11 | 0.80 | 0.40 |
| | 12 | 0.84 | 0.36 | | 12 | | | | 12 | 0.43 | 0.51 |
| 2015 | 1 | 0.78 | 0.45 | 2015 | 1 | – | – | 2015 | 1 | – | – |
| | 2 | 0.80 | 0.40 | | 2 | – | – | | 2 | – | – |
| | 3 | 0.49 | 0.62 | | 3 | –0.24 | 0.50 | | 3 | 0.74 | 0.44 |
| | 4 | 0.36 | 0.57 | | 4 | 0.08 | 0.51 | | 4 | 0.61 | 0.50 |
| | 5 | 0.28 | 0.62 | | 5 | 0.03 | 0.36 | | 5 | 0.74 | 0.46 |
| | 6 | 0.57 | 0.53 | | 6 | –0.03 | 0.33 | | 6 | 0.79 | 0.41 |
| | 7 | 0.77 | 0.44 | | 7 | –0.15 | 0.45 | | 7 | 0.53 | 0.54 |
| | 8 | 0.51 | 0.60 | | 8 | – | – | | 8 | 0.44 | 0.62 |
| | 9 | 0.43 | 0.57 | | 9 | 0.05 | 0.28 | | 9 | 0.59 | 0.52 |
| | 10 | 0.36 | 0.57 | | 10 | –0.01 | 0.41 | | 10 | 0.76 | 0.44 |
| | 11 | 0.51 | 0.52 | | 11 | – | – | | 11 | 0.77 | 0.46 |
| | 12 | 0.23 | 0.60 | | 12 | 0.05 | 0.59 | | 12 | 0.83 | 0.38 |
| 2016 | 1 | 0.73 | 0.48 | 2016 | 1 | 0.08 | 0.27 | 2016 | 1 | 0.92 | 0.27 |
| | 2 | 0.47 | 0.55 | | 2 | 0.22 | 0.44 | | 2 | 0.81 | 0.43 |
| | 3 | 0.83 | 0.39 | | 3 | – | – | | 3 | 0.64 | 0.50 |
| | 4 | 0.75 | 0.51 | | 4 | 0.34 | 0.54 | | 4 | 0.68 | 0.48 |
| | 5 | 0.48 | 0.67 | | 5 | 0.58 | 0.55 | | 5 | 0.84 | 0.37 |
| | 6 | 0.53 | 0.53 | | 6 | – | – | | 6 | 0.69 | 0.48 |
| | 7 | 0.40 | 0.52 | | 7 | – | – | | 7 | 0.57 | 0.51 |
| | 8 | 0.77 | 0.45 | | 8 | – | – | | 8 | 0.77 | 0.43 |
| | 9 | 0.63 | 0.53 | | 9 | 0.15 | 0.40 | | 9 | 0.68 | 0.50 |
| | 10 | 0.82 | 0.43 | | 10 | 0.16 | 0.37 | | 10 | 0.70 | 0.47 |
| | 11 | 0.50 | 0.52 | | 11 | 0.05 | 0.22 | | 11 | 0.62 | 0.69 |
| | 12 | 0.71 | 0.48 | | 12 | – | – | | 12 | 0.86 | 0.37 |
| 2017 | 1 | 0.59 | 0.64 | 2017 | 1 | – | – | 2017 | 1 | 0.48 | 0.57 |
| | 2 | 0.63 | 0.51 | | 2 | 0.07 | 0.27 | | 2 | 0.80 | 0.44 |
| | 3 | 0.73 | 0.49 | | 3 | 0.28 | 0.51 | | 3 | 0.84 | 0.40 |
| | 4 | 0.81 | 0.43 | | 4 | 0.18 | 0.51 | | 4 | 0.58 | 0.58 |
| | 5 | 0.78 | 0.50 | | 5 | – | – | | 5 | 0.79 | 0.49 |
| | 6 | 0.72 | 0.54 | | 6 | – | – | | 6 | 0.66 | 0.61 |
| | 7 | 0.64 | 0.56 | | 7 | 0.18 | 0.40 | | 7 | 0.50 | 0.54 |
| | 8 | 0.70 | 0.53 | | 8 | 0.61 | 0.57 | | 8 | 0.82 | 0.41 |
| | 9 | 0.66 | 0.50 | | 9 | 0.07 | 0.71 | | 9 | 0.82 | 0.40 |
| | 10 | 0.72 | 0.50 | | 10 | 0.66 | 0.54 | | 10 | 0.84 | 0.39 |
| | 11 | 0.55 | 0.58 | | 11 | 0.15 | 0.67 | | 11 | 0.51 | 0.64 |
| | 12 | 0.70 | 0.51 | | 12 | 0.39 | 0.49 | | 12 | 0.89 | 0.34 |
| 2018 | 1 | 0.04 | 0.81 | 2018 | 1 | 0.34 | 0.55 | 2018 | 1 | 0.78 | 0.46 |
| | 2 | 0.74 | 0.55 | | 2 | – | – | | 2 | 0.63 | 0.52 |
| | 3 | 0.68 | 0.56 | | 3 | – | – | | 3 | 0.64 | 0.55 |
| | 4 | 0.77 | 0.47 | | 4 | 0.45 | 0.59 | | 4 | 0.40 | 0.53 |
| | 5 | 0.64 | 0.62 | | 5 | 0.35 | 0.48 | | 5 | 0.64 | 0.53 |
| | 6 | 0.13 | 0.89 | | 6 | 0.15 | 0.39 | | 6 | 0.78 | 0.47 |
| | 7 | 0.55 | 0.66 | | 7 | 0.01 | 0.52 | | 7 | 0.63 | 0.57 |
| | 8 | 0.70 | 0.54 | | 8 | 0.54 | 0.57 | | 8 | 0.79 | 0.43 |
| | 9 | 0.69 | 0.56 | | 9 | 0.76 | 0.44 | | 9 | 0.84 | 0.42 |
| | 10 | 0.71 | 0.51 | | 10 | 0.20 | 0.43 | | 10 | 0.67 | 0.60 |
| | 11 | 0.00 | 0.70 | | 11 | –0.07 | 0.43 | | 11 | 0.53 | 0.62 |
| | 12 | 0.10 | 0.73 | | 12 | –0.12 | 0.60 | | 12 | 0.32 | 0.70 |
| 2019 | 1 | 0.26 | 0.77 | 2019 | 1 | –0.41 | 0.63 | 2019 | 1 | 0.76 | 0.50 |
| | 2 | –0.21 | 0.87 | | 2 | –0.03 | 0.56 | | 2 | 0.70 | 0.53 |
| | 3 | 0.40 | 0.64 | | 3 | –0.14 | 0.49 | | 3 | 0.73 | 0.55 |
| | 4 | 0.69 | 0.56 | | 4 | –0.29 | 0.60 | | 4 | 0.79 | 0.46 |
| | 5 | 0.55 | 0.64 | | 5 | –0.35 | 0.62 | | 5 | 0.79 | 0.47 |

| bacteria | | | | plants | | | | unspecified | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Year** | **Month** | **Sent** | **SD** | **Year** | **Month** | **Sent** | **SD** | **Year** | **Month** | **Sent** | **SD** |
| **2013** | **1** | – | – | **2013** | **1** | | | **2013** | **1** | – | – |
| | **2** | – | – | | **2** | | | | **2** | – | – |
| | **3** | – | – | | **3** | | | | **3** | – | – |
| | **4** | – | – | | **4** | – | – | | **4** | – | – |
| | **5** | – | – | | **5** | – | – | | **5** | – | – |
| | **6** | – | – | | **6** | – | – | | **6** | – | – |
| | **7** | | | | **7** | – | – | | **7** | 0.31 | 0.61 |
| | **8** | – | – | | **8** | – | – | | **8** | 0.57 | 0.53 |
| | **9** | – | – | | **9** | – | – | | **9** | 0.48 | 0.51 |
| | **10** | – | – | | **10** | – | – | | **10** | 0.40 | 0.53 |
| | **11** | – | – | | **11** | – | – | | **11** | 0.64 | 0.51 |
| | **12** | – | – | | **12** | – | – | | **12** | 0.70 | 0.47 |
| **2014** | **1** | – | – | **2014** | **1** | – | – | **2014** | **1** | 0.76 | 0.43 |
| | **2** | – | – | | **2** | – | – | | **2** | 0.56 | 0.50 |
| | **3** | – | – | | **3** | – | – | | **3** | 0.75 | 0.44 |
| | **4** | – | – | | **4** | – | – | | **4** | 0.52 | 0.51 |
| | **5** | – | – | | **5** | – | – | | **5** | 0.47 | 0.52 |
| | **6** | – | – | | **6** | – | – | | **6** | 0.61 | 0.51 |
| | **7** | – | – | | **7** | – | – | | **7** | 0.48 | 0.51 |
| | **8** | – | – | | **8** | – | – | | **8** | 0.56 | 0.52 |
| | **9** | – | – | | **9** | – | – | | **9** | 0.52 | 0.50 |
| | **10** | 0.91 | 0.28 | | **10** | – | – | | **10** | 0.51 | 0.51 |
| | **11** | – | – | | **11** | – | – | | **11** | 0.60 | 0.50 |
| | **12** | – | – | | **12** | – | – | | **12** | 0.37 | 0.51 |
| **2015** | **1** | – | – | **2015** | **1** | – | – | **2015** | **1** | 0.54 | 0.51 |
| | **2** | 0.87 | 0.33 | | **2** | – | – | | **2** | 0.58 | 0.50 |
| | **3** | 0.74 | 0.45 | | **3** | 0.46 | 0.53 | | **3** | 0.45 | 0.59 |
| | **4** | 0.59 | 0.49 | | **4** | 0.56 | 0.53 | | **4** | 0.45 | 0.55 |
| | **5** | 0.78 | 0.42 | | **5** | 0.73 | 0.45 | | **5** | 0.38 | 0.53 |
| | **6** | 0.51 | 0.50 | | **6** | 0.65 | 0.48 | | **6** | 0.44 | 0.63 |
| | **7** | 0.64 | 0.48 | | **7** | 0.61 | 0.49 | | **7** | 0.59 | 0.55 |
| | **8** | – | – | | **8** | 0.42 | 0.49 | | **8** | 0.55 | 0.52 |
| | **9** | 0.73 | 0.46 | | **9** | 0.66 | 0.47 | | **9** | 0.52 | 0.56 |
| | **10** | 0.56 | 0.50 | | **10** | 0.83 | 0.39 | | **10** | 0.40 | 0.54 |
| | **11** | 0.52 | 0.52 | | **11** | 0.61 | 0.49 | | **11** | 0.49 | 0.56 |
| | **12** | 0.71 | 0.46 | | **12** | 0.50 | 0.51 | | **12** | 0.49 | 0.55 |
| **2016** | **1** | 0.56 | 0.57 | **2016** | **1** | 0.71 | 0.45 | **2016** | **1** | 0.40 | 0.55 |
| | **2** | 0.38 | 0.49 | | **2** | 0.72 | 0.46 | | **2** | 0.33 | 0.56 |
| | **3** | 0.48 | 0.51 | | **3** | 0.59 | 0.51 | | **3** | 0.31 | 0.64 |
| | **4** | 0.53 | 0.50 | | **4** | 0.26 | 0.59 | | **4** | 0.49 | 0.54 |
| | **5** | 0.45 | 0.50 | | **5** | 0.38 | 0.56 | | **5** | 0.44 | 0.54 |
| | **6** | 0.62 | 0.49 | | **6** | 0.72 | 0.45 | | **6** | 0.56 | 0.52 |
| | **7** | 0.23 | 0.57 | | **7** | 0.48 | 0.54 | | **7** | 0.31 | 0.54 |
| | **8** | 0.61 | 0.49 | | **8** | 0.73 | 0.45 | | **8** | 0.57 | 0.53 |
| | **9** | 0.65 | 0.52 | | **9** | 0.55 | 0.53 | | **9** | 0.33 | 0.58 |
| | **10** | 0.52 | 0.65 | | **10** | 0.44 | 0.54 | | **10** | 0.33 | 0.56 |
| | **11** | 0.54 | 0.50 | | **11** | 0.66 | 0.48 | | **11** | 0.26 | 0.59 |
| | **12** | 0.58 | 0.50 | | **12** | 0.71 | 0.46 | | **12** | 0.40 | 0.53 |
| **2017** | **1** | 0.36 | 0.49 | **2017** | **1** | 0.63 | 0.49 | **2017** | **1** | 0.56 | 0.52 |
| | **2** | 0.64 | 0.49 | | **2** | 0.67 | 0.47 | | **2** | 0.28 | 0.49 |
| | **3** | 0.76 | 0.44 | | **3** | 0.75 | 0.44 | | **3** | 0.46 | 0.53 |
| | **4** | 0.95 | 0.23 | | **4** | 0.58 | 0.50 | | **4** | 0.44 | 0.55 |
| | **5** | 0.27 | 0.49 | | **5** | 0.85 | 0.37 | | **5** | 0.16 | 0.72 |
| | **6** | 0.86 | 0.36 | | **6** | 0.70 | 0.48 | | **6** | 0.12 | 0.67 |
| | **7** | 0.25 | 0.44 | | **7** | 0.62 | 0.51 | | **7** | 0.34 | 0.60 |
| | **8** | 0.75 | 0.43 | | **8** | 0.66 | 0.47 | | **8** | 0.46 | 0.57 |
| | **9** | 0.84 | 0.37 | | **9** | 0.84 | 0.37 | | **9** | 0.43 | 0.56 |
| | **10** | 0.66 | 0.48 | | **10** | 0.67 | 0.50 | | **10** | 0.52 | 0.54 |
| | **11** | 0.81 | 0.40 | | **11** | 0.66 | 0.55 | | **11** | 0.44 | 0.57 |
| | **12** | 0.80 | 0.40 | | **12** | 0.64 | 0.58 | | **12** | 0.38 | 0.58 |
| **2018** | **1** | 0.49 | 0.70 | **2018** | **1** | 0.63 | 0.53 | **2018** | **1** | 0.42 | 0.60 |
| | **2** | 0.75 | 0.45 | | **2** | 0.68 | 0.47 | | **2** | 0.50 | 0.58 |
| | **3** | 0.87 | 0.35 | | **3** | 0.50 | 0.52 | | **3** | 0.49 | 0.62 |
| | **4** | 0.75 | 0.44 | | **4** | 0.63 | 0.52 | | **4** | 0.51 | 0.63 |
| | **5** | 0.69 | 0.48 | | **5** | 0.73 | 0.47 | | **5** | 0.51 | 0.55 |
| | **6** | 0.74 | 0.48 | | **6** | 0.65 | 0.48 | | **6** | 0.36 | 0.61 |
| | **7** | 0.78 | 0.42 | | **7** | 0.17 | 0.70 | | **7** | –0.03 | 0.85 |
| | **8** | 0.76 | 0.46 | | **8** | 0.34 | 0.76 | | **8** | 0.46 | 0.59 |
| | **9** | 0.74 | 0.44 | | **9** | 0.44 | 0.67 | | **9** | 0.35 | 0.59 |
| | **10** | 0.91 | 0.29 | | **10** | 0.60 | 0.59 | | **10** | 0.49 | 0.57 |
| | **11** | 0.66 | 0.68 | | **11** | 0.47 | 0.69 | | **11** | 0.39 | 0.63 |
| | **12** | 0.24 | 0.56 | | **12** | 0.71 | 0.48 | | **12** | 0.40 | 0.64 |
| **2019** | **1** | 0.74 | 0.51 | **2019** | **1** | 0.69 | 0.49 | **2019** | **1** | 0.49 | 0.59 |
| | **2** | 0.67 | 0.47 | | **2** | 0.77 | 0.45 | | **2** | 0.40 | 0.60 |
| | **3** | 0.70 | 0.46 | | **3** | 0.63 | 0.52 | | **3** | 0.45 | 0.60 |
| | **4** | 0.72 | 0.46 | | **4** | 0.59 | 0.66 | | **4** | 0.47 | 0.63 |
| | **5** | 0.85 | 0.38 | | **5** | 0.65 | 0.53 | | **5** | 0.55 | 0.56 |

Table B.5: Monthly mean sentiments and standard deviations per organism. The table shows the mean sentiments (Sent) and their standard deviations (SD) for every months and organism. A dash (–) indicates that less than 100 tweets were in the respective organism category for that month and that we did not calculate the mean sentiment. Months with empty rows had no tweets in that category. The mean values of this table were used in Figure 4.2.

| Year | Hashtag | Count | Sent | SD |
|---|---|---|---|---|
| **2013** | genome | 94 | 0.81 | 0.40 |
| | dna | 48 | 0.92 | 0.35 |
| | cas9 | 44 | 0.32 | 0.47 |
| | drosophila | 41 | 0.41 | 0.50 |
| | crisp | 38 | 1.00 | 0.00 |
| | genetics | 38 | 0.82 | 0.39 |
| | synbio | 38 | 0.42 | 0.64 |
| | btoty | 34 | 1.00 | 0.00 |
| | science | 31 | 0.74 | 0.44 |
| | rna | 30 | 0.47 | 0.51 |
| | editas | 28 | 0.96 | 0.19 |
| | genomics | 28 | 0.43 | 0.50 |
| | gblocks | 23 | 0.00 | 0.00 |
| | cell | 22 | 0.82 | 0.39 |
| | biotech | 20 | 0.50 | 0.51 |
| **2014** | genomics | 368 | 0.80 | 0.40 |
| | dna | 273 | 0.63 | 0.48 |
| | synbio | 244 | 0.68 | 0.47 |
| | cas9 | 225 | 0.60 | 0.49 |
| | science | 222 | 0.71 | 0.46 |
| | genome | 195 | 0.77 | 0.42 |
| | biotech | 177 | 0.64 | 0.48 |
| | genetics | 175 | 0.73 | 0.45 |
| | nbthighlight | 165 | 0.35 | 0.48 |
| | sciwri14 | 118 | 0.29 | 0.45 |
| | ashg14 | 115 | 0.39 | 0.49 |
| | rna | 110 | 0.53 | 0.50 |
| | nbtinthenews | 91 | 0.92 | 0.27 |
| | genetherapy | 90 | 0.54 | 0.50 |
| | drosophila | 86 | 0.36 | 0.48 |
| **2015** | geneeditsummit | 3337 | 0.20 | 0.44 |
| | science | 2096 | 0.56 | 0.58 |
| | crisprfacts | 1322 | 0.36 | 0.59 |
| | dna | 1148 | 0.33 | 0.73 |
| | geneediting | 1088 | 0.34 | 0.54 |
| | genetics | 1045 | 0.23 | 0.70 |
| | genomeediting | 963 | 0.49 | 0.52 |
| | genome | 962 | 0.55 | 0.59 |
| | biotech | 938 | 0.49 | 0.53 |
| | genomics | 848 | 0.47 | 0.55 |
| | bioethics | 797 | –0.02 | 0.45 |
| | cas9 | 781 | 0.53 | 0.55 |
| | synbio | 722 | 0.51 | 0.54 |
| | gene | 610 | 0.21 | 0.77 |
| | cancer | 441 | 0.85 | 0.36 |
| **2016** | science | 4479 | 0.62 | 0.52 |
| | geneediting | 3920 | 0.42 | 0.53 |
| | tech | 2163 | 0.46 | 0.59 |
| | biotech | 2132 | 0.45 | 0.55 |
| | genetics | 1748 | 0.47 | 0.53 |
| | cancer | 1671 | 0.69 | 0.46 |
| | dna | 1626 | 0.66 | 0.50 |
| | news | 1551 | 0.43 | 0.57 |
| | gmo | 1518 | 0.24 | 0.57 |
| | genomics | 1496 | 0.50 | 0.53 |
| | hiv | 1459 | 0.90 | 0.34 |
| | obesity | 1172 | 0.47 | 0.58 |
| | gene | 1057 | 0.64 | 0.51 |
| | patent | 1038 | 0.02 | 0.28 |
| | cas9 | 972 | 0.57 | 0.52 |

| Year | Hashtag | Count | Sent | SD |
|---|---|---|---|---|
| **2017** | geneediting | 12,648 | 0.38 | 0.56 |
| | genomeediting | 9747 | 0.36 | 0.55 |
| | science | 5190 | 0.60 | 0.56 |
| | biotech | 3374 | 0.53 | 0.57 |
| | tech | 3322 | 0.63 | 0.53 |
| | dna | 2894 | 0.56 | 0.57 |
| | genetics | 2546 | 0.54 | 0.59 |
| | genomics | 2294 | 0.51 | 0.57 |
| | cancer | 2003 | 0.68 | 0.55 |
| | health | 1883 | 0.79 | 0.49 |
| | news | 1804 | 0.62 | 0.55 |
| | ai | 1743 | 0.47 | 0.53 |
| | technology | 1679 | 0.58 | 0.56 |
| | sntop10 | 1533 | 0.25 | 0.64 |
| | gmo | 1504 | 0.14 | 0.75 |
| **2018** | geneediting | 13,000 | 0.37 | 0.63 |
| | genomeediting | 7765 | 0.40 | 0.58 |
| | science | 5210 | 0.52 | 0.64 |
| | biotech | 5052 | 0.41 | 0.60 |
| | genetics | 4468 | 0.52 | 0.58 |
| | dna | 4206 | 0.49 | 0.63 |
| | crisprbabies | 4020 | –0.30 | 0.65 |
| | gmo | 3886 | 0.09 | 0.61 |
| | cancer | 3547 | 0.57 | 0.72 |
| | ai | 3421 | 0.63 | 0.51 |
| | genomics | 3272 | 0.43 | 0.60 |
| | geneeditsummit | 2504 | –0.01 | 0.46 |
| | synbio | 2289 | 0.58 | 0.58 |
| | gmos | 2139 | 0.03 | 0.51 |
| | cas9 | 2009 | 0.45 | 0.65 |
| **2019*** | geneediting | 10,764 | 0.42 | 0.62 |
| | genomeediting | 5950 | 0.40 | 0.57 |
| | biotech | 4778 | 0.55 | 0.53 |
| | science | 4097 | 0.57 | 0.59 |
| | dna | 3734 | 0.54 | 0.62 |
| | genetics | 3720 | 0.53 | 0.55 |
| | technology | 2657 | 0.68 | 0.53 |
| | genomics | 2590 | 0.48 | 0.55 |
| | cancer | 2306 | 0.68 | 0.55 |
| | gmo | 2090 | –0.11 | 0.76 |
| | cas9 | 1841 | 0.61 | 0.53 |
| | researchhighlight | 1822 | 1.00 | 0.06 |
| | ai | 1793 | 0.56 | 0.56 |
| | crisprbabies | 1730 | –0.13 | 0.63 |
| | genetherapy | 1637 | 0.62 | 0.54 |
| **2019** | geneediting | 4485 | 0.42 | 0.62 |
| | genomeediting | 2479 | 0.40 | 0.57 |
| | biotech | 1991 | 0.55 | 0.53 |
| | science | 1707 | 0.57 | 0.59 |
| | dna | 1556 | 0.54 | 0.62 |
| | genetics | 1550 | 0.53 | 0.55 |
| | technology | 1107 | 0.68 | 0.53 |
| | genomics | 1079 | 0.48 | 0.55 |
| | cancer | 961 | 0.68 | 0.55 |
| | gmo | 871 | –0.11 | 0.76 |
| | cas9 | 767 | 0.61 | 0.53 |
| | researchhighlight | 759 | 1.00 | 0.06 |
| | ai | 747 | 0.56 | 0.56 |
| | crisprbabies | 721 | –0.13 | 0.63 |
| | genetherapy | 682 | 0.62 | 0.54 |

Table B.6: Top hashtags' counts and sentiments. List of top 15 hashtags, corresponding counts (Count), sentiments (Sent) and standard deviations (SD) by year. The extrapolated hashtag counts for 2019 are shown under 2019*, the original counts for the first five months under 2019. The mean values of this table were used in Figure 4.4.

| Sentiment | Year | genome | baby | disease | embryo | treatment | mutation |
|---|---|---|---|---|---|---|---|
| **negative** | **2013** | 8 | 0 | 0 | 0 | 0 | 0 |
| | **2014** | 24 | 0 | 0 | 1 | 0 | 1 |
| | **2015** | 1407 | 58 | 18 | 418 | 16 | 40 |
| | **2016** | 1678 | 72 | 149 | 69 | 151 | 58 |
| | **2017** | 6392 | 137 | 289 | 1092 | 35 | 3647 |
| | **2018** | 18,340 | 8363 | 484 | 986 | 970 | 1598 |
| | **2019*** | 8431 | 6586 | 425 | 1382 | 154 | 871 |
| | **2019** | 3513 | 2744 | 177 | 576 | 64 | 363 |
| **neutral** | **2013** | 516 | 0 | 10 | 3 | 5 | 45 |
| | **2014** | 2019 | 7 | 41 | 17 | 15 | 89 |
| | **2015** | 17,894 | 1376 | 331 | 2798 | 124 | 110 |
| | **2016** | 36,039 | 575 | 403 | 3579 | 361 | 294 |
| | **2017** | 40,272 | 3178 | 1111 | 12,293 | 380 | 1353 |
| | **2018** | 53,096 | 19,705 | 1967 | 6551 | 932 | 2485 |
| | **2019*** | 40,241 | 15,794 | 1572 | 2755 | 785 | 898 |
| | **2019** | 16,767 | 6581 | 655 | 1148 | 327 | 374 |
| **positive** | **2013** | 1176 | 2 | 492 | 7 | 41 | 31 |
| | **2014** | 4347 | 5 | 429 | 15 | 60 | 349 |
| | **2015** | 22,036 | 544 | 1402 | 570 | 933 | 469 |
| | **2016** | 40,402 | 320 | 5748 | 1643 | 5868 | 3019 |
| | **2017** | 68,733 | 1218 | 15,754 | 10,762 | 6994 | 8510 |
| | **2018** | 81,303 | 4869 | 17,358 | 2828 | 7757 | 6450 |
| | **2019*** | 82,258 | 5460 | 16,198 | 1315 | 10,284 | 4567 |
| | **2019** | 34,274 | 2275 | 6749 | 548 | 4285 | 1903 |

Table B.7: Top themes found in tweets. List of top 6 themes with highest overall occurrence across sentiment. The table shows the number of occurrences in tweets for every sentiment and year. The year 2019 was extrapolated to determine the top themes, indicated by the star (*), based on the first five months of 2019. These counts were used in Figure 4.5.

# C Supplementary Information: Experts and authorities receive disproportionate attention on Twitter during the COVID-19 crisis

## C.1 Overview

We provide a set of alternative views on the analysis discussed in Figure 2. First, in Fig. C.3, we show how the topic of tweets—whether they are COVID-19-related or not—plays a role in determining the degree of engagement they received during the pandemic, with COVID-19-related tweets consistently receiving more engagement for Healthcare, Government & Politics and Political Supporters. We support this analysis with regression modelling, presented in Fig. C.4 and Fig. C.5. In Fig. C.6, we provide an alternative view of Fig. 2, where each week of the Study Period corresponds to a point connected by an arrow with the previous week. In Fig. C.7, we provide the results for the "other" category, which is excluded in the analysis. Similarly, in Fig. C.9, we examine the robustness of our findings by evaluating the impact of users joining the platform during the Studied Period. Lastly, we provide additional information about a set of supplemental experiments, with the goal of understanding, first, the degree of automated activity within the studied accounts, and second, the between-category interactions that drive the trends in engagement. In order to measure which categories retweet which other categories, we use an automated method for label expansion, detailed below.

## C.2 Bot detection

To assess the degree of bot activity in our data, we used the tool Botometer (Davis et al. 2016). Botometer uses a supervised Machine Learning approach to estimate the so called complete automation probability (CAP), for which a value of 1 indicates complete automation. Botometer extracts features from recent tweets in the account's timeline, such as temporal activity patterns, social networks and sentiment, among others. In this work, we use a CAP threshold of 0.25 in order to decide whether a account is presumed to be a bot. The bot activity data was collected via the Botometer API between July 22 and July 27, 2020. By using the method above on a sample of 5000 accounts in our annotation dataset (dataset A), we find around 3.3% of presumed automated accounts. Bot activity in the annotation dataset was significantly higher for accounts annotated as "other" (4.6% bots) for the category labels and "unclear" (5.3%) in the type of account labeling (that is, when annotators had to classify accounts as belonging to an individual or an institution). Testing was performed using a one-tailed binomial test at significance level $\alpha = 0.5$ (before Bonferroni correction). Based on these numbers, bots seem to only have a marginal influence on the overall validity of the results which are based on the sampled user accounts (dataset B).

## C.3 Who retweets whom?

In Fig. C.10, we looked at all tweets and retweets produced in the week of interest. Recall that here we take advantage of the fact that the data obtained from the stream is complete, that is we are certain to have all the retweets of a given tweet. We deploy an automatic classifier described better in Section 1.3, to automatically label the category of all accounts in the week of interest. Excluded are accounts with user descriptions of less than 3 characters, yielding labels for a total of 39.2M users. With the labels generated by the classifier, we build a retweet digraph $G$. Each node $u$ in this graph is an account, assigned to a single category (the most likely according to the classifier). Each $(u, v)$ edge in this graph stands for a retweet from account $u$ to account $v$. That means that an edge only exists if the tweet by account $v$ was retweeted more than 10 times. Given this graph, we proceeded to explore the number of retweets between categories. This can be thought of as a collapsed graph $G'$ where all nodes with the same category are collapsed into one. Looking at this graph we analyze, for each category, where are the sources of the incoming edges. We also obtain a null model with this graph. This null model assumes that each category is equally likely to connect to any other category. Thus, suppose we want to calculate the percentage of incoming edges from category $X$ to category $Y$. Let $Out(X)$ be the number of outgoing edges from category $X$ and $In(Y)$ be the number of incoming edges from category $Y$. Also, let $B$ be the total number of edges in the graph. Notice that $In(Y)/B$ is the fraction of all edges that are incoming edges towards $Y$. If the assignment of edges from category $X$ is really independent of other

categories, we would expect $Out(X)\frac{In(Y)}{B}$ edges between $X$ and $Y$. This what we consider to be our category-agnostic random null model. Lastly, to obtain confidence intervals over this analysis we bootstrap the whole process, we choose a random sample of the edges in the original graph $G$ to "collapse" generating the category-graph $G'$. We repeat this procedure 1000 times, and obtain confidence intervals for the expected value and the observed value for each category.

## C.4 Label Expansion

In Fig. C.10, we used label expansion, a method in which a Machine Learning classifier is trained on the subset of annotated data to predict the labels for the full data set. The account descriptions consist of unstructured text, including frequent use of emojis, and special Unicode characters. Furthermore, the entire COVID-19 Twitter stream data is multilingual, covering 41 languages from very diverse language families. Given this complexity, two major approaches were tried using the FastText library (Joulin et al. 2017) and models based on the BERT family (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018).

### C.4.1 BERT

BERT is a general-purpose language understanding model which can be used, among other applications, for text classification. BERT models are pretrained on large bodies of plain text (e.g. from Wikipedia) in an unsupervised way. Pretrained models can then be used in a supervised downstream task, such as text classification, in a process called finetuning.

In this work, we started with the pretrained multilingual cased BERT model (`bert-multilang`), a BERT model which was simultaneously pretrained on the Wikipedia corpora of 104 languages. However, our target domain (Twitter account descriptions) is very different from text found in Wikipedia. Therefore, an additional step of unsupervised pretraining, also called domain-specific pretraining, was conducted on our existing corpus of account descriptions. The domain-specific pretraining and finetuning of BERT models was performed with code from the COVID-Twitter-BERT repository (Müller, Salathé, and Kummervold 2020) Account descriptions from dataset A of at least 3 characters length, which were not contained in the annotation dataset (dataset B), were combined into a dataset of 52M account descriptions, comprising of a total of 697M tokens. The data was preprocessed by replacing account names, URLs, and email addresses with generic fillers. Furthermore, emojis were replaced by textual versions (e.g. the American flag emoji would be replaced by `:flag-us:`), using the Python emoji library[I]. From this dataset 593M training examples were generated. Training was run for roughly 1 epoch (600k steps) at a batch size of 1024 and a constant learning rate of 2e-5.

[I] `https://pypi.org/project/emoji/`

Training took roughly 2 days on a TPU v3-8 (8 cores, 128 GB of memory), and resulted in a new model, which we refer to as `bert-multilang-pt`.

A similar procedure was applied for a English-only model, in which pretraining was conducted with account descriptions in English (251M training examples, 21M account descriptions), and pretraining was started from the English BERT-large uncased (whole word masking) variant (`bert-english`). Training for this variant was conducted with the same batch size and learning rate, but for roughly 2 epochs (roughly 5 days of training). We will refer to this variant as `bert-english-pt`.

### C.4.2 FastText

FastText is a lightweight library for text classification and representation learning. It is a shallow model that uses subword information to enrich word vectors. Similar to BERT, it is possible to fine-tune pretrained word representations for text classification purposes. In contrast to BERT, which heavily relies on training on GPUs, it can be trained on a large dataset using multicore CPUs in a matter of minutes. Also, FastText models are much more compact than BERT (in our case, 125 MB vs 700 MB).

For FastText models, we only used account descriptions in English language. Preprocessing was conducted by normalizing texts, replacing accoun tnames, URLs and emails and removing emojis. We then pretrained a FastText skipgram model for 5 epochs, with a learning rate of 0.1, context window size of 5, and n-gram size between 3 and 6. We will refer to the pretrained FastText model as `fasttext-english-pt`.

### C.4.3 Finetuning

Eventually all pretrained models were finetuned on the type (3 classes) and category (13 classes) tasks. The annotation data was deduplicated (accounts may have identical descriptions), and preprocessed in the same way the the pretraining data was prepared for the respective model type. The preprocessed annotation data (100%, $n_{\text{category}} = 9913$, $n_{\text{type}} = 10725$) was split into a training (64%), development (16%), and test set (20%) for both type and category, respectively. Multilingual models were fine-tuned on the original training data, whereas English models were fine-tuned on the translated versions of the account descriptions. Model selection was performed by optimizing the respective F1-macro score on the development set. BERT-like models were fine-tuned in 10 epochs, using a learning rate of 1e-5 (using 10% warm-up with linear decay) and training batch sizes of 32. FastText models were fine-tuned using built-in hyperparameter autotuning available for supervised training with a vector dimension of 100.

### C.4.4 Classifier results

Based on the pretrained models described above, we compare downstream classifier performance scores in Fig. C.13. Unexpectedly, BERT models trained on English-only data outperform the multilingual BERT model. Generally, we also see a performance boost due to domain-specific pretraining. The best English-only model (`bert-english-pt`) gives a F1-macro score of 0.71 and 0.62, on the category and type datasets, respectively. The smaller FastText models (`fasttext-english-pt`) perform comparably to other models on the type dataset but give slightly lower scores on the category dataset. The best multilingual model (`bert-multilang-pt`) yields F1-macro scores of 0.56 (category) and 0.63 (type).

For further analysis we focus on the multilingual BERT model (`bert-multilang-pt`), which was the final model used for label expansion in this work. When inspecting the confusion matrices (Fig. C.11 and Fig. C.12), classifier scores for this model are generally satisfying. Certain classes for which only very few observations are present show lower scores in comparison. In particular, this is concerning the classes "Religion" and "Public Services" (for category) and "Unclear" (for type). The smallest error rates can be expected for the classes "Healthcare", "News Media", and "Government and Politics". No significant deviations from the mean accuracy could be observed for individual languages. Testing was performed using a two-sided binomial test at significance level $\alpha = 0.5$ (before Bonferroni correction).

## References

Davis, Clayton Allen et al. (2016). "Botornot: A system to evaluate social bots". In: *Proceedings of the 25th international conference companion on world wide web*, pp. 273–274.

Joulin, Armand et al. (Apr. 2017). "Bag of Tricks for Efficient Text Classification". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pp. 427–431.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Müller, Martin, Marcel Salathé, and Per E Kummervold (2020). "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter". In: *arXiv preprint arXiv:2005.07503*.

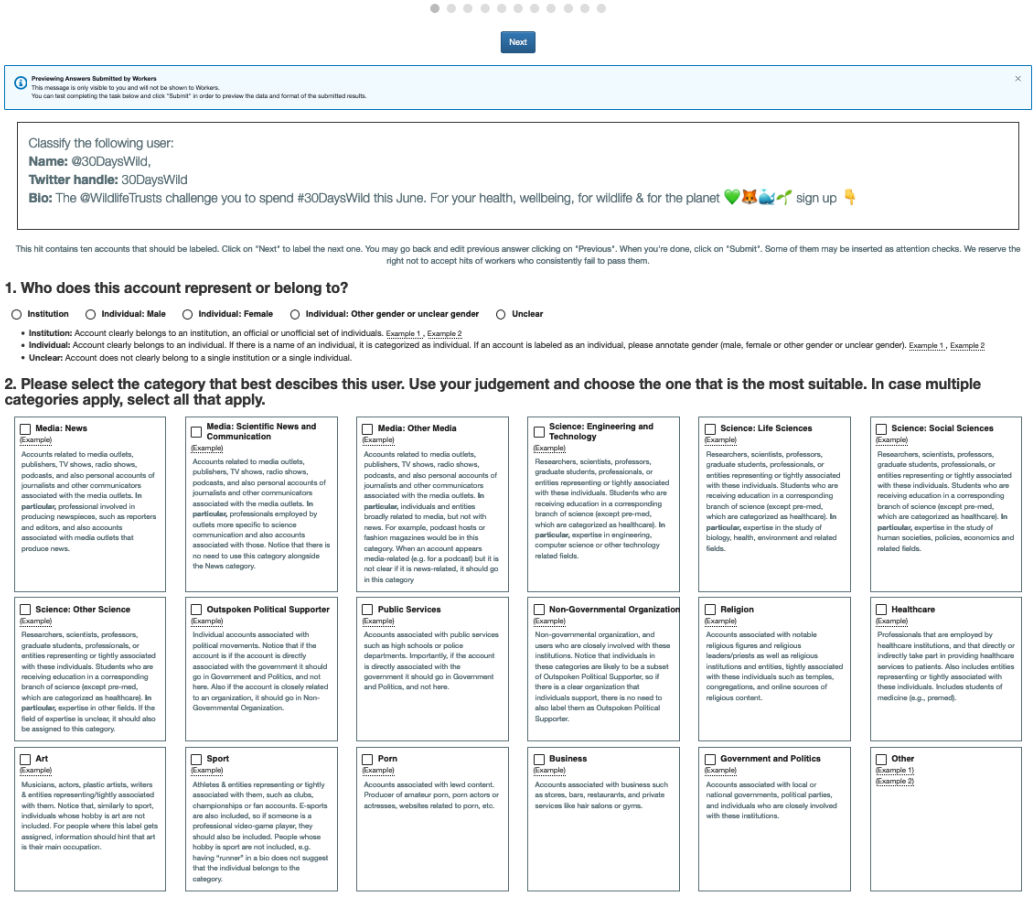



Figure C.1: Screenshot of the annotation interface.

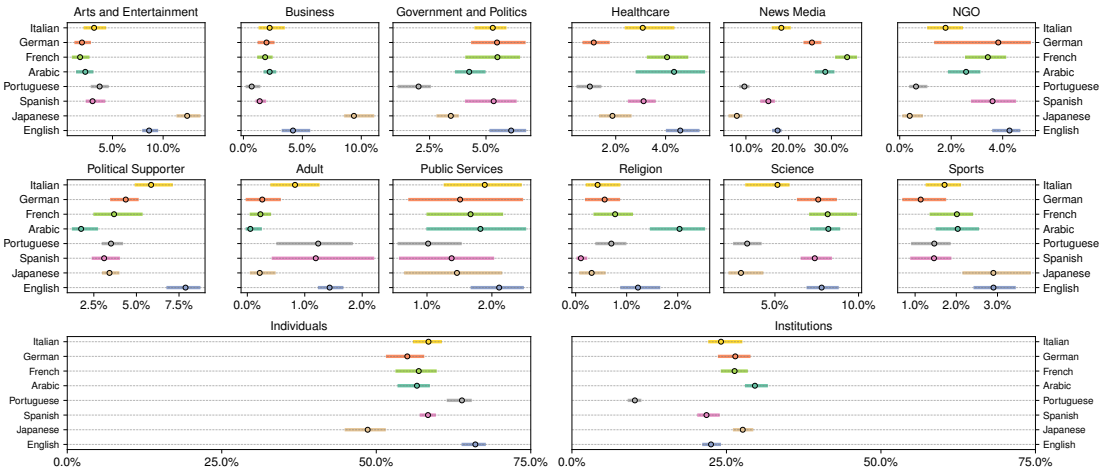



Figure C.2: Category and type prevalence across different languages.

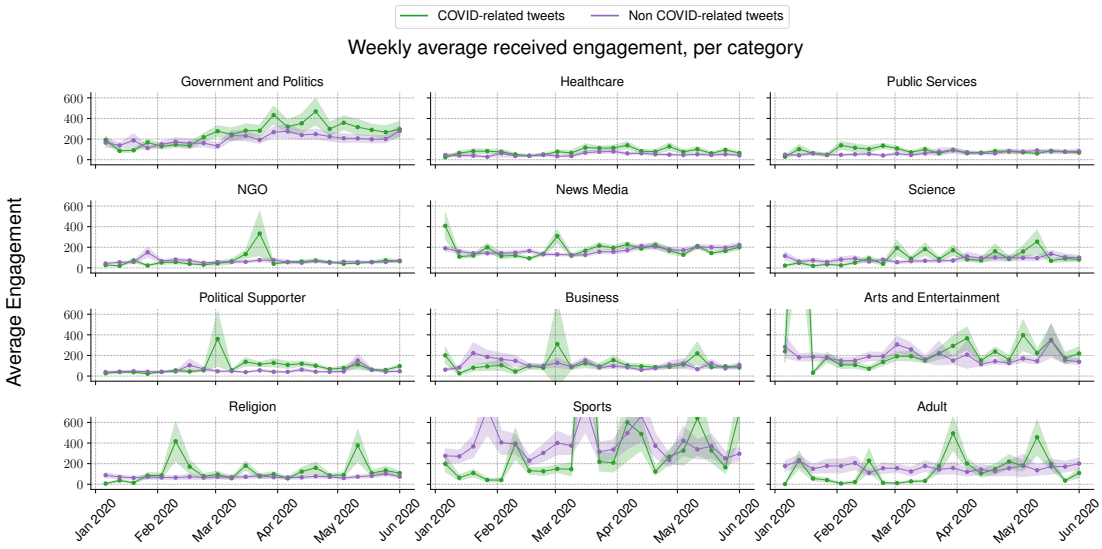



Figure C.3: To further understand the mechanisms behind the change in engagement, we show the results of a complimentary analysis where we look at the effect on engagement of tweeting specifically about COVID-19, identified using the same keywords used by Twitter. COVID-19-related tweets consistently receive more engagement for Healthcare, Government and Politics and Political Supporters.

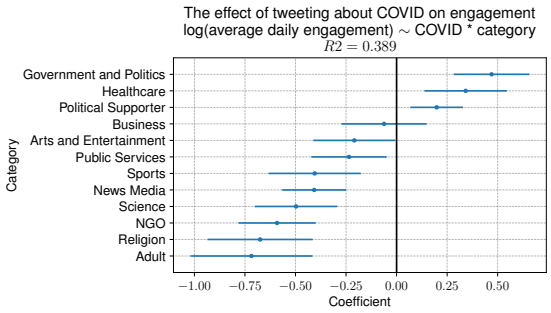



Figure C.4: Daily average received engagement across categories for COVID-19 and non-COVID-19 tweets.

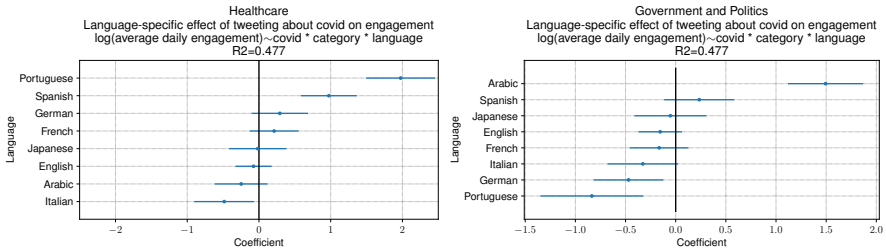



Figure C.5: Language-specific effect of tweeting about COVID-19 on engagement for accounts belonging to Healthcare (left), and Government and Politics (right). Important cultural differences emerge which will require future work. As an example, for COVID-19-related tweets in Portugese (largely from Brazil), Government and Politics is negatively correlated with engagement while Healthcare is positively correlated with engagement.

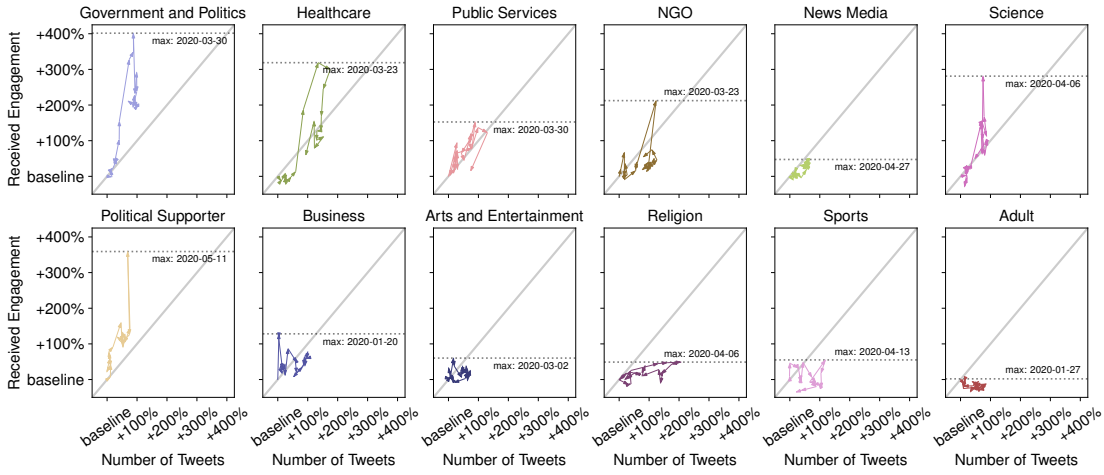
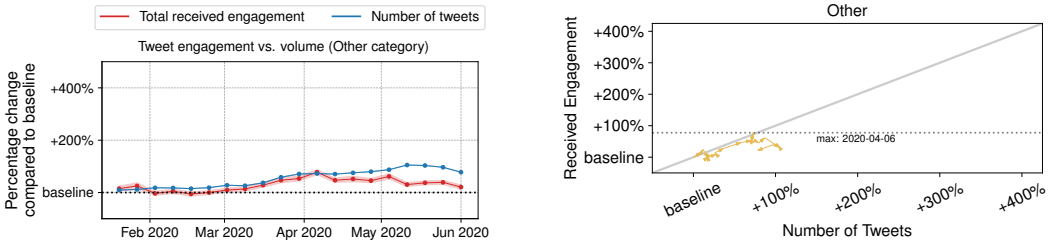



Figure C.6: We show an alternate view of the analysis depicted in Figure 2 of the main text. Here,each week of the Study Period is sequentially connected by arrows in a 2D-plane where the x-axis depicts the weekly average increase in volume, and the y-axis the weekly average increase in engagement.




(a) (b)

Figure C.7: **(a)** The account-averaged percentage change (calculated with IPW) in number of tweets (in blue) and engagement (in red). The change is shown relative to a baseline, calculated using the two weeks of January 2020. **(b)** Each week of the Study Period is sequentially connected by arrows in a 2D-plane where the x-axis depicts the weekly average increase in volume, and the y-axis the weekly average increase in engagement.

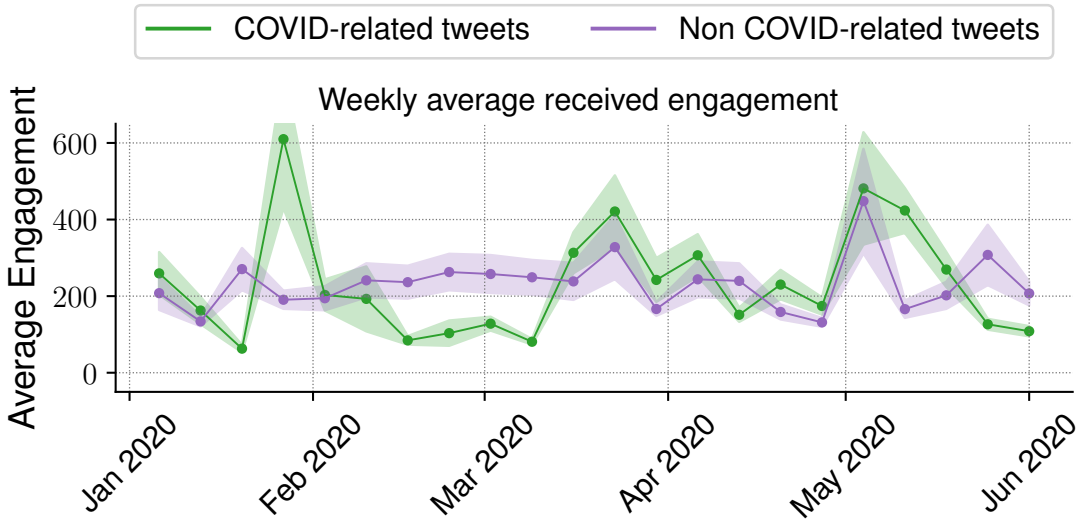



Figure C.8: The effect on average engagement of tweeting specifically about COVID-19.

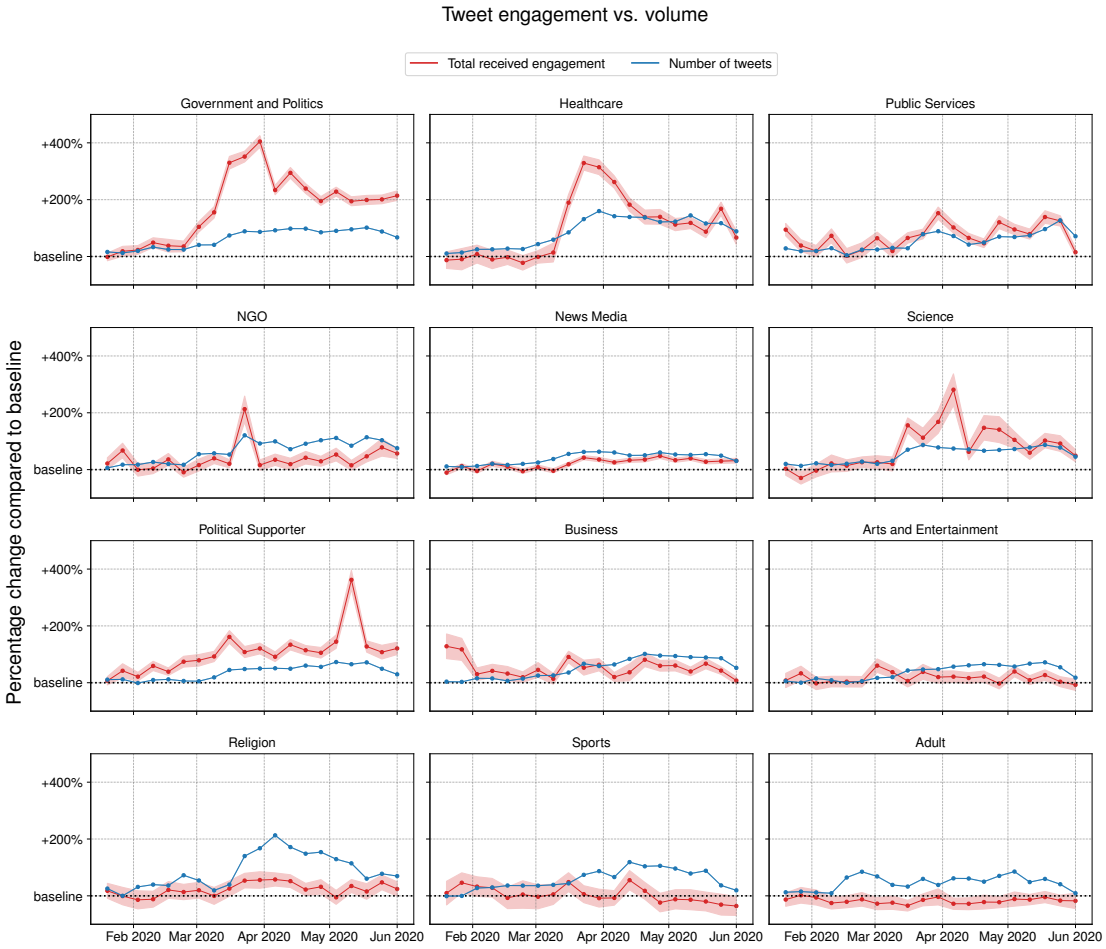
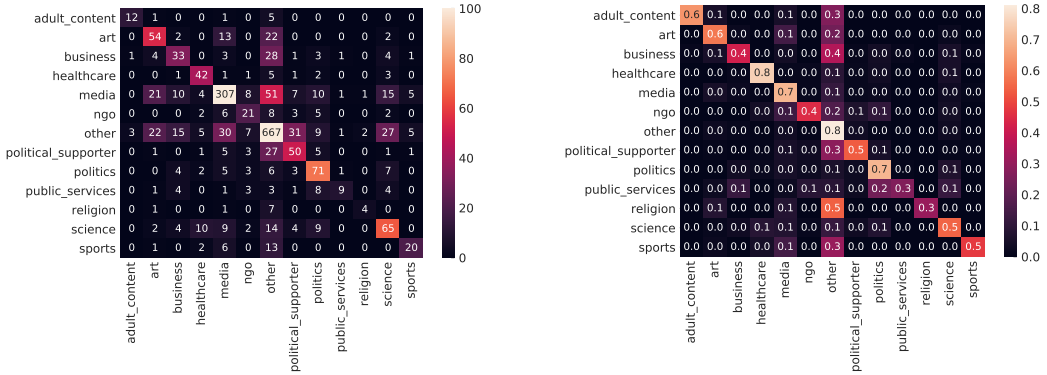


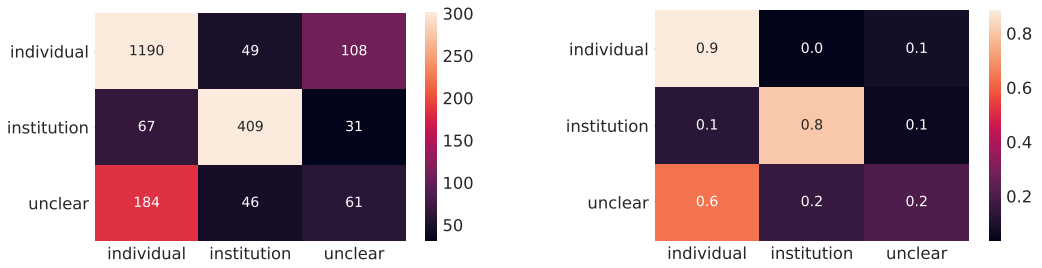

Figure C.9: To alleviate a potential bias in our analysis caused by the hypothetical surge of new users joining the platform during the crisis, we conducted an alternative analysis where we restricted ourselves to a set of users who created their Twitter account before the studied period. In that way, we excluded 482, out of 14000 annotated users. The observed trends are not impacted by the presence of such newcomers.


Figure C.10: We measure, for each category, what is the source of their engagement (who retweets them). We compare this value to a category agnostic null model that assumes each category receives engagement at random, proportionally to their size. The figure shows arrows that start from the expected value, according to the null model, and end at the observed value. Where differences are not significant ($p > 0.05$), arrow ends are replaced by gray circles. The "other" category, while being numerically larger, is a net retweeter of the remaining categories and gets retweeted less frequently than expected. We also observe a strong homophily: all categories retweet significantly ($p < 0.05$) more tweets from their own category than predicted by the null model. The one exception are Political Supporters retweeting Science more than Science retweeting itself.



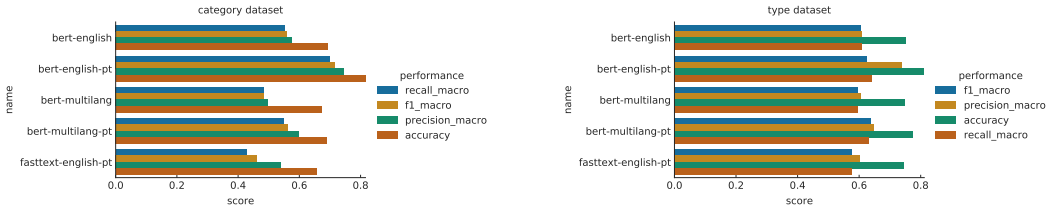

(a) (b)

Figure C.11: Confusion matrix on the held out test set for the multilingual BERT category classifier (fine-tuned version of `bert-multilang-pt`). The y-axis represents the true label (as per annotation data) and the x-axis represents the label predicted by the classifier. Confusion matrix on the left shows absolute counts, whereas on the right normalized counts are shown. Most errors were made by predicting a account description as "other" (which was the most frequent category). The weakest categories are "religion" (often predicted as "other"), and "public_services" (often predicted as "politics"). These categories also have had few training and test examples.




(a) (b)

Figure C.12: Confusion matrix on the held out test set for the BERT type classifier (fine-tuned version of `bert-multilang-pt`). The y-axis represents the true label (as per annotation data) and the x-axis represents the label predicted by the classifier. Confusion matrix on the left shows absolute counts, whereas on the left normalized counts are shown. Predictions for "individual" and "institution" are very accurate. "Unclear" represents a relatively small class, therefore leading to a higher relative error.




(a) (b)

Figure C.13: Comparisons of test scores of BERT and FastText classifiers. Overall, best results are achieved for English-only models. Models which underwent domain-specific pretraining, as indicated by the "pt" suffix, generally outperform the default pretrained models. The model used for the analysis is `bert-multilang-pt`.

Table C.1: The COVID-19 Twitter accounts taxonomy: **category of account**.

**Account category:** Please select the category that best describes this account. Use your judgement and choose the one that is the most suitable. In case multiple categories apply, select all that apply.

| Category of account | Description |
|---|---|
| Media: News | Accounts related to media outlets, publishers, TV shows, radio shows, podcasts, and also personal accounts of journalists and other communicators associated with the media outlets. Professionals employed by large media outlets and also accounts associated with those. |
| Media: Scientific News and Communication | Accounts related to media outlets, publishers, TV shows, radio shows, podcasts, and also personal accounts of journalists and other communicators associated with the media outlets. Professionals employed by outlets more specific to science communication and also accounts associated with those. |
| Media: Other Media | Accounts related to media outlets, publishers, TV shows, radio shows, podcasts, and also personal accounts of journalists and other communicators associated with the media outlets. Individuals and entities broadly related to media, but not with news. For example, podcast hosts or fashion magazines would be in this category. |
| Business | Accounts associated with business such as stores, bars, restaurants, and private services like hair salons or gyms, and individuals associated with businesses. |
| Government and Politics | Accounts associated with local or national governments, political parties, and individuals who are closely involved with these institutions. |
| Public Services | Accounts associated with public services such as high schools or police departments. |
| NGO | Non-governmental political organization, and users who are closely involved with these institutions. Notice that individuals in these categories are likely to be a subset of OPS, so if there is a clear NGO that individuals support, there is no need to also label them as OPS. |
| Political Supporter | Individual accounts associated with political movements. |
| Religion | Accounts associated with notable religious figures and religious leaders/priests as well as religious institutions and entities, tightly associated with these individuals such as temples, congregations, and online sources of religious content. |
| Science: Engineering and Technology | Researchers, scientists, professors, graduate students, professionals, or entities representing or tightly associated with these individuals. Students who are receiving education in a corresponding branch of science (except pre-med, who are categorized as healthcare). Expertise in engineering, computer science or other technology related fields. |
| Science: Life Sciences | Researchers, scientists, professors, graduate students, professionals, or entities representing or tightly associated with these individuals. Students who are receiving education in a corresponding branch of science (except pre-med, who are categorized as healthcare). Expertise in the study of biology, health and environment. |
| Science: Social Sciences | Researchers, scientists, professors, graduate students, professionals, or entities representing or tightly associated with these individuals. Students who are receiving education in a corresponding branch of science (except pre-med, who are categorized as healthcare). Expertise in the study of human societies, policies, economics. |
| Science: Other Sciences | Researchers, scientists, professors, graduate students, professionals, or entities representing or tightly associated with these individuals. Students who are receiving education in a corresponding branch of science (except pre-med, who are categorized as healthcare). Expertise in other fields. If the field of expertise is unclear, it should also be assigned to this category. |
| Healthcare | Professionals that are employed by healthcare institutions, and that directly or indirectly take part in healthcare providing services to patients. Also includes entities representing or tightly associated with these individuals. Includes students of medicine (e.g., premed). |
| Arts and Entertainment | Musicians, actors, plastic artists, writers and entities representing or tightly associated with them. Notice that, similarly to sport, individuals whose hobby is art are not included. If an account belongs to an individual, art is the individual's main occupation. |
| Sports | Athletes and entities representing or tightly associated with them, such as clubs, championships or fan accounts. E-sports are also included, so if someone is a professional video-game player, they should also be included. If an account belongs to an individual, sport is the individual's main occupation. People whose hobby is sport are not included, e.g. having "runner" in a bio does not suggest that the individual belongs to the category. |
| Adult Content | Accounts associated with lewd content. Producer of amateur porn, porn actors or actresses, websites related to porn, and similar. |
| Not in English | Users whose description is not written in English. When labelling these please do not specify the type of account, that is tag them as unclear. |
| Other | Please select this category when none of the others apply. |

Table C.2: The COVID-19 Twitter users taxonomy: **type of account**.

**Account type:** Who does this account represent or belong to?

| Type of account | Description |
|---|---|
| Institution | Account clearly belongs to an institution, an official or unofficial set of individuals. |
| Individual | Account clearly belongs to an individual. |
| Unclear | Account does not clearly belong to a single institution or a single individual. |

Table C.3: The distribution of accounts tweeting about COVID-19 in the complete one week sample, and corresponding number of sampled and annotated accounts, across languages.

| Language | Number of unique accounts | Number of annotated accounts |
|---|---|---|
| English | 89,652 | 1800 |
| Japanese | 33.609 | 1600 |
| Spanish | 36,033 | 1600 |
| Portuguese | 14,813 | 1500 |
| Indonesian | 3291 | 1300 |
| Hindi | 8165 | 1400 |
| French | 4225 | 1300 |
| German | 2205 | 1200 |
| Italian | 1598 | 1200 |
| Arabic | 3357 | 1300 |
| Overall: | 196,948 | 14,200 |

Table C.4: Inter-annotator agreements.

| Language | Category | Type |
|---|---|---|
| Studies languages: | | |
| English | 0.50 | 0.54 |
| Japanese | 0.39 | 0.33 |
| Spanish | 0.39 | 0.51 |
| Portuguese | 0.44 | 0.30 |
| French | 0.25 | 0.34 |
| German | 0.34 | 0.50 |
| Italian | 0.43 | 0.48 |
| Arabic | 0.40 | 0.53 |
| Overall: | 0.43 | 0.44 |
| Omitted languages: | | |
| Hindi | 0.21 | 0.21 |
| Indonesian | 0.22 | 0.24 |

# D Supplementary Information: International expert communities on Twitter become more isolated during the COVID-19 pandemic

## D.1 Data collection

Collection started on January 13, 2020 a few weeks after first reports about a disease outbreak in Wuhan, China surfaced. Throughout the collection period, the keywords were changed in order to accommodate for the various ways the virus was referred to (see table D.2). Initially the virus was referred to as “wuhan virus” and later as 2019-nCoV (2019 novel coronavirus). On February 11 the ICTV (International Committee on Taxonomy of Viruses) changed the official name to sars-cov-2 and COVID-19, for the virus and the disease respectively.

Due to the high volume of data small interruptions occurred during data collection when no data was collected. Four interruptions were for longer than one hour, the longest being 9 hours on April 11.

## D.2 Geo-localization of tweets

In order to geo-localize a tweet the following procedure was performed:

1. Geo coordinates (~0.1% of original tweets in dataset): Tweet contains coordinates (longitude and latitude) information.

2. Place (2.9%): Users can tag a tweet with a named place. Tweets with place indication

contain structured geo information, including a geographical bounding box.

3. Parsable user location (61.9%): We use the Python library `local-geocode`[I] in order to parse the user location field. This field contains unstructured text and may reference one or multiple places and/or countries. It also sometimes contains humorous or imaginary places (e.g. "the end of the universe"). The `local-geocode` library makes use of the geonames database and performs substring matching against place names in this database in order to obtain structured geographical information (also known as geocoding). In the matching, only places with a population larger than 30k are considered. `local-geocode` has been compared against `geopy`[II] (using the Nominatim library), which is frequently used for this task. Visual inspections of the country-level disagreements between both tools, indicate that `local-geocode` generally performs better in this task. This is likely due to the fact that `local-geocode` only considers relatively well known places, therefore ignoring imaginary names whereas geopy attempts to provide a (wrong) result in these cases. However, human-level benchmarking would need to be conducted in order to come to a final conclusion on the performance of both tools for Twitter user location decoding.

## D.3 Network analysis

**Community detection.** We applied Louvain's community detection algorithm (implemented in Python's Networkit package, PLM function), setting the default resolution parameter $\gamma = 1$. Since each run of the algorithm produces different results, we run the algorithm for 50 trials and assigned each user to the community it was mostly found into. On average, about 15 communities reached a size larger than $10^5$ ($15.42 \pm 0.09$) (see Supplementary Fig. D.2). In order to assign each user to a community, we counted how many times each node appeared in the same community along the 50 trials (the same community was hypothesized to be that of maximal overlap within all trials). The ratio of times each node was found in the same community was used as a 0-1 score ("community score") about goodness of identification of the community associated to each node (Supplementary Fig. D.6). Furthermore, we analyzed the overlap of user IDs in communities obtained from the full network and from networks reconstructed with data aggregated per month. Retweets posted during January and February were lower than the rest of observational period, so we joined the two months into a single time-window. This means that four temporary networks were built aggregating the retweets sent during January-February, March, April, May 2020 separately. A fair stability over time was observed overall (see Supplementary Fig. D.7). Temporal stability was highest for the largest communities (labelled from A to H), having an average overlap of 72% (min 44%, max 94%)

[I] `https://github.com/mar-muel/local-geocode`
[II] `https://github.com/geopy/geopy`

with the most overlapping temporary communities. Also smaller communities, in particular L, M, and J, showed a fair temporal stability (avg. overlap 57%, min 20%, max 89%).

**Top users characterization.** The network's communities are composed by users with different roles and centrality inside the network. For a finer characterization of the authorities of this retweet network, we selected as top users the 1000 most retweeted users for each super-community. We computed well-known centrality measures, with the Python's package Networkit, and show their correlation in Supplementary Figure D.9. The node betweenness centrality for each user in the network was estimated considering the shortest paths between 100k randomly sampled nodes. Correlation between centrality measures do not display different patterns for different super-communities. Out-degree and in-degree have the meaning of the number of retweets respectively received and sent. The distribution of received retweets is centered on the highest value for the Political super-community, meaning more attention received. Clustering coefficient is centered on the lowest value for Other, meaning a sparser and less modular community.

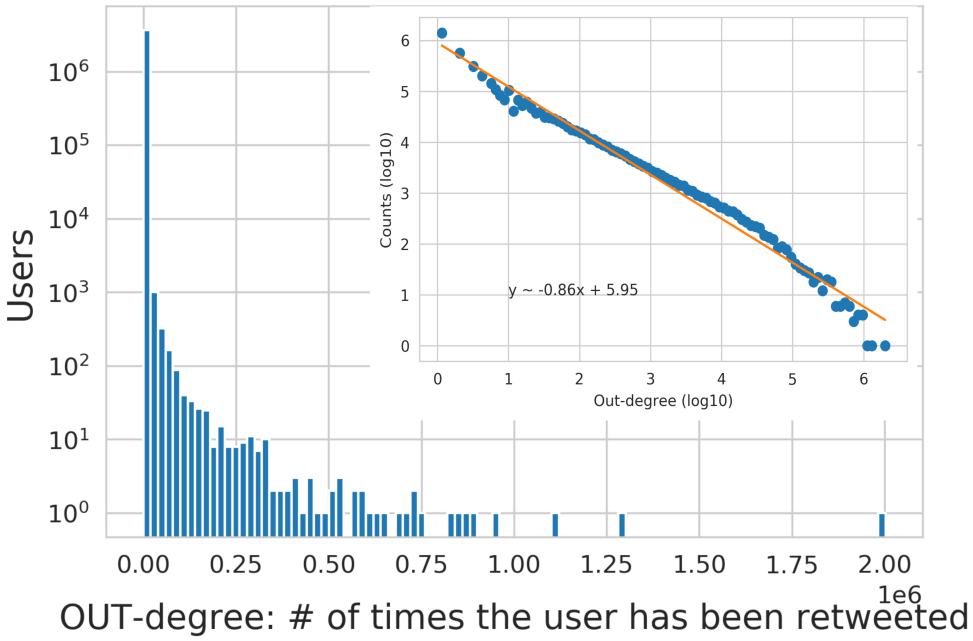



Figure D.1: Weighted out-degree of the retweet network (inbox: log-log plot of the same data).

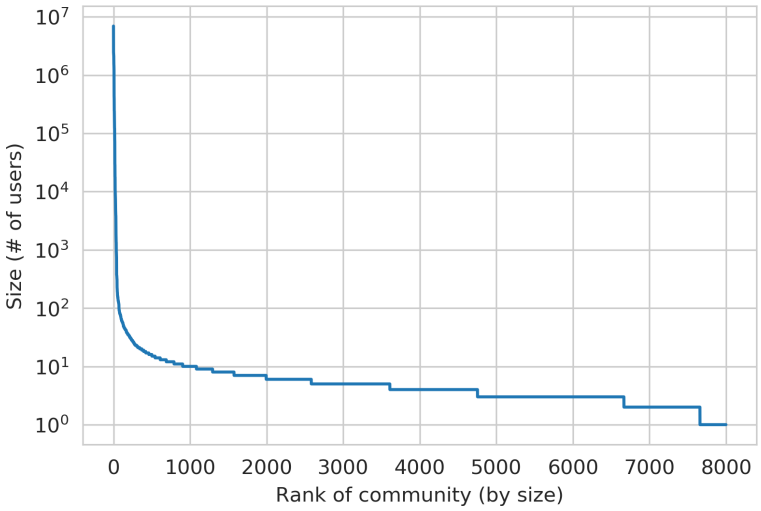



Figure D.2: Size of the communities obtained within one single run of Louvain's community detection algorithm.

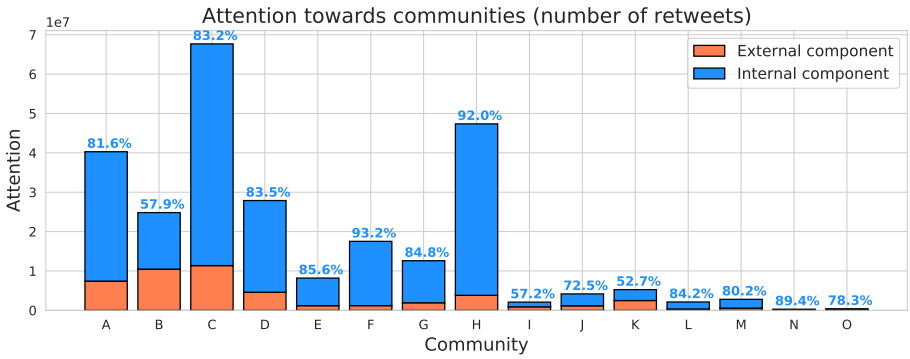



Figure D.3: Internal and external components of the attention towards each community. Each bar represents the number of total retweets received by users in each community, divided into retweets from the same (blue, with the percentage represented on top of each bar) and from other (orange) communities. Community C and H, assigned to the political super-community, are the most retweeted communities, though not the largest. The attention received by these two communities is mainly internal(more than 80%). Community B, assigned to the scientific experts super-community, is the second largest one in terms of number of users but received few overall attention ($5^{th}$ most retweeted in total). Nevertheless, it has a high level of reach, ranking $2^{nd}$ on the highest external attention component.

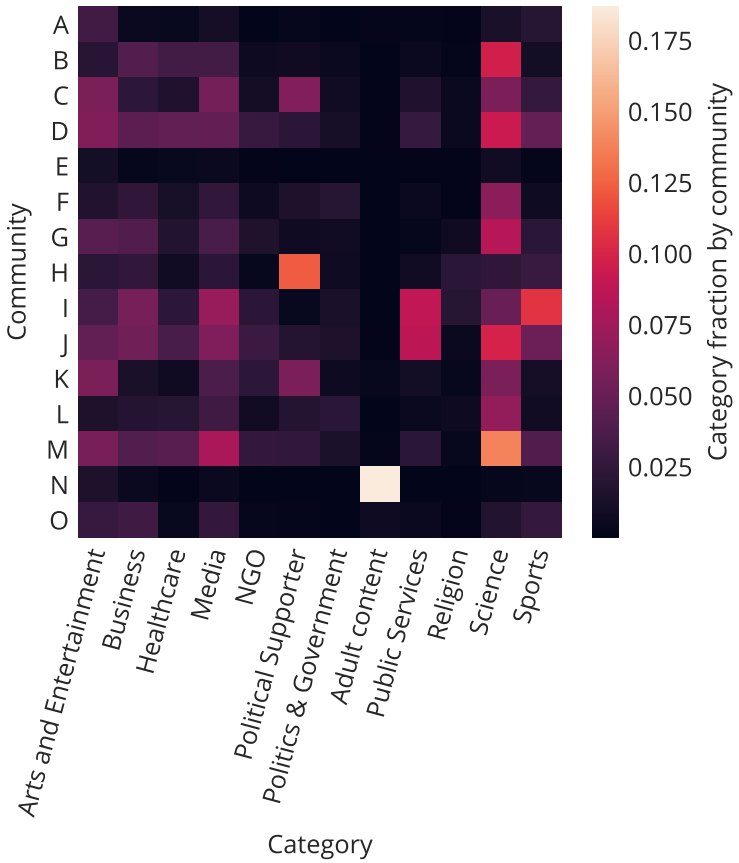



Figure D.4: Heatmap of category fraction by community. Category "Other" is the largest fraction in all communities but not shown in the figure.

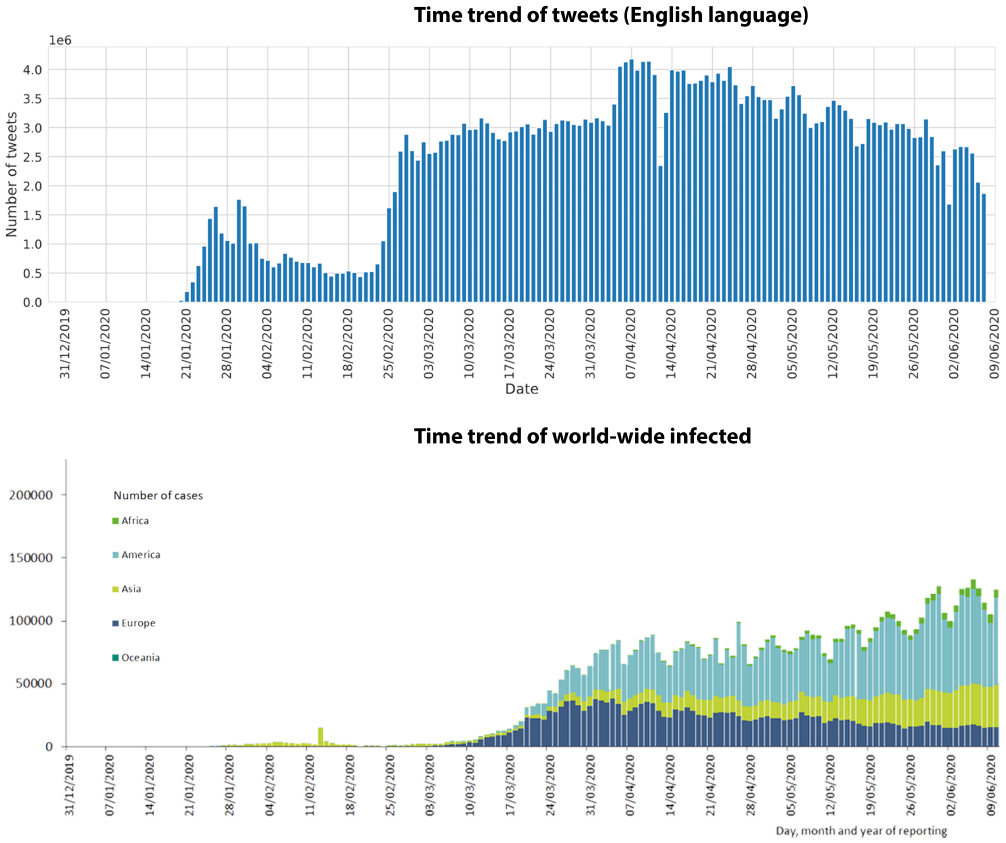



Figure D.5: Top: count of tweets collected daily during the period we observed. Both original tweets and retweets are counted. Bottom: distribution of COVID-19 cases worldwide (website: `https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases`).

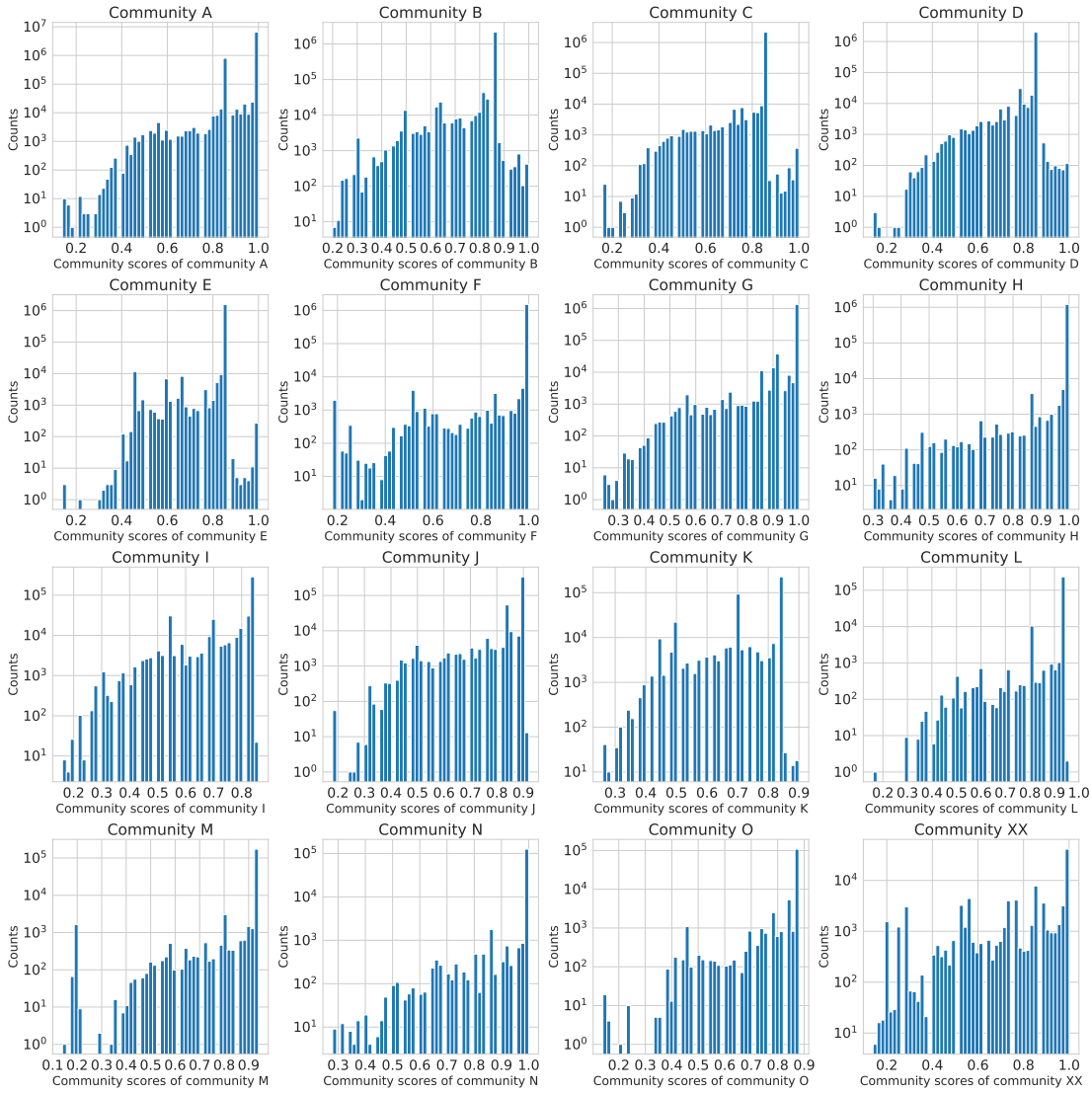



Figure D.6: Community score for the 15 largest communities. Score distribution is shown in semilogarithmic scale: the majority of nodes have a score very close to 1.

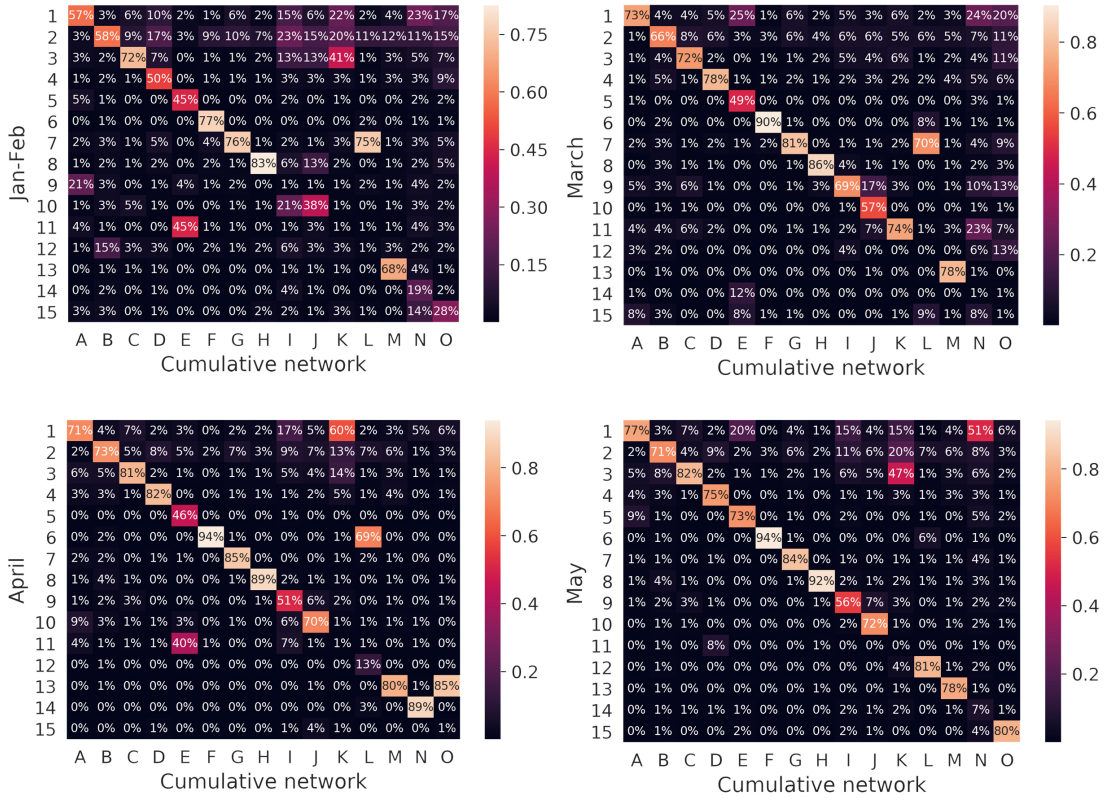



Figure D.7: Overlap of communities detected on the aggregated network (cumulative communities) with respect to communities detected on four time-windows (temporary communities). Each column shows how each cumulative community (x-axis) was distributed through the temporary communities (y-axis). For each heatmap, percentages are computed respect to the total number of users in that time window.

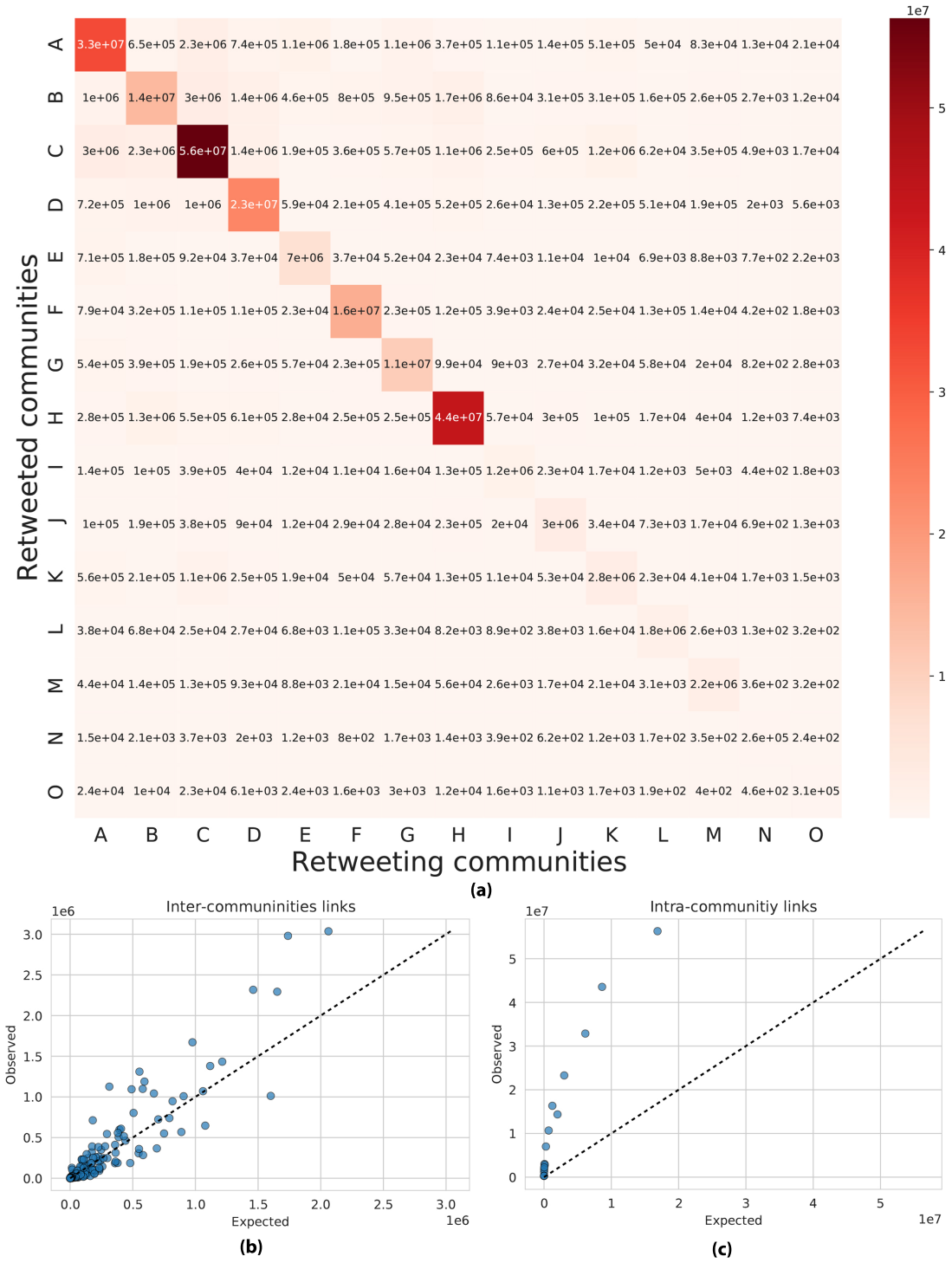



Figure D.8: (a) Mixing matrix of the network, obtained by collapsing all the users belonging to a community into a single node. (b-c) Scatter plot of observed links and expected number of links assuming a random mixing null model between communities. (b) Inter-community links, without considering intra-community retweets. (c) Intra-community links. Communities of users, by construction, have more intra-links than expected by a random mixing null model.

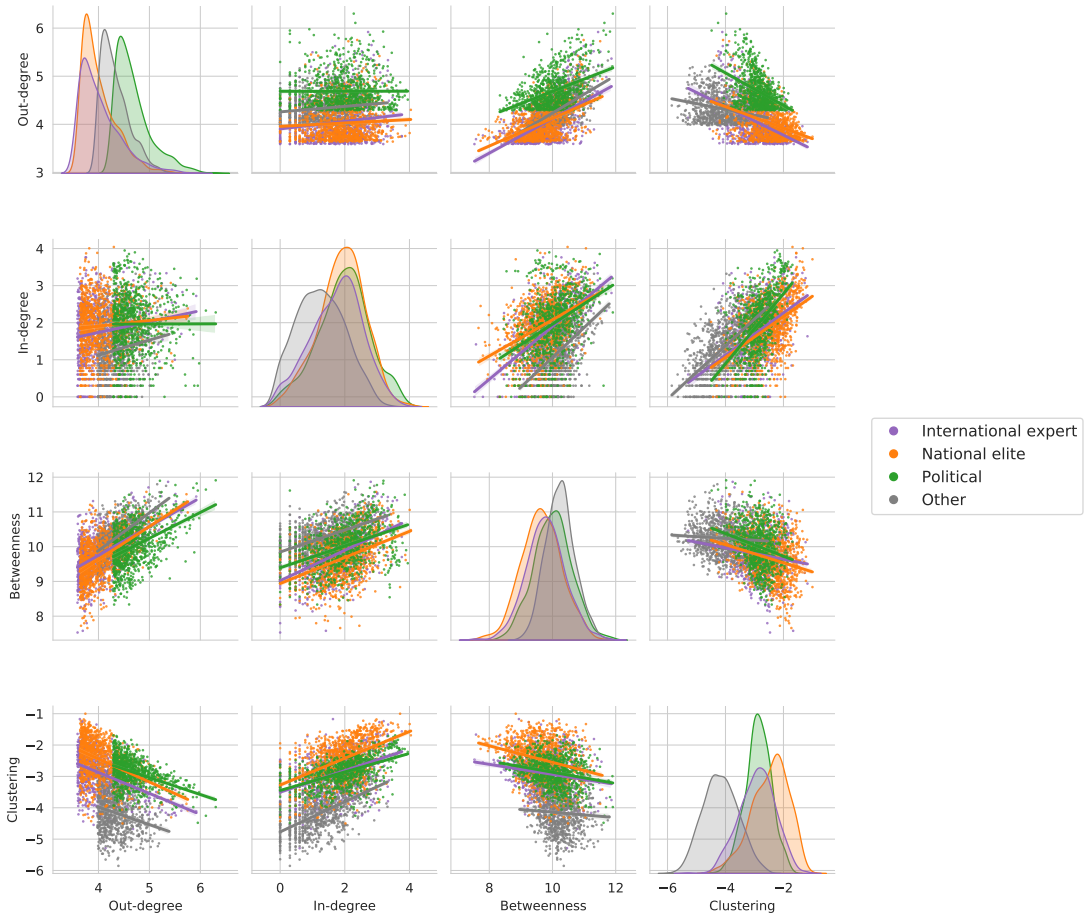



Figure D.9: Centrality measures of the top users in each super-community, portrayed in log-log scale. Top users are chosen as the most retweeted 1000 for each super-community.

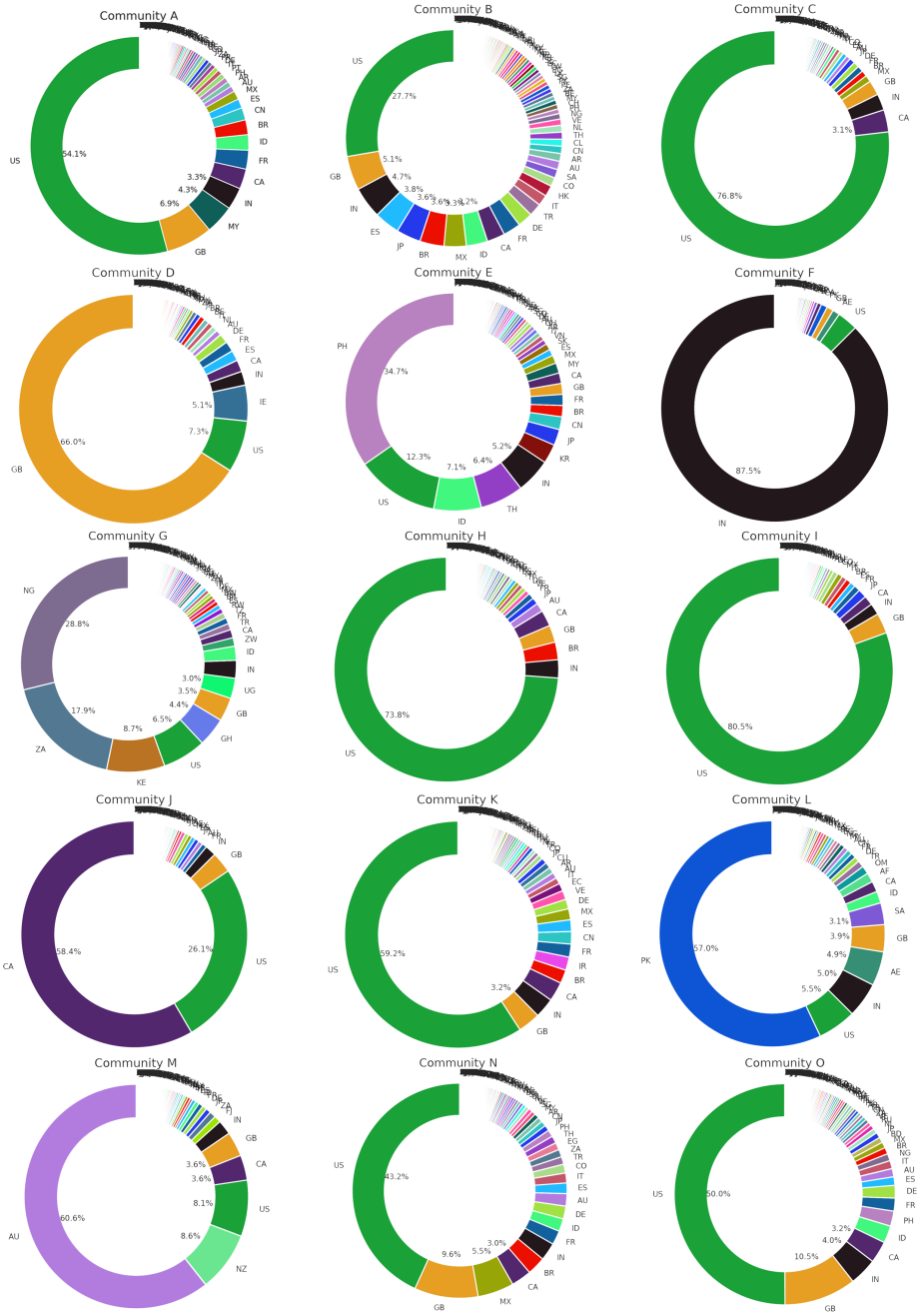



Figure D.10: Pie chart of the location of users, at country level. Each user was assigned to the country code mostly represented in its tweets. Percent value is shown only for countries recurring more than 3% in the community users.

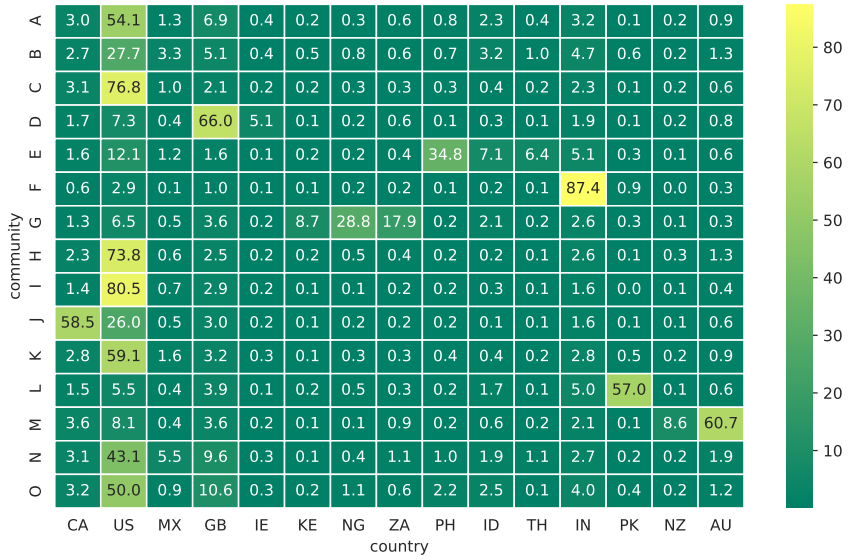



Figure D.11: Heatmap of community locations. Each user was assigned to the country code mostly represented in its tweets. Only country codes represented at least by 5% in a community are displayed.

| **Community** | **2nd largest user category** | **Majority location** | **Number of users** | **Super-community** |
|---|---|---|---|---|
| A | Arts & Entertainment (3.3%) | US | 7,464,665 (33.3%) | Other |
| B | Science (9.7%) | int. | 2,366,768 (10.6%) | International expert |
| C | Political Supporter (6.2%) | US | 2,231,259 (10.0%) | Political |
| D | Science (9.3%) | GB | 2,117,691 (9.4%) | National elite |
| E | Arts & Entertainment (1.0%) | int. | 1,616,006 (7.2%) | Other |
| F | Science (6.5%) | IN | 1,538,840 (6.9%) | Political |
| G | Science (8.4%) | int. | 1,436,377 (6.4%) | International expert |
| H | Political Supporter (12.3%) | US | 1,217,933 (5.4%) | Political |
| I | Sports (10.7%) | US | 465,125 (2.1%) | National elite |
| J | Science (9.9%) | CA | 456,399 (2.0%) | National elite |
| K | Arts & Entertainment (5.9%) | US | 423,077 (1.9%) | Political |
| L | Science (6.9%) | PK | 252,111 (1.1%) | Political |
| M | Science (13.8%) | AU | 186,216 (0.8%) | National elite |
| N | Adult content (18.7%) | US | 133,771 (0.6%) | Other |
| O | Business (3.1%) | US | 124,110 (0.6%) | Other |

Table D.1: Overview of key properties of the 15 largest communities detected in the retweet network. Community name is ordered alphabetically by increasing size. $2^{nd}$ largest category was reported in the table, since category "Other" was the most abundant on for all the communities. Majority location was explicitly reported only when exceeding 50%, indicating "int." for *international* otherwise.

| Date of change | Keywords |
|---|---|
| 2020-01-13 | wuhan |
| 2020-01-14 | wuhan, ncov |
| 2020-01-21 | wuhan, ncov, coronavirus |
| 2020-02-11 | wuhan, ncov, coronavirus, covid |
| 2020-02-18 | wuhan, ncov, coronavirus, covid, sars-cov-2 |

Table D.2: Keywords used to collect data on the Twitter filter stream. Keywords used represent the way the sars-cov-2 virus was referred to at different points in time.

# Curriculum Vitae

**Martin Müller**, born July 20th, 1988

## Education

| | |
|---|---|
| **2016 - (2020)** | **PhD, Digital Epidemiology Lab, EPFL**<br>with Prof. Marcel Salathé |
| **2012 - 2015** | **MSc in Computational Biology & Bioinformatics, ETH Zurich**<br>MSc thesis with Prof. Sebastian Bonhoeffer |
| **2009 - 2012** | **BSc in Biochemistry, ETH Zurich** |

## Experience

| | |
|---|---|
| **2020** | **Part-time consultant**<br>Coteries SA |
| **2016 - 2020** | **Teaching assistant**<br>Introduction to C++, Issues in Global Health |
| **2015 - 2016** | **Research assistant**<br>Theoretical Biology, Lab of Prof. Sebastian Bonhoeffer |

## Extracurricular activity

| | |
|---|---|
| **2020** | **Co-organized "Meet your Artificial Self: Generate text that sounds like you"**<br>Top-rated workshop at Applied Machine Learning Days 2020 |
| **2018-2019** | **Co-organized "Open Science in Practice"**<br>EPFL summer school |

## Publications

**Müller, M.**, Salathé, M. (2020). *Addressing machine learning concept drift reveals declining vaccine sentiment during the COVID-19 pandemic.* arXiv preprint arXiv:2012.02197.

Nsoesie, E. O., Cesare, N., **Müller, M.**, Ozonoff, A. (2020). *Characterizing the Spread of COVID-19 Misinformation in Eight Countries Using Exponential Growth Models.* Journal of Medical Internet Research.

Durazzi, F., **Müller, M.**, Salathé, M., Remondini, D. (2020). *International expert communities on Twitter become more isolated during the COVID-19 pandemic.* arXiv preprint arXiv:2011.06845.

Gligorić, K., Ribeiro, M. H., **Müller, M.**, Altunina, O., Peyrard, M., Salathé, M., Colavizza, G., West, R. (2020). *Experts and authorities receive disproportionate attention on Twitter during the COVID-19 crisis.* arXiv preprint arXiv:2008.08364.

Edelstein, M., **Müller, M.**, Ladhani, S., Yarwood, J., Salathé, M., Ramsay, M. (2020). *Keep calm and carry on vaccinating: Is anti-vaccination sentiment contributing to declining vaccine coverage in England?.* Vaccine, 38(33), 5297-5304.

**Müller, M.**, Salathé, M., Kummervold, P. E. (2020). *COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter.* arXiv preprint arXiv:2005.07503.

**Müller, M.**, Salathé, M. (2019). *Crowdbreaks: Tracking health trends using public social media data and crowdsourcing.* Frontiers in public health, 7, 81.

**Müller, M.**, Schneider, M., Salathé, M., Vayena, E. (2020). *Assessing Public Opinion on CRISPR-Cas9: Combining Crowdsourcing and Deep Learning.* Journal of medical Internet research, 22(8), e17830.

Kogan, N. E., Bolon, I., Ray, N., Alcoba, G., Fernandez-Marquez, J. L., **Müller, M.**, Mohanty, S. P., de Castañeda, R. R. (2019). *Wet markets and food safety: TripAdvisor for improved global digital surveillance.* JMIR public health and surveillance, 5(2), e11477.