2021

Developing natural language processing instruments to study sociotechnical systems

Thayer Alshaabi
University of Vermont

Developing natural language processing instruments to study sociotechnical systems

A Dissertation Presented

by

Thayer Alshaabi

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Complex Systems and Data Science

August, 2021

Defense Date: July 8th, 2021
Dissertation Examination Committee:

Peter Sheridan Dodds, Ph.D., Advisor
Christopher M. Danforth, Ph.D., Advisor
Randall Harp, Ph.D., Chairperson
Emma Tosch, Ph.D.
Jeremiah Onaolapo, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

Identifying temporal linguistic patterns and tracing social amplification across communities has always been vital to understanding modern sociotechnical systems. Now, well into the age of information technology, the growing digitization of text archives powered by machine learning systems has enabled an enormous number of interdisciplinary studies to examine the coevolution of language and culture. However, most research in that domain investigates formal textual records, such as books and newspapers. In this work, I argue that the study of conversational text derived from social media is just as important. I present four case studies to identify and investigate societal developments in longitudinal social media streams with high temporal resolution spanning over 100 languages. These case studies show how everyday conversations on social media encode a unique perspective that is often complementary to observations derived from more formal texts. This unique perspective improves our understanding of modern sociotechnical systems and enables future research in computational linguistics, social science, and behavioral science.

Citations

Material from this dissertation has been published in the following form:

Alshaabi, T., Adams, J. L., Arnold, M. V., Minot, J. R., Dewhurst, D. R., Reagan, A. J., Danforth, C. M., and Dodds, P. S.. (2021). Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. *Science Advances*, 7(29), eabe6534.

Alshaabi, T., Dewhurst, D. R., Minot, J. R., Arnold, M. V., Adams, J. L., Danforth, C. M., and Dodds, P. S.. (2021). The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. *EPJ Data Science*, 10(1), 1-28.

Alshaabi, T., Arnold, M. V., Minot, J. R., Adams, J. L., Dewhurst, D. R., Reagan, A. J., Muhamad, R., Danforth, C. M., and Dodds, P. S.. (2021). How the world's collective attention is being paid to a pandemic: COVID-19 related n-gram time series for 24 languages on Twitter. *Plos One*, 16(1), e0244476.

Alshaabi, T., Van Oort, C. M., Fudolig, M., Arnold, M. V., Danforth, C. M., and Dodds, P. S.. (2021). Augmenting semantic lexicons using word embeddings and transfer learning. Manuscript in preparation.

To the memory of the kindest women I have ever met in my life.
To my sisters, Haya and Yara, for their unconditional love and support.
To my parents who supported me every step along the way.
To Ken Stavisky and Janice Smith for their soothing love and companionship.

P.S. *It has long been said that the mind abhors a vacuum of interpretation. It is fascinating how we (humans) are too afraid of not knowing the texture of our realities, extremely so, that we enclose our knowledge horizons to avoid our relentless fear of uncertainty. Yet, I have learned that we have to ultimately get comfortable with that to be scientists—thriving at the event horizon of knowledge, and the edge of ignorance. Every time we unfold a chapter of the unknown world, we simultaneously widen our circle of knowledge and dive deeper into the world of mysteries. I lose sleep wondering how not knowing the answers for most of our questions is, of course, conceptually frightening and unsettling, but the questions that we have not learned to ask yet may shatter our inception altogether...*

Acknowledgements

Looking back at the past three years, my journey in graduate school has been relatively short. Needless to say, it did not fall short of unforgettable memories. While it has been undoubtedly exhausting, not to mention living through a global pandemic, I am deeply honored to have spent this time surrounded by wonderful and smart people. So, please bear with me as I attempt to sort through the utter madness and scattered thoughts in my mind, stumbling over my words to thank you all for being part of my journey.

This would have not been remotely possible without the personal and academic support and friendship of my dear advisors, Peter Dodds and Chris Danforth. I am genuinely thrilled to have had the opportunity to work with you. Over the years, you have gathered a team of exceptionally brilliant minds, nurturing a lovely work environment for everyone to thrive and succeed. There are no words that can truly portray my respect and appreciation for your guidance.

I would like to express my deepest gratitude to my committee members, Randall Harp, Emma Tosch, and Jeremiah Onaolapo. Thank you for committing to be on my dissertation committee. Your kindness and thoughtful advice made it possible for me to be here. Many thanks go to the supportive community and professors with whom I took classes, and discussed a numerous number of thoughts and reflections. I very much appreciate the generous support from Safwan Wshah, Brian Tivnan, and Mads Almassalkhi. Thank you in particular to James Bagrow, Jason Bates, Laurent Hébert-Dufresne, Joe Near, and Chris Skalka for their collaboration on many exciting projects throughout my time at UVM.

Without friends, even the greatest pursuits become infinitely meaningless. I am grateful to Narine and Brian Hall for their inspiring mentorship and friendship. Thank you for building my confidence and recommending that I pursue a graduate degree. Without you, I would have not been here struggling to write the last few words of my dissertation to thank everyone. Big thank you to Victoria Manukyan for her devoted kindness and enthusiastic support. Knowing you have taught me that enduring the impossible is possible in pursuit of our passions and dreams.

I am especially thankful to my best mates, Eric Cacciavillani and Colin Van Oort for the countless hours we spent working and hanging out together. I am always inspired by their superb work ethic. It has been truly an honor to work with you gentlemen. To Anne Marie Stupinski and Vanessa Myhaver for their wonderful companionship and support. I am incredibly proud of you and so grateful for being part of your world. To John Ring and Axel Masquelin for their heartfelt friendship. And, of course, I can never forget to thank the one and only Melissa Rubinchuk. I would have not made it here without your exceptional help.

Special thanks to my workmates and friends at the Computational Story Lab. To Michael Arnold for the unforgettable and fun days we spent together, debugging code and hashing ridiculous thoughts. To Josh Minot, Jane Adams, and David Dewhurst for their outstanding work on the Storywrangler project. Many more thanks to my colleagues Mikaela Fudolig, Sage Hahn, Henry Wu, Kelly Gothard, Juniper Lovato, Sophie Hodson, Max Green, Kelsey Linnell, Julia Zimmerman, Adam Fox, Marc Maier, Yi Li, Xiangdong Gu, and Andy Reagan. It has been an absolute pleasure to have worked with you all.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

A sociotechnical system is a complex system that involves social and technological elements. Historically, the term is grounded in system theories that were primarily developed to optimize the workflow of industrial workers during the aftermath of World War II [90, 287]. In the age of information technology, the term provides a macroscopic description of social and interactive digital infrastructures (e.g., online services, e-commerce, social media) [103]. Indeed, social structures have evolved because of a growing rate of automation and innovations [107, 286]. Studying the interplay between society and technology is vital in light of these rapid technological changes. Understanding the throughput of a sociotechnical system is particularly helpful to study long-term societal changes and the coevolution of technical domains and social sciences [18].

Language is a key component of these sociotechnical systems—the technology that enables participants to communicate through a complex system. Consequently, a sociotechnical system is a medium through which language and culture co-evolve [97, 137]. Recent advances in natural language processing (NLP) such as optical character recognition (OCR), and automatic speech recognition (also known as, speech-to-text), have enabled a vast collection of rich and interdisciplinary studies (e.g., books [197] and news [132]). The emerging digitization of text archives has also empowered new NLP applications such as machine translation, speech tagging, question answering, text summarization, and sentiment analysis—most of which are primarily powered by artificial neural networks and machine learning [189, 315].

1.1.1 Culturomics

Data mining of digital archives enables users to scan through an extensive collection of databases for words, tracing their inception and tracking their rate of usage over time. The term 'culturomics'—originally proposed by Michel et al. [197]—is often used to refer to this vast collection of studies of human behavior and sociocultural trends portrayed in language usage. The Google n-grams viewer [197], which inspires the work presented here, is a notable example that provides year-scale n-gram usage time series derived from most books printed over the past century. While the cultural significance of the time series provided by the Google Books project has been duly noted in the scientific literature, researchers have also highlighted some pitfalls [156, 195].

The Google Books framework records n-gram usage once for each book per year, which fails to encode the cultural popularity of these n-grams, ignoring the millions of copies sold of these books [221]. Of course, measuring cultural popularity is remarkably difficult. Databases that track sales of books over time are sparse, and large-scale archives that attempt to capture the popularity of n-grams are rare because the data needed to compile such records is limited, prohibitively expensive, and often proprietary [4]. Nevertheless, identifying linguistic patterns in digitized archives has proven to have a tremendous societal impact beyond the readily seen observations. Researchers show it is possible to predict political unrest by analyzing temporal trends found in a large digital news archive [132, 173]. Other studies extend this form of computational lexicology, assessing the foreseeable cultural impact of conservation interventions [166], and examining gender bias in newspapers [96].

Availability of data is arguably one of the key drivers of these technical advances. The positive feedback loop between machine learning systems and computational linguistics has been essential to further enlighten our understanding of language usage in sociotechnical systems. Despite the recent accomplishments, most large-scale studies focus on stylized and copyedited corpora such as books, news articles, and other formal records [132, 197, 198, 257, 308]. Google has also created a similar framework to track n-gram usage time series in search data, aptly named Google Trends.[1] Many studies have showed the utility of using search data to predict and forecast sociotechnical trends [263], such as economic indicators [54], disease outbreaks [45], and health care [211].

1.1.2 Social media

Over the past decade, however, we have observed an unprecedented growth of social media, giving birth to large-scale sociotechnical systems with billions of activities taking place in real time across the entire globe [152]. The rise of social media platforms has enabled people, media outlets, organizations, and chatbots to share content freely, transcending physical boundaries. Messages are shared instantly within nanoseconds, featuring a high temporal resolution to capture and study sociocultural phenomena.

The decentralized nature of various activities shared on these platforms forms a distributed sociotechnical sensor system, providing a rich lexicon, with emerging perils, to track and trace trending storylines in real time [9, 146, 212]. Unlike mainstream media outlets, social media platforms provide unique conversational data streams to

[1] https://googleblog.blogspot.com/2007/09/its-all-about-today.html

examine daily discussions and reactions by millions of people on a scale that is inadequately captured and analyzed [13, 260]. Social media encodes casual daily conversation in a format that is simply unavailable through other outlets such as newspapers, and books. The ever-growing compendium of daily discourse on Twitter, and other social media platforms, have already made an unprecedented impact on modern industrial societies (e.g., #ArabSpring, #MeToo, and #BlackLivesMatter) [146].

Mining social media data, particularly Twitter, is useful for marketing, product development, risk management, and brand interactions [43, 218]. Combined with sentiment analysis, researchers show that analyzing social media data can be used to predict box office sales [204], and global financial trends [35]. Beyond these notable applications, researchers also show how the complementary signal derived from social media can be brought into play for gauging public opinion on pending policies [171, 283], and monitoring inflammatory discourse [217].

Importantly, these platforms support various communication channels through which people can discuss and share content—encoding an essential property that allows researchers to quantify the popularity (i.e., social amplification) of trending topics [147]. When sharing mechanisms are native to these platforms, we can quantify the collective attention of the public to certain topics. Social amplification is also rarely encoded in historical text archives. It not only allows us to track and potentially predict trending storylines, but it can also help us disentangle how information flows and spreads across communities, identifying misinformation, and tracing disinformation. Many studies have addressed theoretical models of social amplification but without adequate data [40, 118, 126, 277].

1.2 A brief overview

With the exception of Google Trends, existing tools have only focused on easily accessible and formal digital archives, such as folklore [198], government records [308], newspapers [257], books [197], radio transcripts [19], and TV news transcripts [132]. However, there is a growing volume of text data on social media that renders human annotation infeasible for real-time data streams. Sophisticated instruments would be needed to help us understand how information flows and persists across communities on social media platforms. While some data streams that are derived from predominant outlets such as Facebook and Instagram are locked behind corporate doors, Twitter and Reddit, among a few others, share their data with researchers across disciplines.

Building on the state-of-the-art research of culturomics, I describe the development of a couple of new tools to help researchers, journalists, and data scientists study the vast universe of stories on social media, particularly Twitter. I present Storywrangler, an evolving research instrument powered by machine learning and a staggering computing cluster to track and trace usage rates of words for the majority of spoken languages in trillions of messages on Twitter. In a series of case studies, I show how Storywrangler can document world events in real time, capturing narratively trending storylines to examine heated political debates, rising social movements, emerging developments and outbreaks, neologisms, memes, emojis, and the quotidian.

Vitally, and problematically absent from existing text corpora such as books and news archives, Storywrangler also encodes popularity (i.e. social amplification) of n-grams by tracking their rate of usage across retweets. To explore the interplay of social

contagion on Twitter, we investigate the relative amplification of n-grams visualized through 'contagiograms', a bespoke Python package to examine word rankings and social amplification. Although Storywrangler leverages Twitter data, our method of extracting and tracking dynamic changes of n-grams can be extended to any similar social media platform.

Finally, I describe the integration of Storywrangler into existing sociotechnical instruments, particularly, the Hedonometer—an instrument that measures the daily rate of happiness on Twitter [80]. Using word embeddings and transfer learning, I present a new tool for augmenting semantic dictionaries, such as labMT [81]. The proposed framework reduces the need for crowdsourcing annotations and provides better accuracy when compared with a random set of reviewers from Amazon Mechanical Turk. While the new method can be fine-tuned to predict scores for any semantic lexicon, we focus on predicting happiness scores for the Hedonometer to capture the emotional valance of various events on social media.

1.2.1 Sociocultural significance

Although other topics, such as breaking news and major stories, are documented in mainstream media, social media provides a unique outlet for ephemeral conversations that are not well captured using other data sources. While it might seem frivolous to track discourse on Twitter ranging from political movements to K-pop to sports to music and favorite movie stars, the banal everyday sociocultural trends encode a unique perspective that is often complementary to observations derived from more formal texts to examine the coevolution of language and culture. Storywrangler features a data-driven approach to index what regular people are talking about in

everyday conversations, in addition (or contrast) to what reporters and authors have shared with the public.

Storywrangler tracks the periodic signal of words related to religious festivals and the collective attention for international sports events. It shows how n-grams connected with new movies and TV series burst into social media then slowly decay, making direct comparisons with their worldwide box-office earnings. Storywrangler revels how marketing campaigns can take advantage of the periodic nature of narratively trending n-grams, exploiting popular hashtags to amplify their message (e.g., including #FF and #TGIF as trending hashtags for Friday promotions).

The time series derived from Storywrangler can also be cross-referenced with other data repositories to enable data-driven, computational versions of journalism, linguistics, history, economics, and political science. For example, changes in salient word usage (e.g., 'rebellion, and 'crackdown') are significantly associated with future changes in geopolitical risk and lower stock returns (see Appendix 3.E). We show how words that are linked to the COVID-19 outbreak and the percent change in how frequently they are used can be correlated with similar volatility in the number of reported coronavirus cases and deaths a couple of weeks later across 24 languages on Twitter [6]. Expressions and hashtags associated with political movements can be directly compared with recorded incidents of fatal police violence [309]. Tracking hurricane name mentions, we find different temporal patterns of collective attention correlated with deaths and damage reportings [8].

Social media, as an example of a growing large-scale social structure, is essentially a platform whereby individuals can form, share, and reshape their social behavior and others [107]. For instance, a single n-gram can embody a sociocultural change

(e.g., #MeToo, #BlackLivesMatter), leaving a remarkable footprint on social media before it ever became part of the formal, mainstream conversation. Storywrangler creates a digital record of the collective attention on social media to these emerging sociotechnical phenomena, documented in the daily n-gram usage rates. Having a large-scale daily record of such sociotechnical trends creates a transformative potential across disciplines.

1.2.2 Implications and limitations

Sample size For a primary social media source, we use Twitter as it provides an open research platform while acknowledging its limitations. Twitter's userbase is not representative of all voices [193], but It provides an outlet for all people to voluntarily carry out conversations that matter to their lives. We use roughly 10% of all tweets ever posted on the platform, thus our tool presents an approximate daily leaderboard of heavy-tailed n-gram Zipf distributions [322]. Researchers have inspected ways to study Zipf distributions and estimate the robustness and stability of their tails [33, 120, 229, 234]. Investigators have also examined various aspects of Twitter's Sample API [227], and how that may affect the observed daily n-gram frequency-of-usage distributions [303]. We provide a brief analysis of the lexicon in Appendix 3.A, describing the distributions of n-grams in our dataset. It is a modest attempt to explore and examine the temporal lexical distributions of n-grams.

However, half of the words that appear in a corpus will appear only once [120, 234, 322]. While the length of phrases is limited on Twitter (140 characters prior to the last few months of 2017, 280 thereafter[2]), the numbers of unique bigrams and

[2] https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html

trigrams strongly outweigh the number of unique unigrams because of the combinatorial properties of language. Future work can shed light on the changes in the lexicon over time. Tweet sampling may render this task difficult in practice, especially for less-used languages on the platform. Regardless, we are unable to make assertions about the size of Twitter's user base or message volume. We do not have knowledge of Twitter's overall volume (and do not seek to per Twitter's terms of service). We deliberately focus on ranks and relative usage rates for n-grams away from the tails of their distributions. Raw frequencies of exceedingly rare words are roughly one-tenth of the true values regarding all of Twitter, however, rankings are likely to be subject to change.

Language coverage In the first case study, we describe how we detect language labels of all tweets in our collection using FastText-LID [32, 144]. A uniform language re-identification is necessary as Twitter's real-time identification algorithm was introduced in late 2012 and then adjusted over time, resulting in temporal inconsistencies. The word embeddings provided by FastText span a wide set of languages, including some regional dialects (see Table 2.A.1 for the full list of languages detectable by FastText-LID). However, language detection of short-text remains an outstanding challenge in NLP. While we hope to expand our language detection in future work, we still classify messages based on the languages identified by FastText-LID. We use FastText-LID as a light, fast, and reasonably accurate language detection tool to overcome the challenge of missing language labels in our Twitter historical feed. Although Storywrangler can detect continuous-script languages, such as Japanese and Chinese, it is unable to parse their tweets into n-grams because of technical limitations. However, we would like to emphasize that Storywrangler still features a wide

range of non-western languages such as Hindi, Indonesian, Korean, Bengali, Nepali, Arabic, Turkish, Persian, and Hebrew.

Social amplification There are substantive limitations to Twitter data, some of which are evident in many large-scale text corpora. Our n-gram dataset contends with popularity, allowing for the examination of story amplification, and we emphasize the importance of using contagiograms as visualization tools that go beyond presenting simple time series. Popularity, however, is notoriously difficult to measure. The main proxy we use for popularity is the relative rate of usage of a given n-gram across originally authored tweets, examining how each term or phrase is socially amplified via retweets. While Twitter attempts to measure popularity by counting impressions, it is increasingly difficult to capture the number of people exposed to a tweet. Twitter's centralized trending feature is yet another dimension that alters the popularity of terms on the platform, personalizing each user timeline and inherently amplifying algorithmic bias. We have also observed a growing passive behavior across the platform, leading to an increasing preference for retweets over original tweets for most languages on Twitter during the past few years [7].

Ethical considerations In building Storywrangler, we have prioritized privacy by aggregating statistics to day-scale resolution for individual languages, truncating distributions, ignoring geography, and masking all metadata. We have also endeavored to make our work as transparent as possible by releasing all code associated with the API. Although we frame Storywrangler as a research focused instrument akin to a microscope or telescope for the advancement of science, it does not have built-in ethical guardrails.

There is potential for misinterpretation and mischaracterization of the data, whether purposeful or not. We strongly caution against cherry picking isolated time series that might suggest a particular story or social trend. Words and phrases may drift in meaning and other terms take their place. For example, 'coronavirus' gave way to 'covid' as the dominant term of reference on Twitter for the COVID-19 pandemic in the first six months of 2020 [6]. To in part properly demonstrate a trend, researchers would need to at least marshal together thematically related n-grams, and do so in a data-driven way, as we have attempted to do for our case studies. Thoughtful consideration of overall and normalized frequency of usage would also be needed to show whether a topic is changing in real volume.

1.2.3 Outline

The work is organized into a series of four case studies, illustrating the crucial value of developing NLP instruments to study social media platforms at scale. In the first case study, we investigate the dynamics of social amplification in a sociotechnical system, examining the daily usage of over 100 languages on Twitter throughout the past decade. Building on our previous work, we present Storywrangler in the second case study, an instrument that extracts sociotechnical time series of words and phrases from social media data streams, automatically capturing narratively trending storylines.

In the third case study, we curate a set of 2000 day-scale time series of unigrams and bigrams across 24 languages that are most salient to the COVID-19 pandemic as a data repository for current and retrospective investigations. Using Storywrangler and the proposed dataset of the most narratively dominant n-grams we upgrade the

Hedonometer [80], amplifying the tool's utility to capture the sentiment of unfolding events in real time.

In the fourth and last case study, we propose a framework for augmenting semantic lexicons using transfer learning, reducing the need for crowdsourcing scores from human annotators. Although our framework can be used in a more general sense, we focus on predicting *happiness scores* for the Hedonometer and the labMT dataset [81].

Throughout each chapter, I discuss direct applications of the proposed instruments in multilingual and longitudinal social media data streams. Comparing the utility of these instruments with similar frameworks, I highlight their benefits while acknowledging existing limitations and outlining future developments.

Chapter 2

Exploring sociolinguistic amplification in textual archives

2.1 Abstract

Working from a dataset of 118 billion messages running from the start of 2009 to the end of 2019, we identify and explore the relative daily use of over 150 languages on Twitter. We find that eight languages comprise 80% of all tweets, with English, Japanese, Spanish, Arabic, and Portuguese being the most dominant. To quantify social spreading in each language over time, we compute the 'contagion ratio': The balance of retweets to organic messages. We find that for the most common languages on Twitter there is a growing tendency, though not universal, to retweet rather than share new content. By the end of 2019, the contagion ratios for half of the top 30 languages, including English and Spanish, had reached above 1—the naive contagion threshold. In 2019, the top 5 languages with the highest average daily ratios were, in order, Thai (7.3), Hindi, Tamil, Urdu, and Catalan, while the bottom 5 were Russian, Swedish, Esperanto, Cebuano, and Finnish (0.26). Further, we show that over time, the contagion ratios for most common languages are growing more strongly than those of rare languages.

2.2 Introduction

Users of social media are presented with a choice: post nothing at all; post something original; or re-post ("retweet" in the case of Twitter) an existing post. The simple amplifying mechanism of reposting encodes a unique digital and behavioral aspect of social contagion, with increasingly important ramifications as interactions and conversations on social media platforms such as Twitter tend to mirror the dynamics of major global and local events [40, 125, 208, 277].

Previous studies have explored the role of retweeting in the social contagion literature, though the vast majority of this research is limited to either a given language (e.g., English tweets) or a short period [40, 118, 126, 277]. Here, drawing on a 10% random sample from over a decade's worth of tweets, we track the rate of originally authored messages, retweets, and social amplification for over 100 languages.

We describe distinct usage patterns of retweets for certain populations. For example, Thai, Korean, and Hindi have the highest contagion ratios, while Japanese, Russian, Swedish, and Finish lie at the other end of the spectrum. While there is a wide range of motives and practices associated with retweeting, our object of study is the simple differentiation of observed behavior between the act of replication of *anything* and the act of *de novo* generation (i.e., between retweeted and what we will call organic messages).

We acknowledge two important limitations from the start. First, while it may be tempting to naively view ideas spreading as infectious diseases, the analogy falls well short of capturing the full gamut of social contagion mechanisms [25, 47, 66, 69, 77, 78, 108, 111, 259, 295], and a full understanding of social contagion remains to be

established. And second, while higher contagion ratios are in part due to active social amplification by users, they may also, for example, reflect changes in Twitter's design of the retweet feature, changes in demographics, or changes in a population's general familiarity with social media. Future work will shed light on the psychological and behavioral drivers for the use of retweets in each language across geographical and societal markers, including countries and communities.

2.3 Background and motivation

Social contagion has been extensively studied across many disciplines including marketing [15, 139, 289, 297], finance [59, 93, 123, 150], sociology [39, 114], and medicine [55, 219, 233]. Because it can be easier to access data on human social behavior from social media outlets than from other sources such as in-person or text-message conversations, social contagion dynamics are often examined in the context of messages posted and subsequently re-posted on social media platforms [36, 89, 94, 159]. Indeed, the flow of information in the context of social contagion on digital media outlets, especially Twitter, has been widely studied over the last decade [126, 174], with attention paid to the spreading of certain kinds of messages, such as rumours [38, 145, 165, 214, 324], misinformation and "fake news" [70, 262, 268, 285]. Several models have also been proposed to predict the spread of information on Twitter [318], while other models have shown the differences in which various topics can propagate throughout social networks [247, 301]. Studies have also investigated the extent to which information spread on Twitter can have an echo chamber effect [14, 61].

The body of research shows overwhelming evidence that retweeting is a key instru-

ment of social contagion on Twitter [208, 273]. One of the earliest analysis of Twitter by Kwak et al. [164] suggests that a retweet can reach an average of a thousand users regardless of the social network of its original author, spreading its content instantly across different hubs of the full Twitter social network. While seemingly simple, there are different styles and drivers of retweeting [40]. The practice of retweeting has become a convention on Twitter to spread information, especially for celebrities. Researchers argue celebrities can act as hubs of social contagion by studying the flow of retweets across their focal networks [118]. Recent work shows how retweets of officials can be either alarming or reassuring amid the COVID–19 pandemic [207, 239]. Statistical features of retweets reveal a strong association between links and hashtags in most retweeted messages [277]. Retweeting is not only an act in which users can spread information, but a mechanism for actors to become involved in a conversation without being active participants [40]. The use of retweets empirically alters the visibility of information and how fast messages can spread on the platform [125].

Other studies have quantified language usage on social media [48, 95], particularly on Twitter [34, 153]. While investigators have studied the use of retweets in the context of social contagion using network-based approaches [91, 174, 207, 247], little research has been done regarding the statistical variability of retweets across the vast majority of languages. In this study, by applying an updated language identification (LID) process to over a decade of Twitter messages, we explore a macroscopic description of social contagion through the use of retweets across languages on Twitter. Our study addresses a unique property of social contagion on Twitter by statistically quantifying the rate of retweets in each language. We show how the practice of retweeting varies across different languages and how retweeting naturally lends itself

to micro-level discussions of social contagion on Twitter, which can also be extended to other social media outlets with similar features.

We structure this chapter as follows. First, we discuss the state-of-the-art tools presently used for language detection of short and informal messages (e.g., tweets). We then describe our dataset and processing pipeline to answer some key questions regarding social contagion through the use of retweets. Based on our considerations, we deploy FastText-LID [32, 143] to identify and explore the evolution of 100+ languages in over 118 billion messages collected via Twitter's 10% random sample (decahose) from 2009 to 2020 [292].

For messages posted after 2013, we also analyze language labels provided by Twitter's proprietary LID algorithm and justify using FastText-LID as an alternative LID tool to overcome the challenge of missing language labels in the historical feed from Twitter (see also Hong et al. [133]).

We study the empirical dynamics of replication: The rate at which users choose to retweet instead of generating original content; and how that rate varies across languages temporally. We quantify the ratio of retweets to new messages (contagion ratio) in each language. In most common languages on Twitter, we show that this ratio reveals a growing tendency to retweet.

Finally, we present a detailed comparison with the historical data feed in Appendix 2.A. We conclude with an analytical validation of our contagion ratios (Appendix 2.B), and the impact of tweet-length on language detection (Appendix 2.C). We also provide an online appendix at: `http://compstorylab.org/storywrangler/papers/tlid/`.

2.4 Data and methods

Twitter is a well-structured streaming source of sociotechnical data, allowing for the study of dynamical linguistics and cultural phenomena [75, 323]. Of course, like many other social platforms, Twitter represents only a subsample of the publicly declared views, utterances, and interactions of millions of individuals, organizations, and automated accounts (e.g., social bots) around the world [149, 194, 205, 305]. Researchers have nevertheless shown that Twitter's collective conversation mirrors the dynamics of local and global events [216] including earthquakes [254], flu and influenza [67, 168], crowdsourcing and disaster relief [102, 231], major political affairs [271], and fame dynamics for political figures and celebrities [82]. Moreover, analyses of social media data and digital text corpora over the last decade have advanced natural language processing (NLP) research [122, 245, 246] such as language detection [23, 183, 184, 302], sentiment analysis [57, 79, 151, 161, 163], word embeddings [72, 112, 201, 223], and machine translation [11, 186, 220].

LID is often referred to as a solved problem in NLP research [113, 135, 182, 185, 191], especially for properly formatted documents, such as books, newspapers, and other long-form digital texts. Language detection for tweets, however, is a challenging task due to the nature of the platform. Every day, millions of text snippets are posted to Twitter and written in many languages along with misspellings, catchphrases, memes, hashtags, and emojis, as well as images, gifs, and videos. Encoding many cultural phenomena semantically, these features contribute to the unique aspects of language usage on Twitter that are distinct from studies of language on longer, edited corpora [196].

A key challenge of LID on Twitter data is the absence of a large, public, annotated corpus of tweets covering most languages for training and evaluation of LID algorithms. Although researchers have compiled a handful of manually labeled datasets of Twitter messages, the proposed datasets were notably small compared to the size of daily messages on Twitter and limited to a few common languages [23, 184, 302]. They showed, however, that most off-the-shelf LID methods perform relatively well when tested on annotated tweets.

As of early 2013, Twitter introduced language predictions classified by their internal algorithm in the historical data feed [248]. Since the LID algorithm used by Twitter is proprietary, we can only refer to a simple evaluation of their own model.[1] Our analysis of Twitter's language labels indicates Twitter appears to have tested several language detection methods, or perhaps different parameters, between 2013 and 2016.

Given access to additional information about the author of a tweet, the LID task would conceivably be much more accurate. For example, if the training data for prediction included any or all of the self-reported locations found in a user's 'bio', the GPS coordinates of their most recent tweet, the language they prefer to read messages in, the language associated with individuals they follow or who follow them, and their collective tweet history, we expect the predictions would improve considerably. However, for the present investigation, we assume the only available predictors are found in the message itself. Our goal is to use the state-of-the-art language detection tools to get consistent language labels for messages in our data set to enable us to investigate broader questions about linguistic dynamics and the growth of retweets

[1] https://blog.twitter.com/engineering/en_us/a/2015/evaluating-language-identification-perfo
html

on the platform over time.

2.4.1 Open-source tools for language detection

Several studies have looked closely at language identification and detection for short-text [46, 88, 110, 210, 242, 272, 288, 299], particularly on Twitter where users are limited to a few characters per tweet (140 prior to the last few months of 2017, 280 thereafter [249]). Researchers have outlined common challenges specific to this platform [17, 106].

Most studies share a strong consensus that language identification of tweets is an exceptionally difficult task for several reasons. First, language classification models are usually trained over formal and large corpora, while most messages shared on Twitter are informal and composed of 140 characters or fewer [23, 184] (see Appendix 2.C for more details). Second, the informal nature of the content is also a function of linguistic and cultural norms; some languages are used differently over social media compared to the way they are normally used in books and formal documents. Third, users are not forced to choose a single language for each message; indeed messages are often posted with words from several languages found in a single tweet. Therefore, the combination of short, informal, and multilingual posts on Twitter makes language detection much more difficult compared to LID of formal documents [232]. Finally, the lack of large collections of verified ground-truth across most languages is challenging for data scientists seeking to fine-tune language detection models using Twitter data [23, 29, 325].

Researchers have evaluated off-the-shelf LID tools on substantial subsets of Twitter data for a limited number of languages [23, 29, 184]. For example, Google's Compact

Language Detector (versions CLD-1[2] and CLD-2[3]) offer open-source implementations of the default LID tool in the Chrome browser to detect language used on web pages using a naive Bayes classifier. In 2012, Lui and Baldwin [183] proposed a model called langid that uses an n-gram-based multinomial naive Bayes classifier. They evaluated langid and showed that it outperforms Google's CLD on multiple datasets. A majority-vote ensemble of LID models is also proposed by Lui and Baldwin [184] that combines both Google's CLD and langid to improve classification accuracy for Twitter data.

Although using a majority-vote ensemble of LID models may be the best option to maximize accuracy, there are a few critical trade-offs including speed and uncertainty. The first challenge of using an ensemble is weighing the votes of different models. One can propose treating all models equally and taking the majority vote. This becomes evidently complicated in case of a tie, or when models are completely unclear on a given tweet. Treating all models equally is an arguably flawed assumption given that not all models will have the same confidence in each prediction—if any is reported. Unfortunately, most LID models either decline to report a confidence score, or lack a clear and consistent way of measuring their confidence. Finally, running multiple LID classifiers on every tweet is computationally expensive and time-consuming.

Recent advances in word embeddings powered by deep learning demonstrate some of the greatest breakthroughs across many NLP tasks including LID. Unlike previous methodologies, Devlin et al. [72] introduces a new language representation model called BERT. An additional output layer can be added to the pre-trained model to harvest the power of the distributed language representations, which enables the

[2] http://code.google.com/p/chromium-compact-language-detector/
[3] https://github.com/CLD2Owners/cld2

model to carry out various NLP tasks such as LID.

FastText [32, 143] is a recently proposed approach for text classification that uses n-gram features similar to the model described by Mikolov et al. [199]. FastText employs various tricks [32, 112, 201] in order to train a simple neural network using stochastic gradient descent and a linearly decaying learning rate for text classification. While FastText is a language model that can be used for various text mining tasks, it requires an additional step of producing vector language representations to be used for LID. To accomplish that, we use an off-the-shelf language identification tool [32] that uses the word embeddings produced by the model. The proposed tool uses a hierarchical softmax function [143, 199] to efficiently compute the probability distribution over the predefined classes (i.e., languages). For convenience, we will refer to the off-the-shelf LID tool [32] as FastText-LID throughout the paper. The authors show that FastText-LID is on par with deep learning models [62, 319] in terms of accuracy and consistency, yet orders of magnitude faster in terms of inference and training time [32]. They also show that FastText-LID outperforms previously introduced LID tools such as langid.[4]

2.4.2 Processing pipeline

While there are many tools to consider for LID, it is important for us to ensure that the language classification process stays rather consistent to investigate our key question about the growth of retweets over time. In light of the technical challenges discussed in the previous section, we have confined this work to using FastText-LID [32] due to its consistent and reliable performance in terms of inference time and accuracy.

[4] https://fasttext.cc/blog/2017/10/02/blog-post.html

To avoid biasing our language classification process, we filter out Twitter-specific content prior to passing tweets through the FastText-LID model. This is a simple strategy originally proposed by Tromp and Pechenizkiy [288] to improve language classification [24, 184]. Specifically, we remove the prefix associated with retweets ("RT"), links (e.g., "https://twitter.com"), hashtags (e.g., "#newyear"), handles (e.g., "@username"), html codes (e.g., "&gt"), emojis, and any redundant whitespaces.

Once we filter out all Twitter-specific content, we feed the remaining text through the FastText-LID neural network and select the predicted language with the highest confidence score as our ground-truth language label. If the confidence score of a given prediction is less than 25%, we label that tweet as Undefined (`und`). Similarly, if no language classification is made by the Twitter-LID model, Twitter flags the language of the message as undefined [228, 293]. We provide a list of all language labels assigned by FastText-LID compared to the ones served by Twitter-LID in Table 2.A.1.

We subsequently extract day-scale time series and Zipf distributions for unigrams, bigrams, and trigrams and make them available through an analytical instrument entitled Storywrangler. Our tool is publicly available online at: `https://storywrangling.org`. See Alshaabi et al. [5] for technical details about our project.

2.5 Results

2.5.1 Temporal and empirical statistics

We have collected a random 10% sample of all public tweets posted on the Twitter platform starting January 1, 2009. Using the steps described in Sec. 2.4.2, we have implemented a simple pipeline to preprocess messages and obtain language labels using FastText-LID [32]. Our source code along with our documentation is publicly available online on a Gitlab repository.[5] Here, we evaluate our results by comparing the language labels obtained by FastText-LID to those found in the metadata provided by Twitter's internal LID algorithm(s). Our initial analysis of the Decahose metadata indicated missing language labels until 2013, when Twitter began offering a language prediction (we offer an approach to detecting corrupted time series [84]).

We find that our classification of tweets using FastText-LID notably improves the consistency of language labels when compared to the labels served with the historical streaming feed. In Fig. 2.1A, we display a weekly rolling average of the daily number of languages detected by each classifier over time. We see that Twitter's language detection has evolved over time. The number of languages stabilized but continued to fluctuate in a manner that is not consistent, with uncommon languages having zero observations on some given days. By contrast, the FastText-LID time series of the number of languages shows some fluctuations in the earlier years—likely the result of the smaller and less diverse user base in the late 2000s—but stabilizes before Twitter introduces language labels. We note that the fluctuations in the time series during the

[5] https://gitlab.com/compstorylab/storywrangler

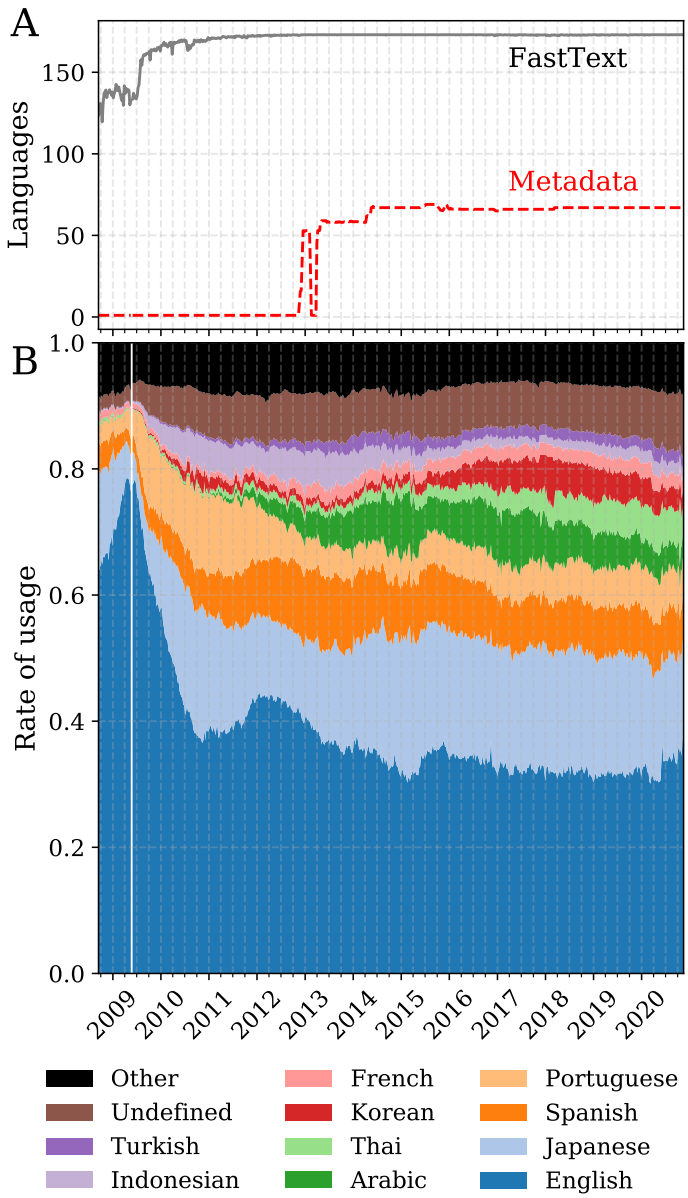

Figure 2.1: ***Language time series for the Twitter historical feed and FastText-LID classified tweets. A.*** *Number of languages reported by Twitter-LID (red) and classified by FastText-LID (black) since September 2008. Fluctuations in late 2012 and early 2013 for the Twitter language time series are indicative of inconsistent classifications.* ***B.*** *Rate of usage by language using FastText-LID maintains consistent behavior throughout throughout that period. The change in language distribution when Twitter was relatively immature can be readily seen—for instance, English accounted for an exceedingly high proportion of activity on the platform in 2009, owing to Twitter's inception in an English-speaking region.*

early years of Twitter (before 2012) and the first week of 2017 are primarily caused by unexpected service outages which resulted in missing data.

FastText-LID classifies up to 173 languages, some of which are rare, thus the occasional dropout of a language seen in this time series is expected. On the other hand, Twitter-LID captures up to 73 languages, some of which are experimental and no longer available in recent years. Nonetheless, Fig. 2.1B shows that the overall rate of usage by language is not impaired by the missing data, and maintained consistent behavior throughout the last decade.

We compute annual confusion matrices to examine the language labels classified by FastText-LID compared to those found in the historical data feed. Upon inspection of the computed confusion matrices, we find disagreement during the first few years of Twitter's introduction of the LID feature to the platform. As anticipated, the predicted language for the majority of tweets harmonizes across both classifiers for recent years (see Fig. 2.A.1). We notice some disagreement between the two classifiers on expected edge-cases such as Italian, Spanish, and Portuguese where the lexical similarity among these languages is very high [37, 142, 244, 255]. Overall, our examination of average language usage over time demonstrates that FastText-LID is on par with Twitter's estimation. We show the corresponding Zipf distribution of language usage for each classifier, and highlight the normalized ratio difference between them for the most used languages on the platform in Figs. 2.A.2–2.A.3. We point the reader's attention to Appendix 2.A for further details of our comparison.

Furthermore, we display a heatmap of the number of messages for each language as classified by FastText-LID in our data set (see Fig. 2.2). We have over 118 billion messages between 2009-01-01 and 2019-12-31 spanning 173 languages. English is the

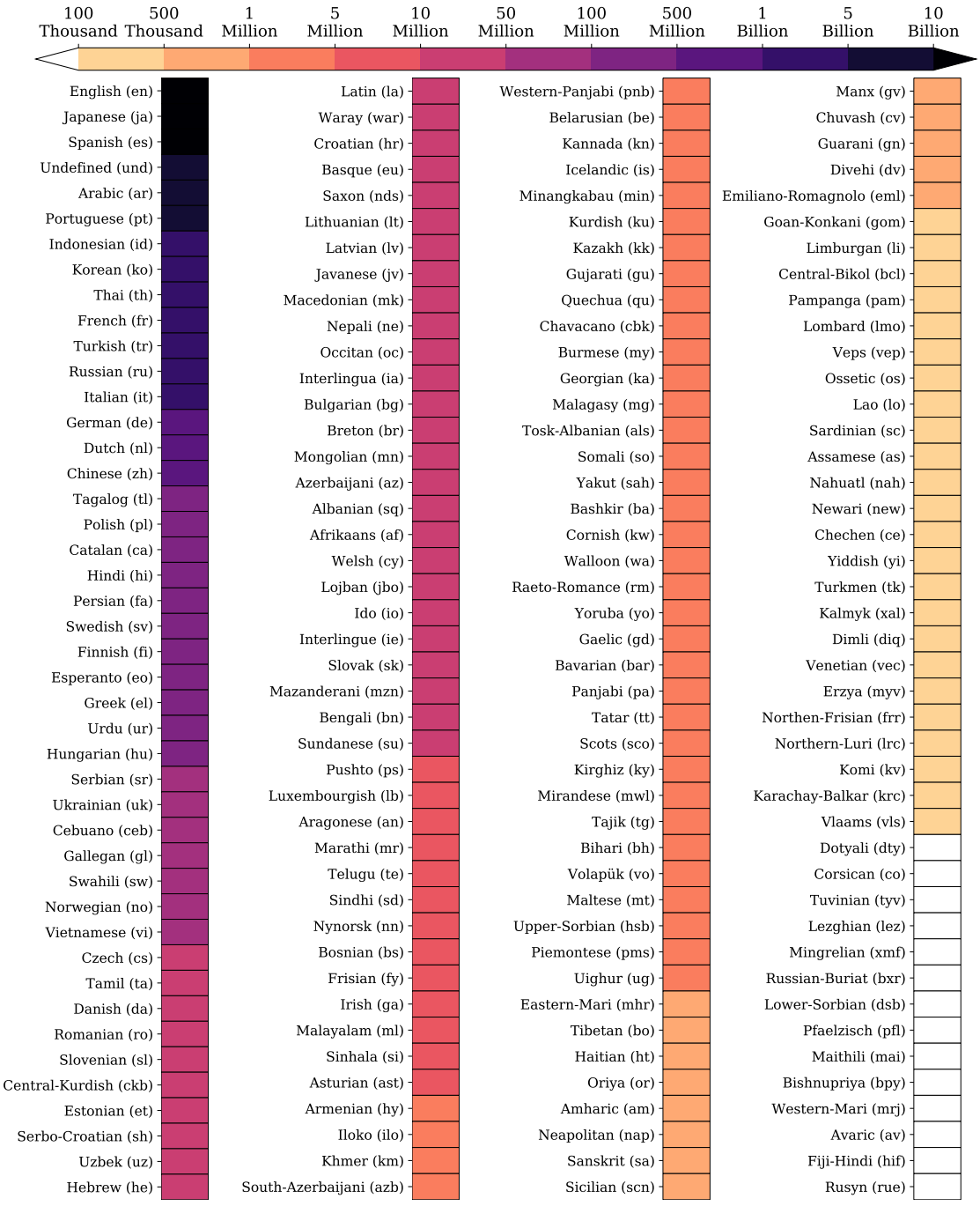

Figure 2.2: ***Overall dataset statistics.*** *Number of messages captured in our dataset as classified by the FastText-LID algorithm between 2009-01-01 and 2019-12-31, which sums up to a approximately 118 billion messages throughout that period (languages are sorted by popularity). This collection represents roughly 10% of all messages ever posted.*

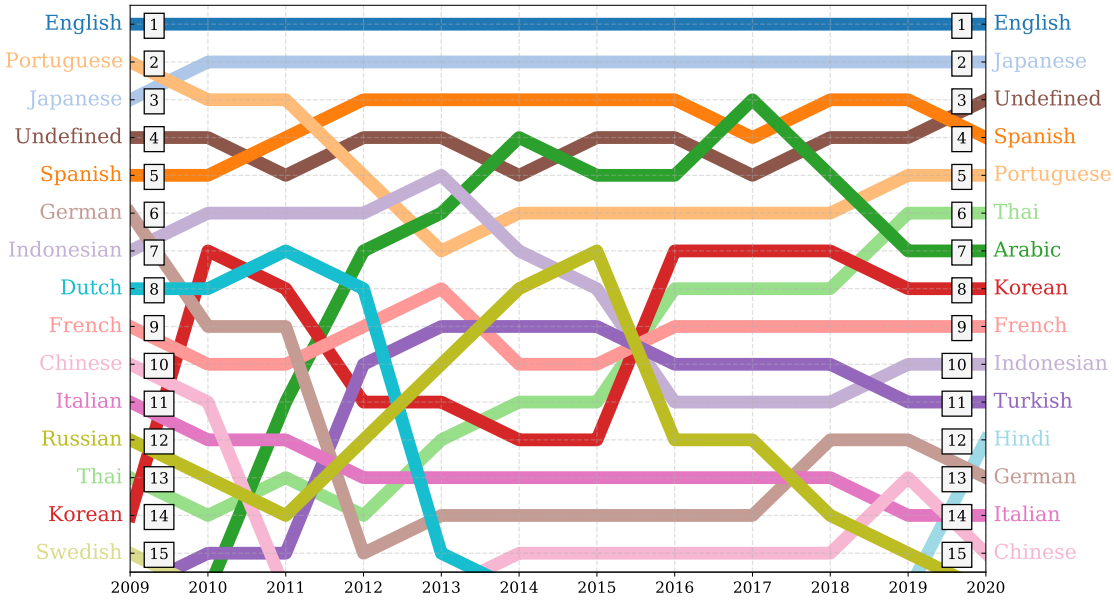

Figure 2.3: ***Annual average rank of the most used languages on Twitter between 2009 and 2020.*** *English and Japanese show the most consistent rank time series. Spanish, and Portuguese are also relatively stable over time. Undefined—which covers a wide variety of content such as emojis, links, pictures, and other media—also has a consistent rank time series. The rise of languages on the platform correlates strongly with international events including Arab Spring and K-pop, as evident in both the Arabic and Korean time series, respectively. Russian, German, Indonesian, and Dutch moved down in rank. This shift is not necessarily due to a dramatic drop in the rate of usage of these languages, but is likely an artifact of increasing growth of other languages on Twitter such as Thai, Turkish, Arabic, Korean, etc.*

most used language on the platform with a little under 42 billion messages throughout the last decade. Although the number of Japanese speakers is much smaller than the number of English speakers around the globe, Japanese has approximately 21 billion messages. Spanish—the third most prominent language on Twitter—is shy of 11 billion messages. Arabic and Portuguese rank next with about 7 billion messages for each. We note that the top 10 languages comprise 85% of the daily volume of messages posted on the platform.

In Fig. 2.3, we show the flow of annual rank dynamics of the 15 most used languages on Twitter between 2009 and 2020. For ease of description, we will refer to Undefined as a language class. The top 5 most common languages on Twitter (English, Japanese, Spanish, Undefined, and Portuguese) are consistent, indicating a steady rate of usage of these languages on the platform. The language rankings correspond with worldwide events such as the Arab Spring [65, 74, 136, 307], K-pop, and political events [82]. "Undefined" is especially interesting as it covers a wide range of content such as emojis, memes, and other media shared on Twitter but can't necessarily be associated with a given language. Russian, however, starts to grow on the platform after 2011 until it peaks with a rank of 7 in 2015, then drops down to rank 15 as of the end of 2019. Other languages such as German, Indonesian, and Dutch show a similar trend down in ranking. This shift is not necessarily caused by a drop in the rate of usage of these languages, but it is rather an artifact prompted by the growth of other languages on Twitter.

2.5.2 Separating organic and retweeted messages

We take a closer look at the flow of information among different languages on the platform, specifically the use of the "retweet" feature as a way of spreading information. Encoding a behavioral feature initially invented by users, Twitter formalized the retweet feature in November 2009 [275]. Changes in platform design and the increasing popularity of mobile apps promoted the RT as a mechanism for spreading. In April 2015, Twitter introduced the ability to comment on a retweet message or "Quote Tweet"(QT) [264] a message, distinct from a message reply [274].

To quantify the rate of usage of each language with respect to these different

means by which people communicate on the platform, we categorize messages on Twitter into two types: "Organic Tweets" (OT), and "Retweets" (RT). The former category (OT) encompasses original messages that are explicitly authored by users, while the latter category (RT) captures messages that are shared (i.e. amplified) by users. We break each quote tweet into two separate messages: a comment and a retweet. We exclude retweets while including all added text (comments) found in quote tweets for the OT category.

For each day t and for each language ℓ, we calculate the raw frequency (count) of organic messages $f_{\ell,t}^{(\mathrm{OT})}$, and retweets $f_{\ell,t}^{(\mathrm{RT})}$. We further determine the frequency of all tweets (AT) such that: $f_{\ell,t}^{(\mathrm{AT})} = f_{\ell,t}^{(\mathrm{OT})} + f_{\ell,t}^{(\mathrm{RT})}$. The corresponding rate of usages (normalized frequencies) for these two categories are then:

$$p_{t,\ell}^{(\mathrm{OT})} = \frac{f_{t,\ell}^{(\mathrm{OT})}}{f_{t,\ell}^{(\mathrm{AT})}}, \text{ and } p_{t,\ell}^{(\mathrm{RT})} = \frac{f_{t,\ell}^{(\mathrm{RT})}}{f_{t,\ell}^{(\mathrm{AT})}}. \tag{2.1}$$

2.5.3 Measuring sociolinguistic wildfire through the growth of retweets

To further investigate the growth of retweets, we use the ratio of retweeted messages to organic messages as an intuitive and interpretable analytical measure to track this social amplification phenomenon. We define the 'contagion ratio' as:

$$R_{\ell,t} = f_{\ell,t}^{(\mathrm{RT})} \Big/ f_{\ell,t}^{(\mathrm{OT})}. \tag{2.2}$$

In 2018, the contagion ratio exceeded 1, indicating a higher number of retweeted

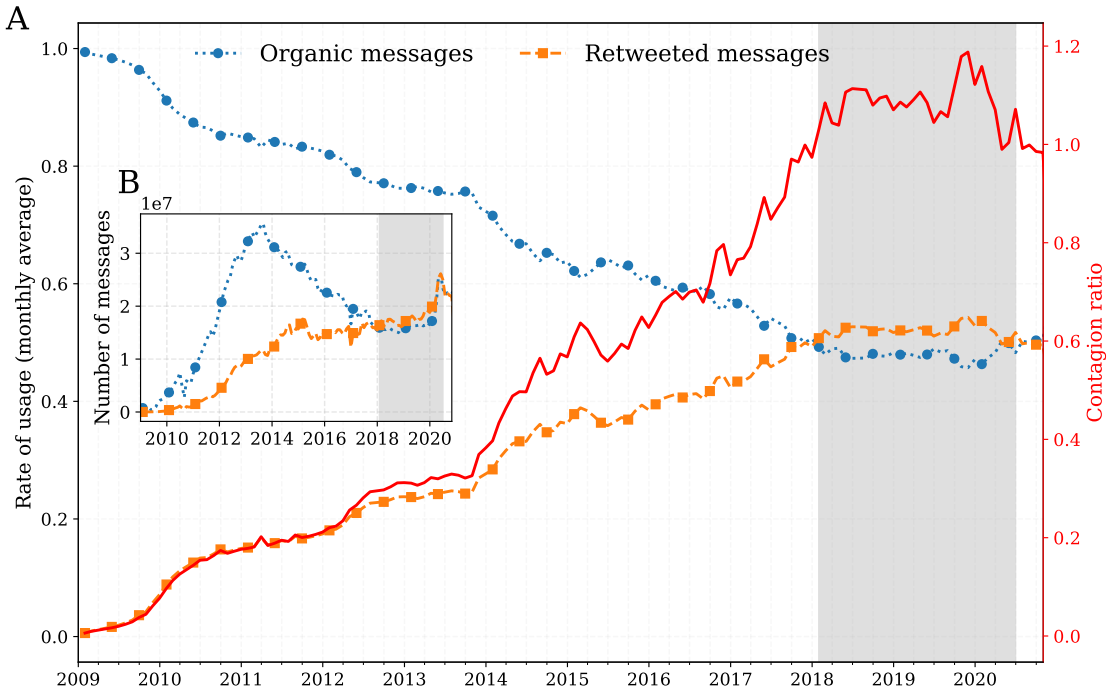

Figure 2.4: ***Timeseries for organic messages, retweeted messages, and average contagion ratio for all languages.*** ***A.*** *Monthly average rate of usage of organic messages ($p^{(\mathrm{OT})}_{t,\ell}$, blue), and retweeted messages ($p^{(\mathrm{RT})}_{t,\ell}$, orange). The solid red line highlights the steady rise of the contagion ratio $R_{\ell,t}$.* ***B.*** *Frequency of organic messages ($f^{(\mathrm{OT})}_{\ell,t}$, blue), compared to retweeted messages ($f^{(\mathrm{RT})}_{\ell,t}$, orange). The areas shaded in light grey starting in early 2018 highlights an interesting shift on the platform where the number of retweeted messages has exceeded the number of organic messages. An interactive version of the figure for all languages is available in an online appendix:* `http://compstorylab.org/storywrangler/papers/tlid/files/ratio_timeseries.html`.

messages than organic messages (Fig. 2.4). The overall count for organic messages peaked in the last quarter of 2013, after which it declined slowly as the number of retweeted messages climbed to approximately 1.2 retweeted messages for every organic message at the end of 2019. Thereafter, the contagion ratio declined through 2020 with the exception of a surge of retweets in the summer amid the nationwide protests sparked by the murder of George Floyd.[6]

In 2020, Twitter's developers redesigned their retweet mechanism, purposefully prompting users to write their own commentary using the Quote Tweet [99], along with several new policies to counter synthetic and manipulated media [98, 250, 251]. While the long upward trend of the contagion ratio is in part due to increasingly active social amplification by users, the recent trend demonstrates how social amplification on Twitter is highly susceptible to systematic changes in the platform design. Twitter has also introduced several features throughout the last decade, such as tweet ranking, and extended tweet length that may intrinsically influence how users receive and share information in their social networks.[7] We investigate the robustness of our findings regarding contagion ratios in light of some of these changes in Appendix 2.B and Appendix 2.C. Future work will shed light on various aspects of social amplification on Twitter with respect to the evolution of the platform design, and behavioral drivers for the use of retweets in each language across communities.

Finally, we show weekly aggregation of the rate of usage of the top 30 ranked languages of 2019 in Fig. 2.5. The time series demonstrate a recent sociolinguistic shift: Several languages including English, Spanish, Thai, Korean, and French have transitioned to having a higher rate of retweeted messages than organic messages.

[6] https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html
[7] https://help.twitter.com/en/using-twitter/twitter-conversations

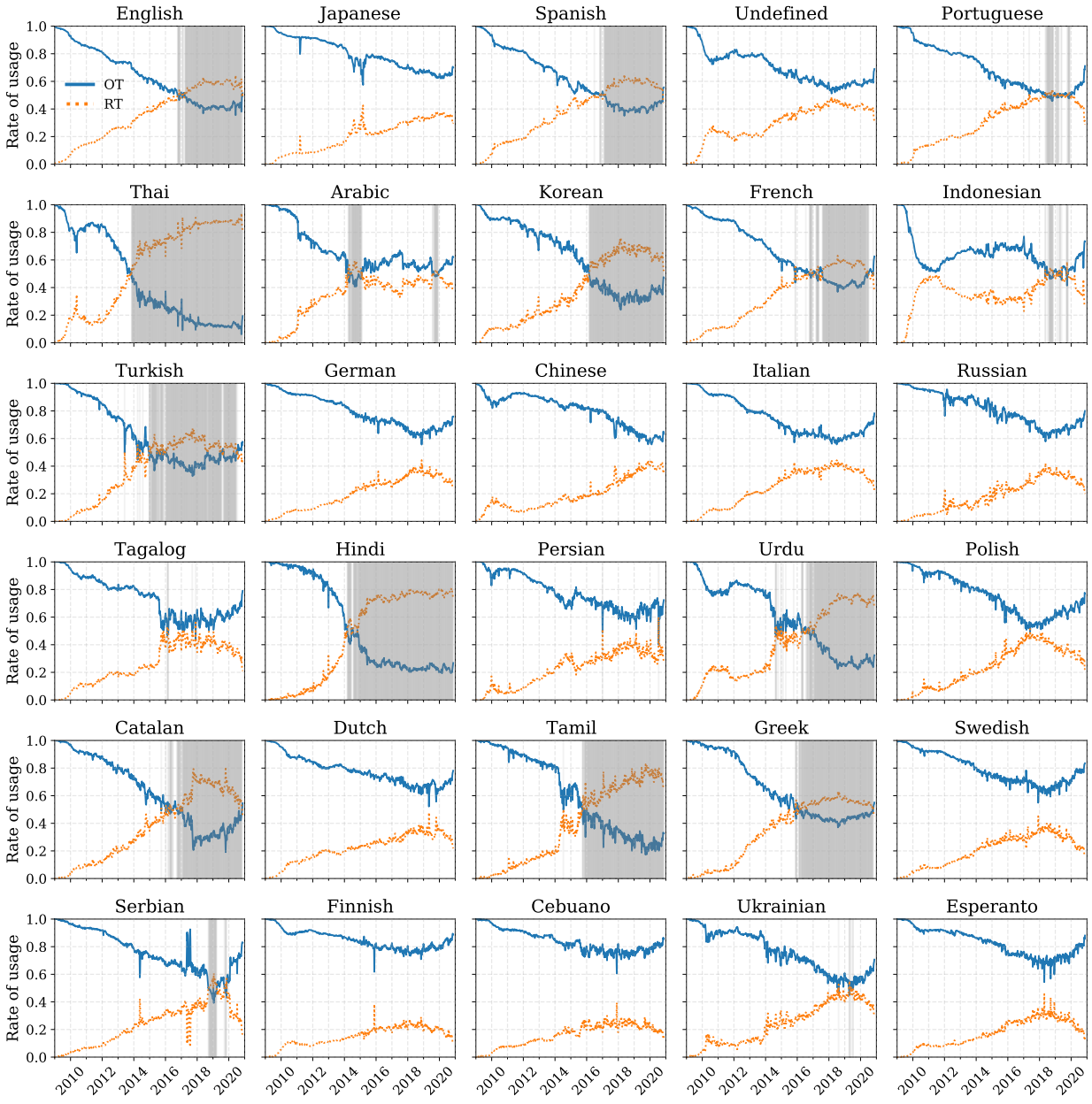

Figure 2.5: ***Weekly rate of usage of the top 30 languages (sorted by popularity).*** *For each language, we show a weekly average rate of usage for organic messages ($p_{t,\ell}^{(\mathrm{OT})}$, blue) compared to retweeted messages ($p_{t,\ell}^{(\mathrm{RT})}$, orange). The areas highlighted in light shades of gray represent weeks where the rate of retweeted messages is higher than the rate of organic messages. An interactive version featuring all languages is available in an online appendix:* `http://compstorylab.org/storywrangler/papers/tlid/files/retweets_timeseries.html`.

Thai appears to be the first language to have made this transition in late 2013. In Fig. 2.6, we show a heatmap of the average contagion ratio for the top 30 most used languages on Twitter per year. With the exception of Indonesian, which showed a small bump between 2010 and 2013, most other languages began adopting a higher ratio of retweeted content in 2014. Thai has the highest number of retweeted messages, with an average of 7 retweeted messages for every organic message. Other languages, for example, Hindi, Korean, Urdu, Catalan, and Tamil average between 2 to 4 retweeted messages for every organic message. On the other hand, Japanese—the second most used language on the platform—does not exhibit this trend. Similarly, German, Italian, and Russian maintain higher rates of organic tweets. The trend of increasing preference for retweeted messages, though not universal, is evident among most languages on Twitter. We highlight the top 10 languages with the highest and lowest average contagion ratio per year in Table 2.B.1 and Table 2.B.2, respectively.

2.6 Discussion

Understanding how stories spread through and persist within populations has always been central to understanding social phenomena. In a time when information can flow instantly and freely online, the study of social contagion has only become more important.

In the sphere of Twitter, the practice of retweeting is complicated from a social and psychological point of view. There is a diverse set of reasons for participants to retweet. For example, scientists and academics can use this elementary feature to share their findings and discoveries with their colleagues. Celebrities and political

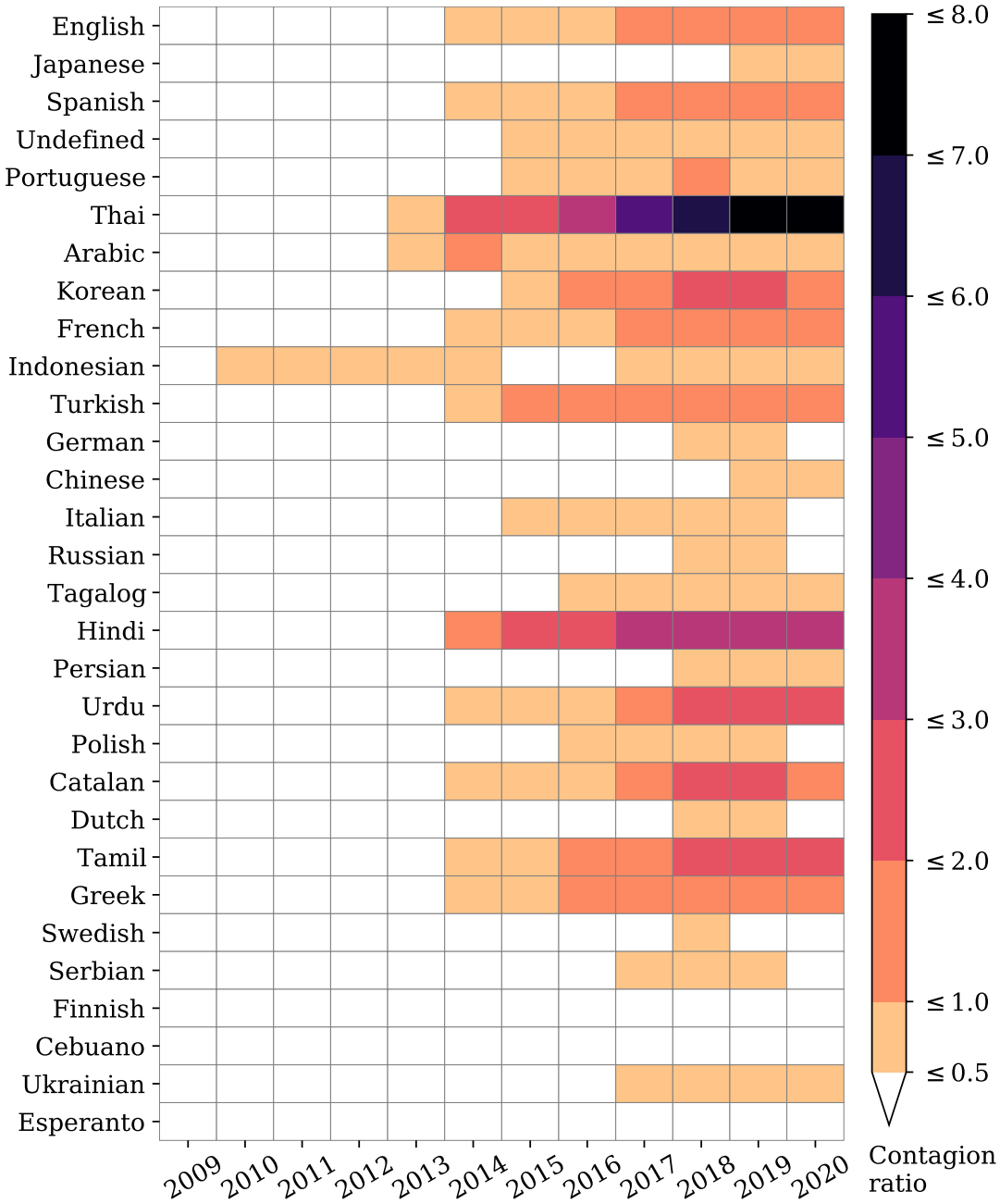

Figure 2.6: ***Timelapse of contagion ratios.*** *The average ratio is plotted against year for the top 30 ranked languages of 2019. Colored cells indicate a ratio higher than 0.5 whereas ratios below 0.5 are colored in white. Table 2.B.1 shows the top 10 languages with the highest average contagion ratio per year, while Table 2.B.2 shows the bottom 10 languages with the lowest average contagion ratio per year.*

actors can benefit from other people retweeting their stories for self-promotion. Attackers can also take advantage of this natural feature of social contagion to pursue malicious intents, deploy social bots, and spread fake news.

In this study, we have analyzed over a hundred billion messages posted on Twitter throughout the last decade. We presented an alternative approach for obtaining language labels using FastText-LID in order to overcome the challenge of missing labels in the Decahose dataset, obtaining consistent language labels for 100+ languages. We acknowledge that shortcomings of language detection for short and informal text (e.g., tweets) are well known in the NLP literature. Using FastText-LID is not necessarily the best approach for language identification. Our analysis may be subject to implicit measurement biases and errors introduced by word embeddings used to train the language detection tool using FastText [32]. We emphasize that we have not intended to reinvent or improve FastText-LID in this work; we have used FastText-LID only as a (well-established and tested) tool to enable the study of social contagion dynamics on Twitter. Nevertheless, we have presented some further analysis of FastText-LID compared to Twitter-LID in Appendix 2.A. Future work will undoubtedly continue to improve language detection for short text, particularly for social media platforms.

Our results comparing language usage over time suggest a systematic shift on Twitter. We found a recent tendency among most languages to increasingly retweet (spread information) rather than generate new content. Understanding the general rise of retweeted messages requires further investigation. Possible partial causes might lie in changes in the design of the platform, increases in bot activity, a fundamental shift in human information processing as social media becomes more familiar to users, and changes in the demographics of users (e.g., younger users joining the platform).

The metrics we have used to compute our contagion ratios are simple but rather limited. We primarily focused on tracking the rate of organic tweets and retweets to quantify social amplification of messages on the platform. While our approach of measuring the statistical properties of contagion ratios is important, it falls short of capturing how retweets propagate throughout the social networks of users. Future work may deploy a network-based approach to investigate the flow of retweets among users and followers. Whether or not the information is differentially propagated across languages, social groups, economic strata, or geographical regions is an important question for future research, and beyond the scope of our present work.

Geolocation information for Twitter is also limited, and here we have only examined contagion ratios at the language level. Language, transcending borders as it does, can nevertheless be used differently across communities. For instance, characterizing the temporal dynamics of contagion ratios for English, which is used all around the globe, is very different from doing so for Thai—a language that is used within a geographically well-defined population. Different social and geographical communities have cultures of communication which will need to be explored in future work.

In particular, it is important to study the relationship between social contagion dynamics, geographical region, and language. It might be the case that contagion dynamics are more homogeneous across geographic regions even when each geographical region displays high language diversity, or *vice versa*. However, in order to conduct this line of research, it is necessary to have accurate geotagging of tweets, which is currently not the case except for a very small subsample [294]. Future research could focus on implementing accurate geotagging algorithms that assign tweets a probabilis-

tic geographical location based on their text and user metadata, while fully respecting privacy through judicious use of masking algorithms.

2.7 Acknowledgments

The authors are grateful for the computing resources provided by the Vermont Advanced Computing Core and financial support from the Massachusetts Mutual Life Insurance Company and Google Open Source under the Open-Source Complex Ecosystems And Networks (OCEAN) project. Computations were performed on the Vermont Advanced Computing Core supported in part by NSF award No. OAC-1827314. We thank Colin Van Oort and Anne Marie Stupinski for their comments on the manuscript.

Appendix

2.A Comparison with the historical feed

We have collected all language labels served in the historical data feed, along with the predicted language label classified by FastText-LID for every individual tweet in our dataset. We provide a list of all language labels assigned by FastText-LID compared to the ones served by Twitter-LID in Table 2.A.1. To evaluate the agreement between the two classifiers, we computed annual confusion matrices starting from 2013 to the end of 2019. In Fig. 2.A.1, we show confusion matrices for the 15 most dominate languages on Twitter for all tweets authored in 2013 (Fig. 2.A.1A) and 2019 (Fig. 2.A.1B).

We observe some disagreement between the two classifiers during the early years of Twitter's introduction of the LID feature to the platform. In Fig. 2.A.2, we show the normalized ratio difference δD_ℓ (i.e., divergence) between the two classifiers for all messages between 2014 and 2019. Divergence is calculated as:

$$\delta D_\ell = \left| \frac{\mathcal{C}_\ell^{\mathrm{F}} - \mathcal{C}_\ell^{\mathrm{T}}}{\mathcal{C}_\ell^{\mathrm{F}} + \mathcal{C}_\ell^{\mathrm{T}}} \right|, \tag{2.3}$$

where $\mathcal{C}_\ell^{\mathrm{F}}$ is the number of messages captured by FastText-LID for language ℓ, and

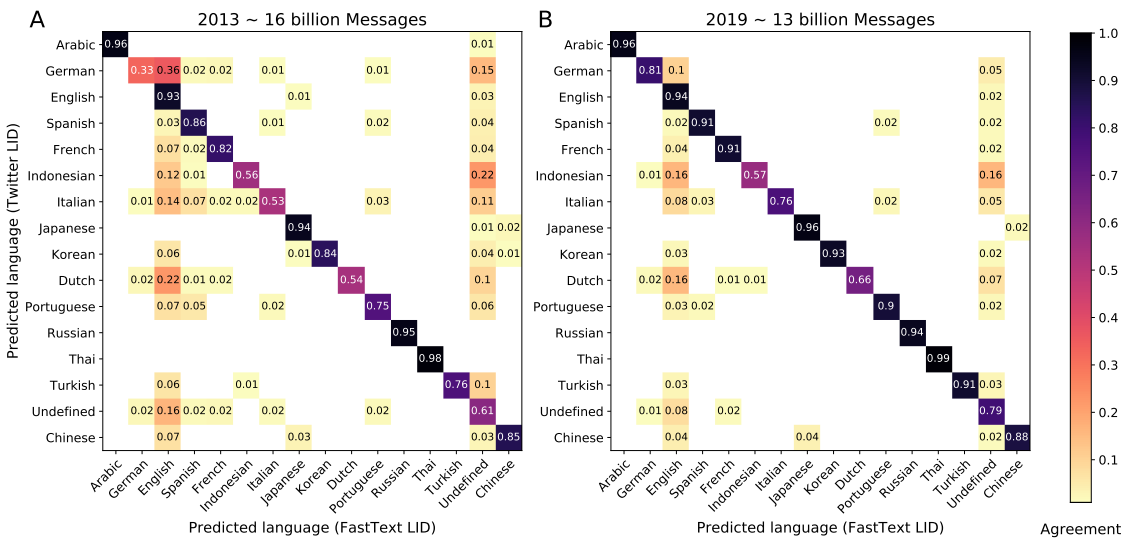

Figure 2.A.1: ***Language identification confusion matrices.*** *We show a subset of the full confusion matrix for top-15 languages on Twitter.* ***A.*** *Confusion matrix for tweets authored in 2013. The matrix indicates substantial disagreement between the two classifiers during 2013, the first year of Twitter's efforts to provide language labels.* ***B.*** *For the year 2019, both classifiers agree on the majority of tweets as indicated by the dark diagonal line in the matrix. Minor disagreement between the two classifiers is evident for particular languages, including German, Italian, and Undefined, and there is major disagreement for Indonesian and Dutch. Cells with values below (.01) are colored in white to indicate very minor disagreement between the two classifiers.*

$\mathcal{C}_\ell^{\mathrm{T}}$ is the number of messages captured by Twitter-LID for language ℓ.

We show Zipf distributions of all languages captured by FastText-LID and Twitter-LID in Fig. 2.A.2A and Fig. 2.A.2B, respectively. FastText-LID recorded a total of 173 unique languages, whereas Twitter-LID captured a total of 73 unique languages throughout that period. Some of the languages reported by Twitter were experimental and no longer available in recent years. In Fig. 2.A.2C, we display the joint distribution of all languages captured by both classifiers. Languages found left of vertical dashed gray line are more prominent using the FastText-LID model (e.g., Chinese (zh), Central-Kurdish (ckb), Uighur (ug), Sindhi (sd)). Languages right of

the line are identified more frequently by the Twitter-LID model (e.g., Estonian (et), Haitian (ht)). Languages found within the light-blue area are only detectable by one classifier but not the other. We note that 'Unknown' is an artificial label that we added to flag messages with missing language labels in the metadata of our dataset. We list divergence values δD_ℓ for all languages identified in our dataset in Fig. 2.A.3.

Table 2.A.1: Language codes for both FastText-LID and Twitter-LID tools

Language	FT	TW	Language	FT	TW	Language	FT	TW	Language	FT	TW
Afrikaans	af	-	**Czech**	cs	cs	**Occitan**	oc	-	**Tagalog**	tl	tl
Albanian	sq	-	**Danish**	da	da	**Oriya**	or	or	**Tajik**	tg	-
Amharic	am	am	**Dimli**	diq	-	**Ossetic**	os	-	**Tamil**	ta	ta
Arabic	ar	ar	**Divehi**	dv	dv	**Pampanga**	pam	-	**Tatar**	tt	-
Aragonese	an	-	**Dotyali**	dty	-	**Panjabi**	pa	pa	**Telugu**	te	te
Armenian	hy	hy	**Dutch**	nl	nl	**Persian**	fa	fa	**Thai**	th	th
Assamese	as	-	**Eastern-Mari**	mhr	-	**Pfaelzisch**	pfl	-	**Tibetan**	bo	bo
Asturian	ast	-	**Egyptian-Arabic**	arz	-	**Piemontese**	pms	-	**Tosk-Albanian**	als	-
Avaric	av	-	**Emiliano**	eml	-	**Polish**	pl	pl	**Turkish**	tr	tr
Azerbaijani	az	-	**English**	en	en	**Portuguese**	pt	pt	**Turkmen**	tk	-
Bashkir	ba	-	**Lojban**	jbo	-	**Pushto**	ps	ps	**Tuvinian**	tyv	-
Basque	eu	eu	**Lombard**	lmo	-	**Quechua**	qu	-	**Uighur**	ug	ug
Bavarian	bar	-	**Lower-Sorbian**	dsb	-	**Raeto-Romance**	rm	-	**Ukrainian**	uk	uk
Belarusian	be	-	**Luxembourgish**	lb	-	**Romanian**	ro	ro	**Undefined**	und	und
Bengali	bn	bn	**Macedonian**	mk	-	**Russian-Buriat**	bxr	-	**Upper-Sorbian**	hsb	-
Bihari	bh	-	**Maithili**	mai	-	**Russian**	ru	ru	**Urdu**	ur	ur
Bishnupriya	bpy	-	**Malagasy**	mg	-	**Rusyn**	rue	-	**Uzbek**	uz	-
Bosnian	bs	bs	**Malay**	ms	msa	**Sanskrit**	sa	-	**Venetian**	vec	-
Breton	br	-	**Malayalam**	ml	ml	**Sardinian**	sc	-	**Veps**	vep	-
Bulgarian	bg	bg	**Maltese**	mt	-	**Saxon**	nds	-	**Vietnamese**	vi	vi
Burmese	my	my	**Manx**	gv	-	**Scots**	sco	-	**Vlaams**	vls	-
Catalan	ca	ca	**Marathi**	mr	mr	**Serbian**	sr	sr	**Volapuk**	vo	-
Cebuano	ceb	-	**Mazanderani**	mzn	-	**Serbo-Croatian**	sh	-	**Walloon**	wa	-
Central-Bikol	bcl	-	**Minangkabau**	min	-	**Shona**	-	sn	**Waray**	war	-
Central-Kurdish	ckb	ckb	**Mingrelian**	xmf	-	**Sicilian**	scn	-	**Welsh**	cy	cy
Chavacano	cbk	-	**Mirandese**	mwl	-	**Sindhi**	sd	sd	**Western-Mari**	mrj	-
Chechen	ce	-	**Mongolian**	mn	-	**Sinhala**	si	si	**Western-Panjabi**	pnb	-
Cherokee	-	chr	**Nahuatl**	nah	-	**Slovak**	sk	-	**Wu-Chinese**	wuu	-
Chinese-Simplified	-	zh-cn	**Neapolitan**	nap	-	**Slovenian**	sl	sl	**Yakut**	sah	-
Chinese-Traditional	-	zh-tw	**Nepali**	ne	ne	**Somali**	so	-	**Yiddish**	yi	-
Chinese	zh	zh	**Newari**	new	-	**South-Azerbaijani**	azb	-	**Yoruba**	yo	-
Chuvash	cv	-	**Northen-Frisian**	frr	-	**Spanish**	es	es	**Yue-Chinese**	yue	-
Cornish	kw	-	**Northern-Luri**	lrc	-	**Sundanese**	su	-			
Corsican	co	-	**Norwegian**	no	no	**Swahili**	sw	-			
Croatian	hr	-	**Nynorsk**	nn	-	**Swedish**	sv	sv			

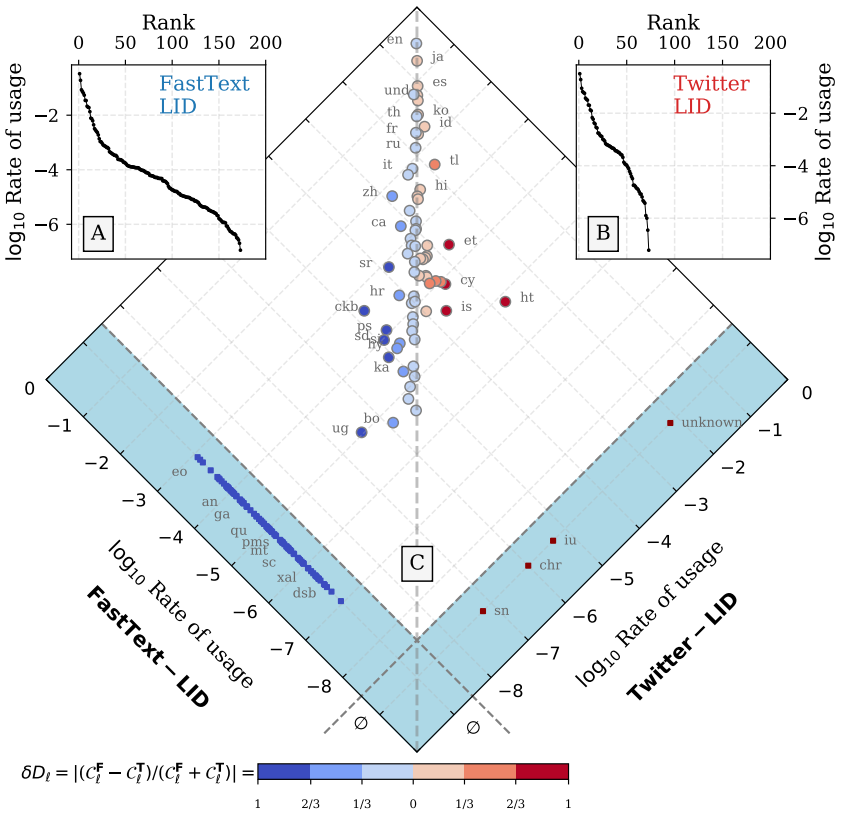

Figure 2.A.2: ***Language Zipf distributions. A.*** *Zipf distribution [321] of all languages captured by FastText-LID model.* ***B.*** *Zipf distribution for languages captured by Twitter-LID algorithm(s). The vertical axis in both panels reports rate of usage of all messages $p_{t,\ell}$ between 2014 and 2019, while the horizontal axis shows the corresponding rank of each language. FastText-LID recorded a total of 173 unique languages throughout that period. On the other hand, Twittert-LID captured a total of 73 unique languages throughout that same period, some of which were experimental and no longer available in recent years.* ***C.*** *Joint distribution of all recorded languages. Languages located near the vertical dashed gray line signify agreement between FastText-LID and Twitter-LID, specifically that they captured a similar number of messages between 2014 and end of 2019. Languages found left of this line are more prominent using the FastText-LID model, whereas languages right of the line are identified more frequently by Twitter-LID model. Languages found within the light-blue area are only detectable by one classifier but not the other where FastText-LID is colored in blue and Twitter is colored in red. The color of the points highlights the normalized ratio difference δD_ℓ (i.e., divergence) between the two classifiers, where C_ℓ^{F} is the number of messages captured by FastText-LID for language ℓ, and C_ℓ^{T} is the number of messages captured by Twitter-LID for language ℓ. Hence, points with darker colors indicate greater divergence between the two classifiers. A lookup table for language labels can be found in the Table 2.A.1, and an online appendix of all languages is also available here:* `http://compstorylab.org/storywrangler/papers/tlid/files/fasttext_twitter_timeseries.html`.

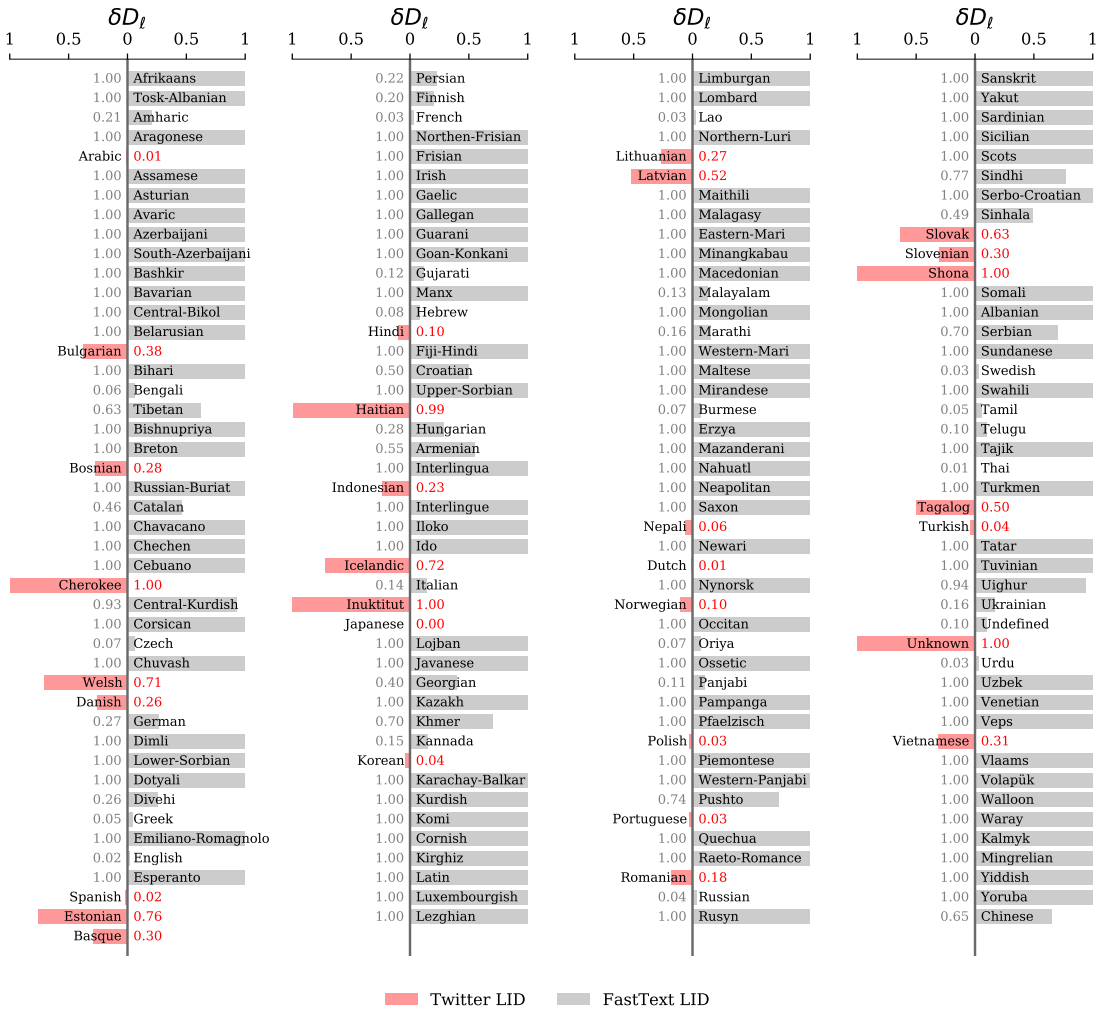

Figure 2.A.3: ***Language identification divergence.*** *A normalized ratio difference value δD_ℓ (i.e., divergence) closer to zero implies strong agreement, whereby both classifiers captured approximately the same number of messages over the last decade. Grey bars indicate higher rate of messages captured by FastText-LID, whereas red bars highlight higher rate of messages captured by Twitter-LID.*

2.B Analytical validation of contagion ratios

To investigate our margin of error for estimating contagion ratios, we find the subset of messages that both classifiers have agreed on their language labels using the annual confusion matrices we discussed in Appendix 2.A. We compute an annual average of the contagion ratios for this subset of messages. We highlight the top 10 languages with the highest and lowest average contagion ratio per year in Table 2.B.1 and Table 2.B.2, respectively. We then compare the new set of annual contagion ratios with the original ones discussed in Sec. 2.5.3. In particular, we compute the absolute difference

$$\delta = |\mathrm{R} - \mathrm{R}_\alpha|,$$

where R indicates the contagion ratios of all messages, and R_α indicates the contagion ratios of the subset of messages that both FastText-LID and Twitter-LID models have unanimously agreed on their language labels.

In Table 2.B.3, we show the top 10 languages with the highest average values of δ's. Languages are sorted by the values of δ's in 2019. Higher values of δ's indicate high uncertainty due to high disagreement on the language of the written messages between FastText-LID and Twitter-LID. Lower values of δ's, on the other hand, highlight better agreement between the two classifiers, and thus better confidence in our estimation of the contagion ratios. We show the bottom 10 languages with the lowest average values of δ's in Table 2.B.4.

Table 2.B.1: Top 10 languages with the highest annual average contagion ratio (sorted by 2019).

	2009	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Greek	0.01	0.05	0.07	0.20	0.42	0.65	0.83	1.11	1.29	1.42	1.27
French	0.02	0.10	0.13	0.22	0.34	0.56	0.76	0.94	1.09	1.40	1.37
English	0.03	0.14	0.20	0.31	0.37	0.56	0.71	0.91	1.15	1.44	1.44
Spanish	0.03	0.16	0.21	0.31	0.42	0.62	0.82	0.94	1.24	1.54	1.52
Korean	0.05	0.11	0.14	0.26	0.30	0.43	0.66	1.28	1.74	2.22	2.07
Catalan	0.01	0.08	0.12	0.21	0.30	0.52	0.74	0.98	1.80	2.44	2.10
Urdu	0.03	0.25	0.25	0.19	0.26	0.64	0.82	0.95	1.51	2.67	2.90
Tamil	0.01	0.04	0.10	0.16	0.22	0.54	0.82	1.30	1.84	2.40	2.96
Hindi	0.01	0.03	0.06	0.15	0.38	1.14	2.26	2.81	3.09	3.58	3.29
Thai	0.07	0.24	0.18	0.32	0.79	2.01	2.54	3.35	5.31	6.52	7.29

Table 2.B.2: Bottom 10 languages with the lowest annual average contagion ratio (sorted by 2019).

	2009	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Finnish	0.02	0.11	0.10	0.11	0.14	0.18	0.23	0.26	0.29	0.31	0.26
Cebuano	0.01	0.07	0.09	0.13	0.14	0.22	0.24	0.29	0.32	0.33	0.30
Esperanto	0.01	0.08	0.09	0.11	0.13	0.18	0.24	0.34	0.41	0.47	0.38
Swedish	0.02	0.07	0.09	0.14	0.20	0.31	0.37	0.41	0.47	0.55	0.45
Russian	0.01	0.04	0.07	0.13	0.13	0.19	0.29	0.31	0.42	0.57	0.50
Dutch	0.02	0.11	0.16	0.23	0.23	0.28	0.32	0.36	0.42	0.52	0.51
German	0.02	0.07	0.09	0.13	0.17	0.26	0.34	0.38	0.42	0.58	0.52
Japanese	0.02	0.08	0.10	0.11	0.16	0.31	0.35	0.31	0.40	0.47	0.53
Polish	0.01	0.06	0.08	0.13	0.22	0.28	0.42	0.60	0.84	0.74	0.57
Persian	0.03	0.07	0.07	0.14	0.22	0.40	0.35	0.41	0.50	0.64	0.57

Table 2.B.3: Top 10 languages with the highest average margin of error for estimating contagion ratios as a function of the agreement between FastText-LID and Twitter-LID (sorted by 2019).

Language	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Undefined	±0.14	±0.14	±0.16	±0.19	±0.17	±0.15
Dutch	±0.11	±0.10	±0.11	±0.12	±0.15	±0.17
Swedish	±0.14	±0.16	±0.18	±0.19	±0.21	±0.20
Serbian	±0.26	±0.27	±0.32	±0.33	±0.35	±0.25
Cebuano	±0.22	±0.24	±0.29	±0.32	±0.33	±0.30
Esperanto	±0.18	±0.24	±0.34	±0.41	±0.47	±0.38
Indonesian	±0.21	±0.18	±0.18	±0.24	±0.39	±0.40
Tagalog	±0.22	±0.34	±0.49	±0.51	±0.48	±0.44
Hindi	±0.08	±0.41	±0.97	±0.76	±0.73	±0.71
Catalan	±0.52	±0.74	±0.98	±1.80	±1.08	±0.75

In Fig. 2.B.1, we display a heatmap of δ's for the top 30 ranked languages. We note low values of δ's for the top 10 languages on the platform. In other words, our contagion ratios for the subset of messages that both classifiers have unanimously predicted their language labels are roughly equivalent to our estimations in Table 2.B.1. By contrast, we note high disagreement on Catalan messages. The two classifiers start off with unusual disagreement in 2014 ($\delta = .52$). The disagreement between the two models continues to grow leading to a remarkably high value of $\delta = 1.80$ in 2017. Thereafter, we observe a trend down in our estimations, indicating that FastText-LID and Twitter-LID have slowly started to harmonize their language predictions for Catalan messages through the past few years. We also note similar trends for Hindi and Tagalog messages.

Table 2.B.4: Bottom 10 languages with the lowest average margin of error for estimating contagion ratios as a function of the agreement between FastText-LID and Twitter-LID (sorted by 2019).

Language	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Tamil	±0.03	±0.01	±0.01	±0.01	±0.01	±0.01
Greek	±0.13	±0.07	±0.01	±0.01	±0.01	±0.01
Japanese	±0.01	±0.01	±0.01	±0.01	±0.02	±0.02
Russian	±0.01	±0.01	±0.01	±0.02	±0.03	±0.03
Persian	±0.10	±0.06	±0.06	±0.05	±0.04	±0.03
Arabic	±0.04	±0.03	±0.02	±0.02	±0.03	±0.04
Chinese	±0.04	±0.04	±0.04	±0.05	±0.06	±0.08
English	±0.04	±0.05	±0.05	±0.06	±0.08	±0.09
Thai	±0.03	±0.03	±0.04	±0.06	±0.08	±0.09
Portuguese	±0.08	±0.10	±0.09	±0.11	±0.11	±0.10

Our results show empirical evidence of inconsistent language labels in the historical data feed between 2014 and 2017. Our margin of error for estimating contagion ratios narrows down as FastText-LID and Twitter-LID unanimously yield their language predictions for the majority of messages authored in recent years. Future investigations can help us shed light on some of the implicit biases of language detection models. Nonetheless, our analysis supports our findings regarding the growth of retweets over time across most languages.

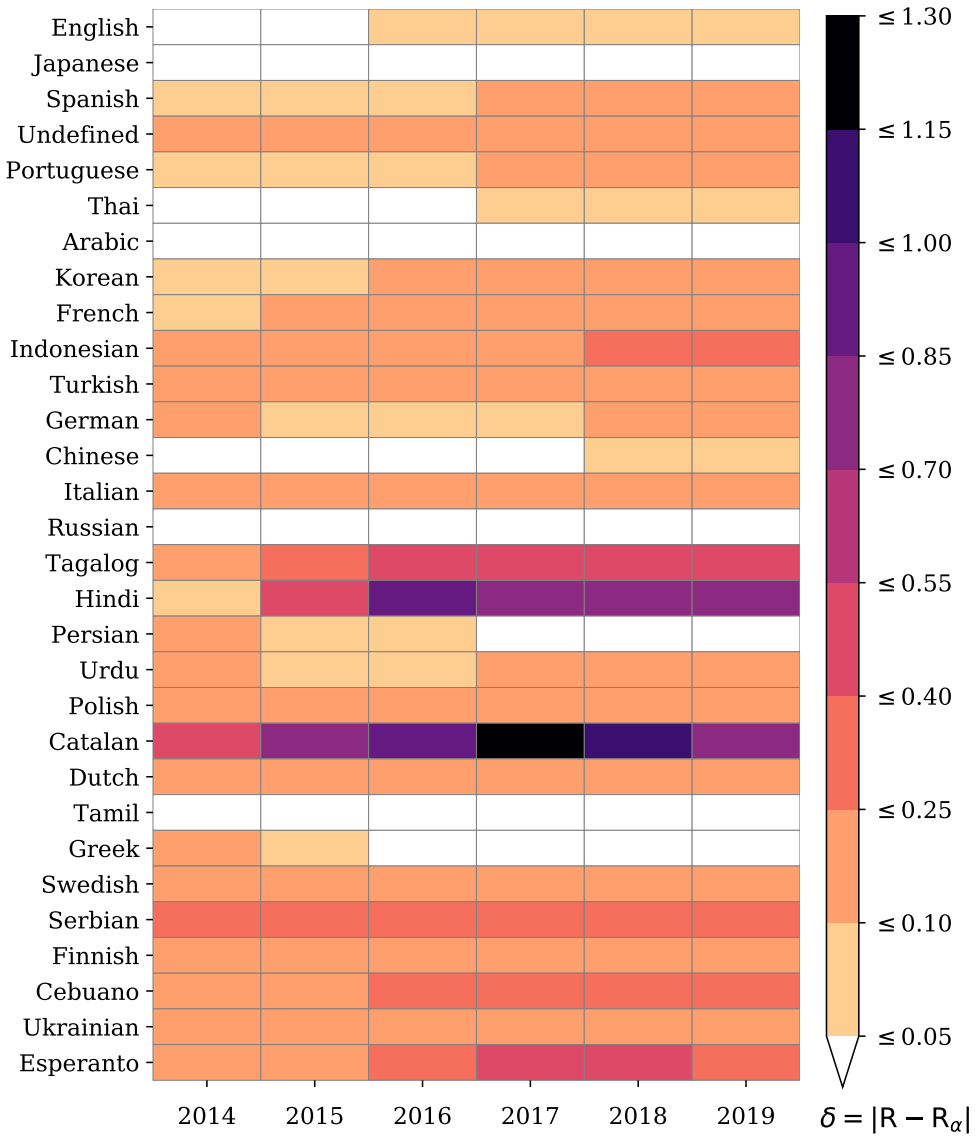

Figure 2.B.1: ***Margin of error for contagion ratios.*** *We compute the annual average of contagion ratios* R *for all messages in the top 30 ranked languages as classified by FastText-LID and described in Sec. 2.5.3. Similarly, we compute the annual average of contagion ratios* R_α *for the subset of messages that both classifiers have unanimously labeled their language labels. We display the absolute difference* $\delta = |R - R_\alpha|$ *to indicate our margin of error for estimating contagion ratios as a function of the agreement between FastText-LID and Twitter-LID models. White cells indicate that* δ *is below* .05*, whereas colored cells highlight values that are equal to, or above* .05*. We show the top 10 languages with the highest average values of* δ*'s per year in Table 2.B.3. We also show the bottom 10 languages with the lowest average values of* δ*'s per year in Table 2.B.4.*

2.C Impact of tweet's length on language detection

The informal and short texture of messages on Twitter—among many other reasons—makes language detection of tweets remarkably challenging. Twitter has also introduced several changes to the platform that notably impacted language identification. Particularly, users were limited to 140 characters per message before the last few months of 2017 and 280 characters thereafter [249]. To investigate the level of uncertainty of language detection as a function of tweet length, we take a closer look at the number of messages that are classified differently by FastText-LID and Twitter-LID for the top 10 most used languages on the platform between 2020-01-01 and 2020-01-07.

In Fig. 2.C.1, we display the daily number of mismatches (grey bars) between 2020-01-01 and 2020-01-07 (approximately 32 million messages for each day for the top-10 used languages), whereas the black line shows an average of that whole week. We also display a histogram of the average number of characters of each message throughout that period. We note that the distribution is remarkably skewed towards shorter messages on the platform. The average length of messages is less than 140 characters, with a large spike around the 140 character mark. Long messages—which include messages with links, hashtags, and emojis—can exceed the theoretical 280 character limit because we do not follow the same guidelines outlined by Twitter for counting the number of characters in each message.[8] For simplicity, we use the

[8] https://developer.twitter.com/en/docs/basics/counting-characters

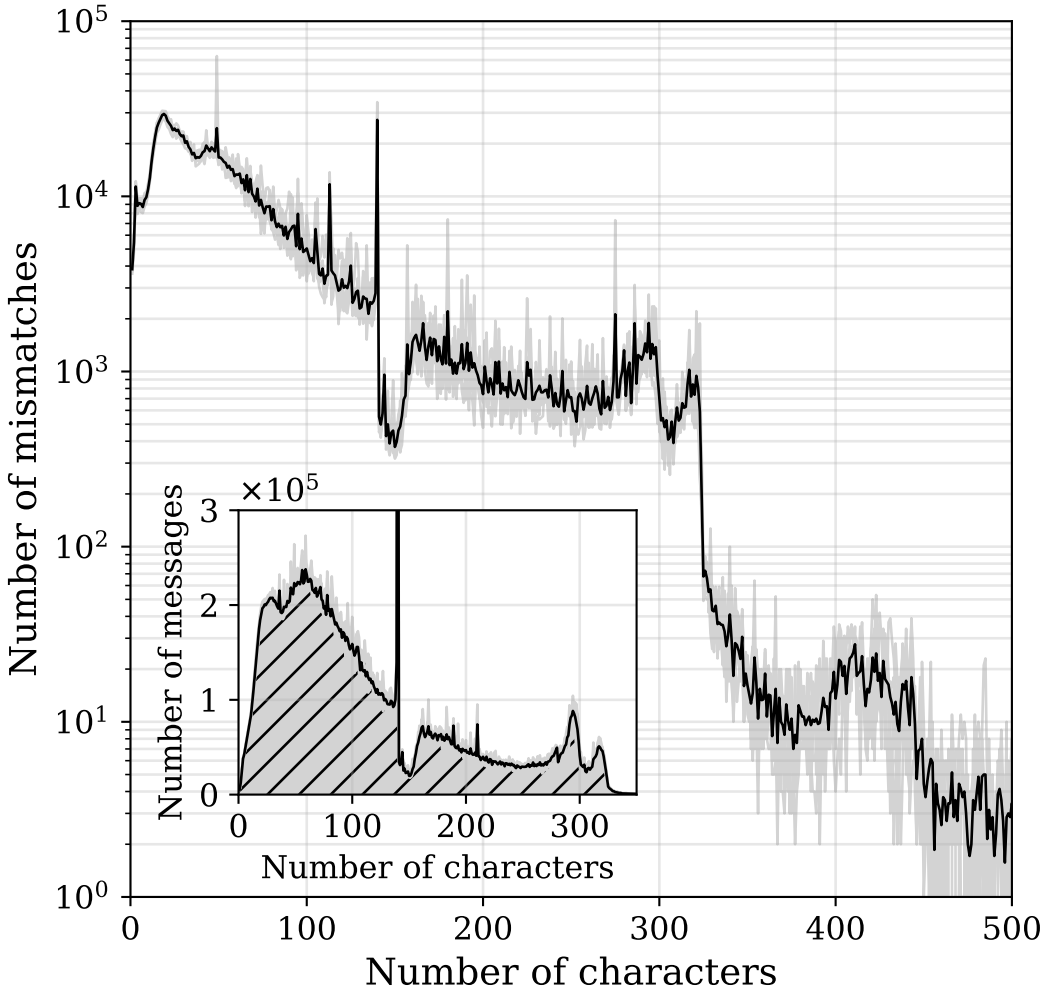

Figure 2.C.1: ***Language identification uncertainty as a function of tweet-length for top 10 most used languages on Twitter.*** *We display the number of messages that were classified differently by Twitter-LID model and FastText-LID for the top-10 prominent languages as a function of the number of characters in each message. Unlike Twitter, we count each character individually, which is why the length of each message may exceed the 280 character limit. The grey lines indicate the daily number of mismatches between 2020-01-01 and 2020-01-07 (approximately* 32 *million messages for each day for the top-10 used languages), whereas the black line shows an average of that whole week.*

built-in Python function to get the exact number of characters in a given message.[9]

[9] https://docs.python.org/3/library/functions.html#len

Table 2.C.1: Average daily messages for the top 10 languages between 2020-01-01 and 2020-01-07 (approximately 32 million messages for each day).

Language	**Messages**	**Mismatches**
English	1.1×10^7	.0853
Japanese	6.8×10^6	.0268
Spanish	2.3×10^6	.0558
Thai	2.2×10^6	.0161
Portuguese	2.1×10^6	.0565
Korean	1.7×10^6	.0085
Arabic	1.5×10^6	.0080
Indonesian	8.1×10^5	.1203
French	7.9×10^5	.1305
Turkish	5.6×10^5	.0325

As anticipated, our results indicate a higher proportion of short messages classified differently by FastText-LID and Twitter-LID models. We highlight the average percentage of mismatches for the top 10 most used languages in Table 2.C.1 (languages are sorted by popularity).

Furthermore, we examine a sample of messages authored through the month before and after the switch to the 280 character limit. We do not observe any distributional changes in FastText-LID's confidence scores between the two months. We categorize messages into four classes based on the confidence scores we get from FastText-LID's neural network. Predictions with confidence scores below .25 are labeled as Undefined (und). On the other hand, messages with scores greater or equal to .25 but less than .5 are flagged as predictions with low confidence (low). Predictions that have scores

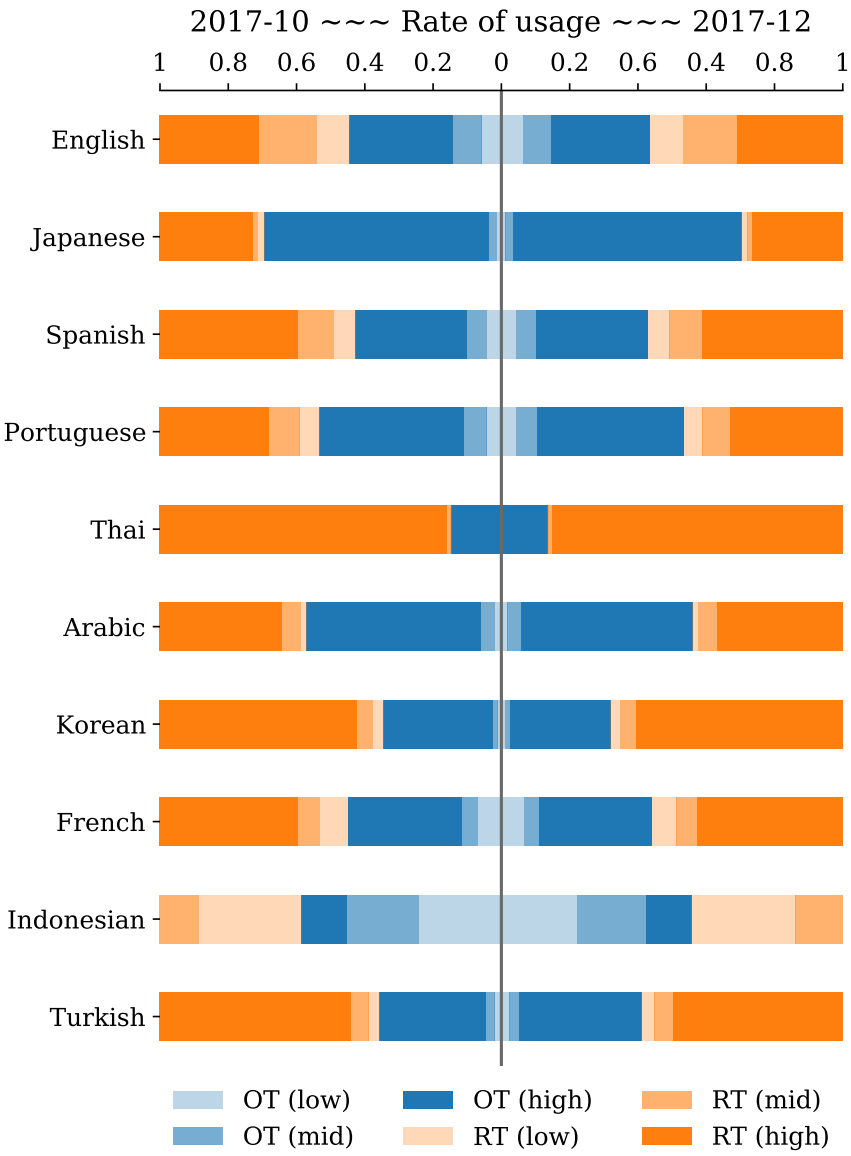

Figure 2.C.2: ***Confidence scores of the FastText-LID neural network predictions for the month before and after the shift to 280 characters.*** *We categorize messages into four classes based on the confidence scores we get from FastText-LID's neural network. Predictions with confidence scores below .25 are labeled as Undefined (und). Messages with scores greater or equal to .25 but less than .5 are flagged as predictions with low confidence (low). Predictions that have scores in the range [.5, .75) are considered moderate (mid), and messages with higher scores are labeled as predictions with high confidence (high). We note a symmetry indicating that the shift did not have a large impact on the network's predictions across organic and retweeted messages.*

in the range [.5, .75) are considered moderate (mid), and messages with higher scores are labeled as predictions with high confidence (high).

In Fig. 2.C.2, we display the relative proportion of messages for each of the confidence classes outlined above. First and foremost, we observe a very symmetrical layout indicating that the shift does not have a notable impact on the network's confidence in its predictions between the two months examined here across organic and retweeted messages.

Moreover, we note that the overall rate of usage for each language does not change before and after the switch to longer messages. To validate that, we take a closer look at the rate of usage for the top 10 most used languages throughout the past three years. In Fig. 2.C.3A, we observe a very consistent frequency of usage across all languages, indicating that the mechanistic shift to allow users to post longer messages does not have a notable impact on the language detection process. Fig. 2.C.3B and Fig. 2.C.3C show the growth of long messages on the platform, while the rate of usage for the most used languages remains consistent. In Fig. 2.C.3C, we see the adoption of longer messages starting in 2017, however, short messages still represent the majority of messages on the platform which comprise 75% of all messages as of 2019.

We observe a much higher ratio of retweets in longer messages than shorter messages. As of 2019, about 25% of all messages are long messages, and surprisingly, 80% of these long messages are retweets. However, we only examined the use of languages over time from a statistical point of view. The use of longer messages and the rate at which they are likely to be retweeted are different across languages. Further investigations will be needed to explore and explain this phenomenon.

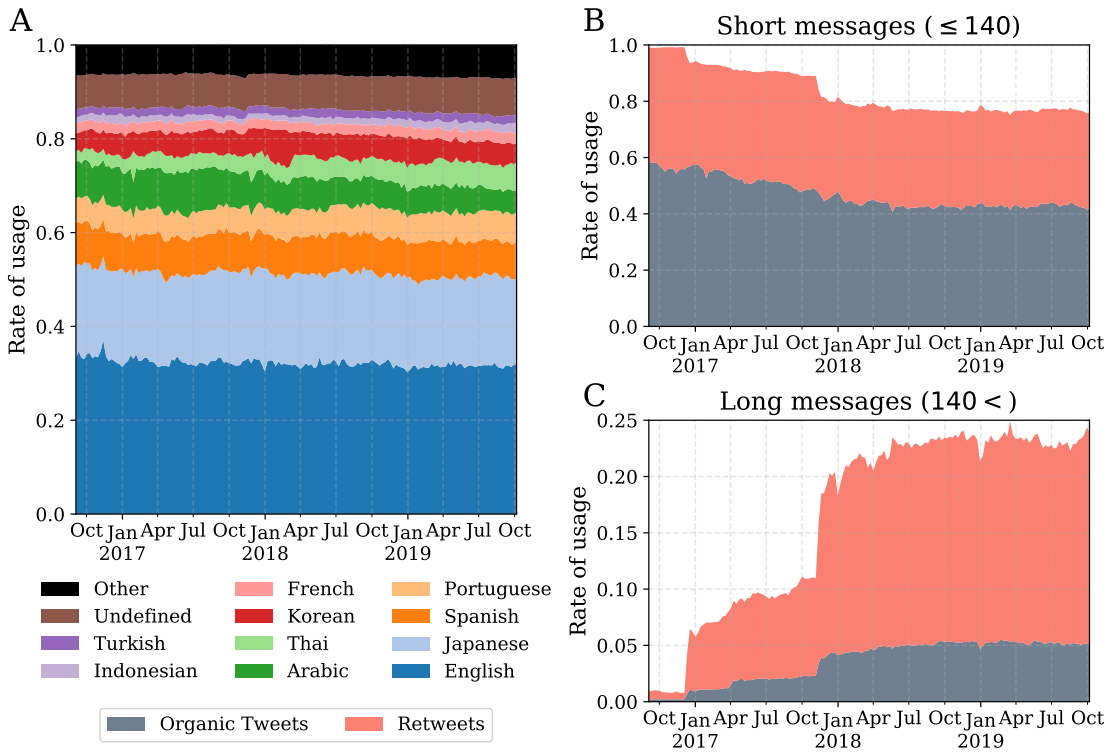

Figure 2.C.3: ***Weekly rate of usage for short and long messages. A.*** *Rate of usage for the top-10 used languages averaged at the week scale for the past three years. The introduction of long messages (i.e., above 140 but below 280 characters) does not change the overall language usage on the platform.* ***B–C.*** *The growth of long messages over time across organic and retweeted messages. We observe a much higher ratio of retweets in longer messages than shorter messages.*

Chapter 3

Day-scale curation of social media data streams

3.1 Abstract

In real-time, Twitter strongly imprints world events, popular culture, and the day-to-day, recording an ever growing compendium of language change. Vitally, and absent from many standard corpora such as books and news archives, Twitter also encodes popularity and spreading through retweets. Here, we describe Storywrangler, an ongoing curation of over 100 billion tweets containing 1 trillion 1-grams from 2008 to 2021. For each day, we break tweets into 1-, 2-, and 3-grams across 100+ languages, generating frequencies for words, hashtags, handles, numerals, symbols, and emojis. We make the dataset available through an interactive time series viewer, and as downloadable time series and daily distributions. Although Storywrangler leverages Twitter data, our method of tracking dynamic changes in n-grams can be extended to any temporally evolving corpus. Illustrating the instrument's potential, we present example use cases including social amplification, the sociotechnical dynamics of famous individuals, box office success, and social unrest.

3.2 Introduction

Our collective memory lies in our recordings—in our written texts, artworks, photographs, audio, and video—and in our retellings and reinterpretations of that which becomes history. The relatively recent digitization of historical texts, from books [56, 105, 197, 221] to news [19, 128, 132, 257] to folklore [2, 198, 280, 300] to governmental records [308], has enabled compelling computational analyses across many fields [2, 4, 237]. But books, news, and other formal records only constitute a specific type of text—carefully edited to deliver a deliberate message to a target audience. Large-scale constructions of historical corpora also often fail to encode a fundamental characteristic: popularity (i.e., social amplification). How many people have read a text? How many have retold a news story to others?

For text-based corpora, we are confronted with the challenge of sorting through different aspects of popularity of n-grams—sequences of n 'words' in a text that are formed by contiguous characters, numerals, symbols, emojis, etc. An n-gram may or may not be part of a text's lexicon, as the vocabulary of a text gives a base sense of what that text may span meaning-wise [256]. For texts, it is well established that n-gram frequency-of-usage (or Zipf) distributions are heavy-tailed [322]. Problematically, this essential character of natural language is readily misinterpreted as indicating cultural popularity. For a prominent example, the Google Books n-gram corpus [197], which in part provides inspiration for our work here, p resents year-scale, n-gram frequency time series where each book, in principle, counts only once [221]. All cultural fame is stripped away. The words of George Orwell's 1984 or Rick Riordan's Percy Jackson books, indisputably read and re-read by many people

around the world, count as equal to the words in the least read books published in the same years. And yet, time series provided by the Google Books n-gram viewer have regularly been erroneously conflated with the changing interests of readers (e.g., the apparent decline of sacred words [30, 156, 195, 221, 222]). Further compounded with an increase of scientific literature throughout the 20th Century, the corpus remains a deeply problematic database for investigations of sociolinguistic and cultural trends. It is also very difficult to measure cultural popularity. For a given book, we would want to know sales of the book over time, how many times the book has been actually read, and to what degree a book becomes part of broader culture. Large-scale corpora capturing various aspects of popularity exist [4], but are hard to compile as the relevant data is either prohibitively expensive or closed (e.g., Facebook), and, even when accessible, may not be consistently recorded over time (e.g., Billboard's Hot 100).

Now, well into the age of the internet, our recordings are vast, inherently digital, and capable of being created and shared in the moment. People, news media, governmental bodies, corporations, bots, and many other entities all contribute constantly to giant social media platforms. When open, these services provide an opportunity for us to attempt to track myriad statements, reactions, and stories of large populations in real-time. Social media data allows us to explore day-to-day conversations by millions of ordinary people and celebrities at a scale that is scarcely conventionalized and recorded. And crucially, when sharing and commenting mechanisms are native to a social media platform, we can quantify popularity of a trending topic and social amplification of a contemporary cultural phenomenon.

Here, we present Storywrangler, a natural language processing framework that

extracts, ranks, and organizes n-gram time series for social media. Storywrangler provides an analytical lens to examine discourse on social media, carrying both the voices of famous individuals—political figures and celebrities—and the expressions of the many. With a complex, ever-expanding fabric of time-stamped messages, Storywrangler allows us to capture storylines in over 150 languages in real time.

For a primary social media source, we use Twitter for several reasons, while acknowledging its limitations. Our method of extracting and tracking dynamic changes of n-grams can in principle be extended to any social media platform (e.g., Facebook, Reddit, Instagram, Parler, 4Chan, and Weibo).

Twitter acts as a distributed sociotechnical sensor system [134, 316]. Using Storywrangler, we can trace major news events and stories, from serious matters such as natural disasters [67, 101, 167, 230, 253] and political events [270] to entertainment such as sports, music, and movies. Storywrangler also gives us insights into discourse around these topics and myriad others including, violence, racism, inequality, employment, pop culture (e.g., fandom), fashion trends, health, metaphors, emerging memes, and the quotidian.

We can track and explore discussions surrounding political and cultural movements that are born and nurtured in real-time over social media with profound ramifications for society (e.g., #MeToo, #BlackLivesMatter, #QAnon). Modern social movements of all kinds, may develop a strong imprint on social media, over years in some cases, before becoming widely known and discussed.

Twitter and social media in general differ profoundly from traditional news and print media in various dimensions. Although amplification is deeply uneven, vast numbers of people may now express themselves to a global audience on any subject

that they choose (within limits of a service, and not without potential consequences). Unlike journalists, columnists, or book authors, people can instantly record and share messages in milliseconds. Importantly from a measurement perspective, this is a far finer temporal resolution than would be reasonably needed to explore sociocultural phenomena or reconstruct major events. The eye witness base for major events is now no longer limited to those physically present because of growing, decentralized live-streaming through various social media platforms. Social media thus enables, and not without peril, a kind of mass distributed journalism.

A crucial feature of Storywrangler is the explicit encoding of n-gram popularity, which is enabled by Twiter's social amplification mechanisms: retweets and quote tweets. For each day and across languages, we create Zipf distributions for the following: (1) n-grams from originally authored messages (OT), excluding all retweeted material (RT); and (2) n-grams from all Twitter messages (AT). For each day, we then have three key levels of popularity: n-gram lexicon, n-gram usage in organic tweets (originally authored tweets), and the rate at which a given n-gram is socially amplified (i.e., retweeted) on the platform. Our data curation using Storywrangler yields a rich dataset, providing an interdisciplinarity resource for researchers to explore transitions in social amplification by reconstructing n-gram Zipf distributions with a tunable fraction of retweets.

We structure this chapter as follows. In Sec. 3.3, we briefly describe our instrument, dataset, and the Storywrangler site which provides day-scale n-gram time series datasets for n=1, 2, and 3, both as time series and as daily Zipf distributions. In Sec. 3.4, we showcase a group of example analyses, arranged by increasing complication: Simple n-gram rank time series (Sec. 3.4.1); qualitative comparison to other

prominent social signals of Google Trends and cable news (Sec. 3.4.2); Contagiograms, time series showing social amplification (Sec. 3.4.3); analysis for identifying and exploring narratively trending storylines (Sec. 3.4.4); and an example set of case studies bridging n-gram time series with disparate data sources to study famous individuals, box office success, and social unrest (Sec. 3.4.5). In our concluding remarks in Sec. 3.5, we outline some potential future developments for Storywrangler.

3.3 Data and methods

3.3.1 Overview of Storywrangler

We draw on a storehouse of messages comprising roughly 10% of all tweets collected from 2008-09-09 onwards, and covering 150+ languages. In previous work [7], we described how we re-identified the languages of all tweets in our collection using FastText-LID[1] [32, 143], uncovering a general increase in retweeting across Twitter over time. A uniform language re-identification was needed as Twitter's own real-time identification algorithm was introduced in late 2012 and then adjusted over time, resulting in temporal inconsistencies for long-term streaming collection of tweets [84]. While we can occasionally observe subtle cues of regional dialects and slang, especially on a non-mainstream media platform like Twitter, we still classify them on the basis of their native languages. Date and language are the only metadata that we incorporate into our database. For user privacy in particular, we discard all other information associated with a tweet.

For each day t (Eastern Time encoding) and for each language ℓ, we categorize

[1] https://fasttext.cc/docs/en/language-identification.html

tweets into two classes: organic tweets (OT), and retweets (RT). To quantify the relative effect of social amplification, we group originally authored posts—including the comments found in quote tweets but not the retweeted content they refer to—into what we call organic tweets. We break each tweet into 1-grams, 2-grams, and 3-grams. Although we can identify tweets written in continuous-script-based languages (e.g., Japanese, Chinese, and Thai), our current implementation does not support breaking them into n-grams.

We accommodate all Unicode characters, including emojis, contending with punctuation as fully as possible (see Appendix 3.A for further details). For our application, we designed a custom n-gram tokenizer to preserve handles, hashtags, date/time strings, and links [similar to the tweet tokenizer in the Natural Language Toolkit (NLTK) [180]]. Although some older text tokenization toolkits followed different criteria, our protocol is consistent with modern computational linguistics for social media data [26, 132].

We derive three essential measures for each n-gram: raw frequency (or count), normalized frequency (interpretable as probability), and rank, generating the corresponding Zipf distributions [322]. We perform this process for all tweets (AT), organic tweets (OT), and (implicitly) retweets (RT). We then record n-grams along with ranks, raw frequencies, and normalized frequencies for all tweets and organic tweets in a single file, with the default ordering according to n-gram prevalence in all tweets.

3.3.2 Notation and measures

We write an n-gram by τ and a day's lexicon for language ℓ—the set of distinct n-grams found in all tweets (AT) for a given date t—by $\mathcal{D}_{t,\ell;n}$. We write n-gram raw frequency as $f_{\tau,t,\ell}$, and compute its usage rate in all tweets written in language ℓ as

$$p_{\tau,t,\ell} = \frac{f_{\tau,t,\ell}}{\sum_{\tau' \in \mathcal{D}_{t,\ell;n}} f_{\tau',t,\ell}}. \tag{3.1}$$

We further define the set of unique language ℓ n-grams found in organic tweets as $\mathcal{D}^{(\mathrm{OT})}_{t,\ell;n}$, and the set of unique n-grams found in retweets as $\mathcal{D}^{(\mathrm{RT})}_{t,\ell;n}$ (hence $\mathcal{D}_{t,\ell;n} = \mathcal{D}^{(\mathrm{OT})}_{t,\ell;n} \cup \mathcal{D}^{(\mathrm{RT})}_{t,\ell;n}$). The corresponding normalized frequencies for these two subsets of n-grams are then:

$$p^{(\mathrm{OT})}_{\tau,t,\ell} = \frac{f^{(\mathrm{OT})}_{\tau,t,\ell}}{\sum_{\tau' \in \mathcal{D}^{(\mathrm{OT})}_{t,\ell;n}} f^{(\mathrm{OT})}_{\tau',t,\ell}}, \text{ and} \tag{3.2}$$

$$p^{(\mathrm{RT})}_{\tau,t,\ell} = \frac{f^{(\mathrm{RT})}_{\tau,t,\ell}}{\sum_{\tau' \in \mathcal{D}^{(\mathrm{RT})}_{t,\ell;n}} f^{(\mathrm{RT})}_{\tau',t,\ell}}. \tag{3.3}$$

We rank n-grams by raw frequency of usage using fractional ranks for ties. The corresponding notation is:

$$r_{\tau,t,\ell},\ r^{(\mathrm{OT})}_{\tau,t,\ell},\ \text{and}\ r^{(\mathrm{RT})}_{\tau,t,\ell}. \tag{3.4}$$

3.3.3 User interface

We make interactive times series based on our n-gram dataset viewable at storywrangling.org. In Fig. 3.3.1, we show a screenshot of the site displaying rank time series

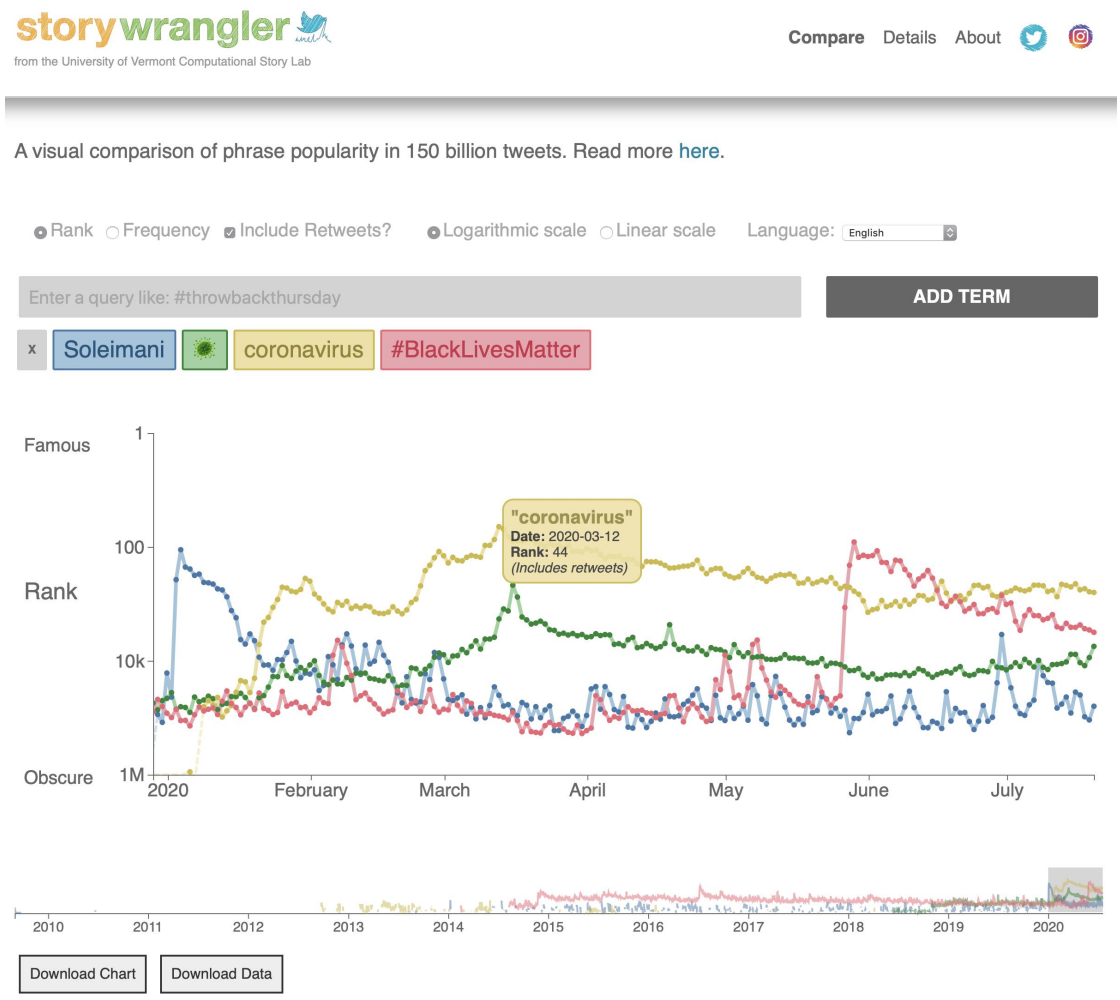

Figure 3.3.1: ***Interactive online viewer.*** *Screenshot of the Storywrangler site showing example Twitter n-gram time series for the first half of 2020. The series reflect three global events: The assassination of Iranian general Qasem Soleimani by the United States on 2020-01-03, the COVID-19 pandemic (the virus emoji and 'coronavirus'), and the Black Lives Matter protests following the murder of George Floyd by Minneapolis police ('#BlackLivesMatter'). The n-gram Storywrangler dataset for Twitter records the full ecology of text elements, including punctuation, hashtags, handles, and emojis. The default view is for n-gram (Zipfian) rank at the day scale (Eastern Time), a logarithmic y-axis, and for retweets to be included. These settings can be respectively switched to normalized frequency, linear scale, and organic tweets (*OT*) only. The displayed time range can be adjusted with the selector at the bottom, and all data is downloadable.*

for the first half of 2020 for 'Soleimani', the virus emoji, 'coronavirus', and '#BlackLivesMatter'. Ranks and normalized frequencies for n-grams are relative to n-grams with the same n, and in the online version we show time series on separate axes below the main comparison plot.

For each time series, hovering over any data point will pop up an information box. Clicking on a data point will take the user to Twitter's search results for the n-gram for the span of three days centered on the given date. All time series are shareable and downloadable through the site, as are daily Zipf distributions for the top million ranked n-grams in each language. Retweets may be included (the default) or excluded, and the language, vertical scale, and time frame may all be selected.

3.4 Results

3.4.1 Basic rank time series

In Fig. 3.4.1, we show rank time series for eight sets of n-grams from all tweets (i.e., including retweets). The n-gram groups move from simple to increasingly complex in theme, span a number of languages, and display a wide range of sociotechnical dynamics. Because of an approximate obeyance of Zipf's law ($f \sim r^{-\theta}$), we observe that normalized frequency of usage time series match rank time series in basic form. We use rank as the default view for its straightforwardness.

Starting with time and calendars, Fig. 3.4.1A gives a sense of how years are mentioned on Twitter. The dynamics show an anticipatory growth, plateau, and then rapid decay, with each year's start and finish marked by a spike.

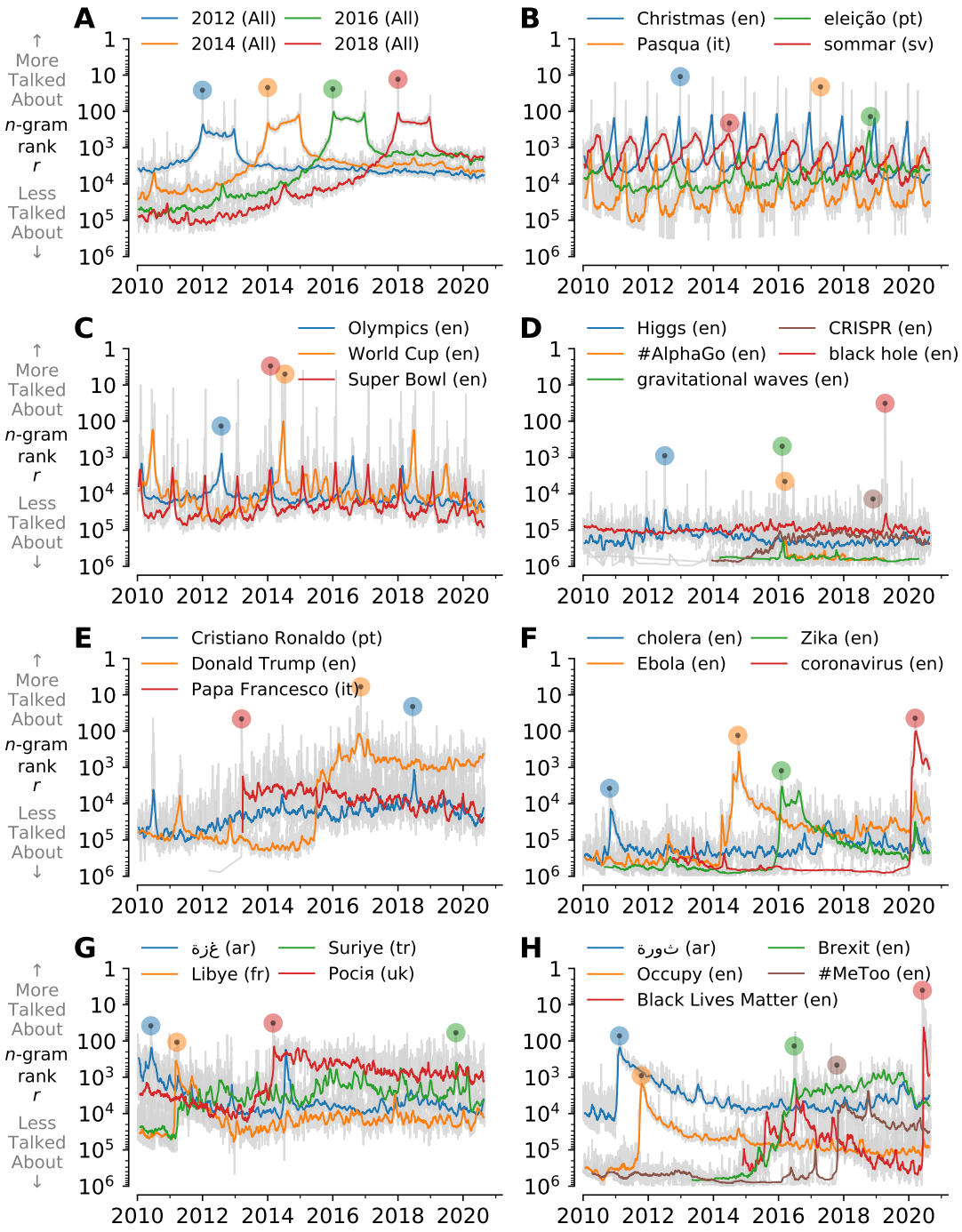

Figure 3.4.1: ***Thematically connected n-gram time series.***

Figs. 3.4.1B and C show calendrically anchored rank time series for seasonal, religious, political, and sporting events that recur at the scale of years in various languages. Periodic signatures at the day, week, and year scale are prominent on Twitter, reflecting the dynamics of the Earth, Moon, and Sun. Easter (shown in Italian) in particular combines cycles of all three. Major sporting events produce time series with strong anticipation, and can reach great heights of attention as exemplified by a peak rank of $r = 3$ for 'Super Bowl' on 2014-02-02.

We move to scientific announcements in Fig. 3.4.1D with the 2012 discovery of the Higgs boson particle (blue), detection of gravitational waves (green), and the first imaging of a black hole (red). For innovations, we show the time series of '#AlphaGo'—the first artificial intelligence program to beat the human Go champion (orange), along with the development of CRISPR technology for editing genomes (brown). We see that time series for scientific advances generally show shock-like responses with little anticipation or memory [75]. CRISPR is an exception for these few examples as through 2015, it moves to a higher, enduring state of being referenced.

Fame is the state of being talked about and famous individuals are well reflected on Twitter [82]. In Fig. 3.4.1E, we show time series for the Portuguese football player Cristiano Ronaldo, the 45th US president Donald Trump, and Pope Francis (Papa Francesco in Italian). All three show enduring fame, following sudden rises for both Trump and Pope Francis. On November 9, 2016, the day after the US election, 'Donald Trump' rose to rank $r = 6$ among all English 2-grams.

In Fig. 3.4.1F, we show example major infectious disease outbreaks over the past decade. Time series for pandemics are shocks followed by long relaxations, resurging both when the disease returns in prevalance and also in the context of new pandemics.

Cholera, Ebola, and Zika all experienced elevated discussion within the context of the COVID-19 pandemic.

In Fig. 3.4.1G, we show n-gram signals of regional unrest and fighting. The word for Gaza in Arabic tracks events of the the ongoing Israeli-Palestinian conflict. The time series for 'Libye' points to Opération Harmattan, the 2011 French and NATO military intervention in Libya. Similarly, the time series for 'Syria' in Turkish indicates the dynamics of the ongoing Syrian civil war on the region, and the build up and intervention of the Russian military in Ukraine is mirrored by the use of the Ukrainian word for 'Russia'.

In Fig. 3.4.1H, we highlight protests and movements. Both the time series for 'revolution' in Arabic and 'Occupy' in English show strong shocks followed by slow relaxations over the following years. The social justice movements represented by #MeToo and 'Black Lives Matter' appear abruptly, and their time series show slow decays punctuated by shocks returning them to higher ranks. Black Lives Matter resurged after the murder of George Floyd, with the highest one day rank of $r = 4$ occurring on 2020-06-02. By contrast, the time series of 'Brexit', the portmanteau for the movement to withdraw the United Kingdom from the European Union, builds from around the start of 2015 to the referendum in 2016, and then continues to climb during the years of complicated negotiations to follow.

3.4.2 Comparison to other signals

To highlight key differences that Storywrangler offers in contrast to other data sources, we display a few example comparisons in Fig. 3.4.2. In particular, we compare the usage rate for a set of n-grams using Storywrangler, Google Trends [54], and the

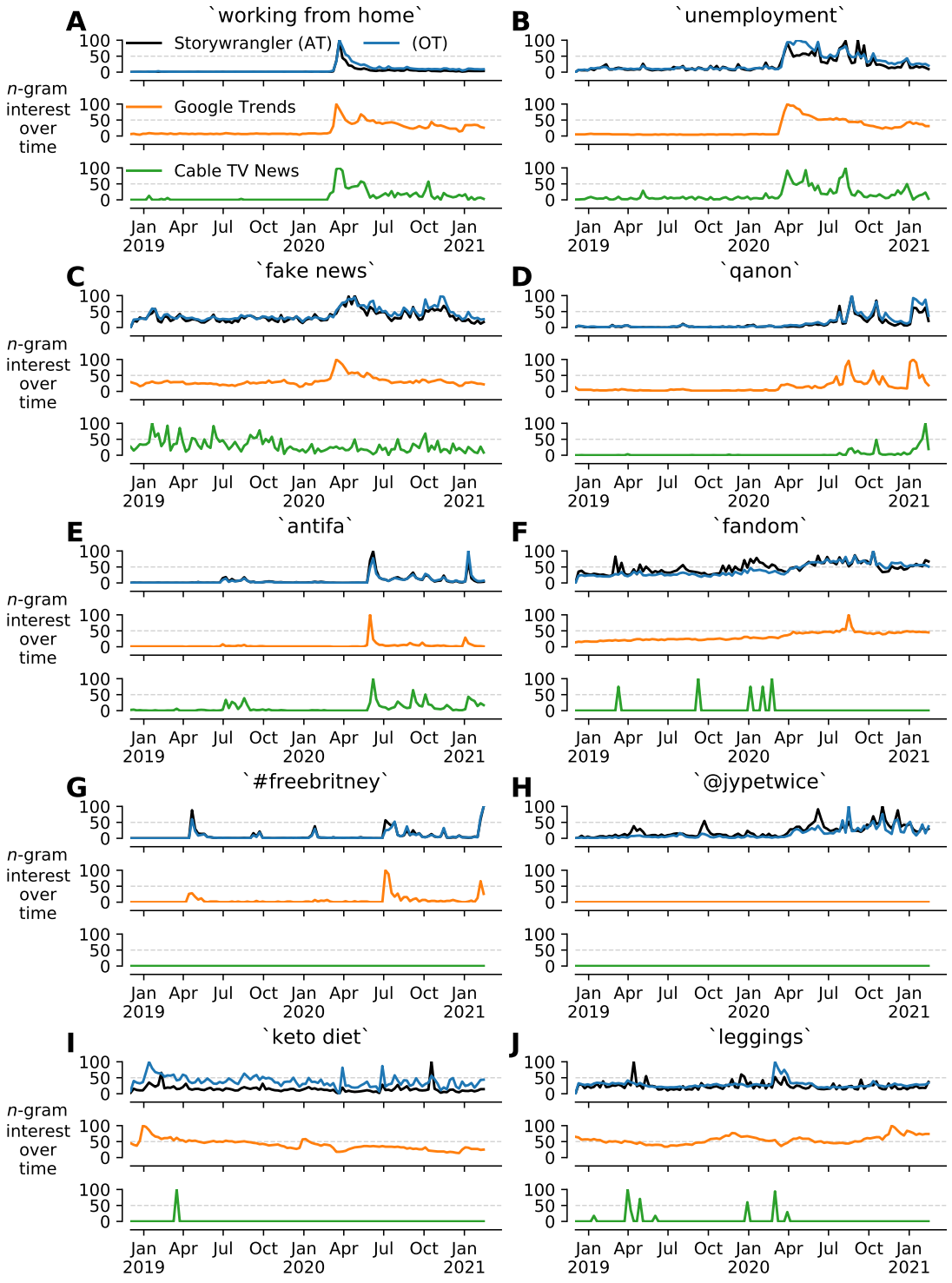

Figure 3.4.2: ***Comparison between Twitter, Google Trends, and Cable News.***

Stanford cable TV news analyzer [132].

Each data source has its own unique collection scheme that is most appropriate to that venue. Google Trends provides search interest scaled relative to a given region and time.[2] While Storywrangler is based on daily n-gram Zipf distributions, the Stanford cable TV news analyzer collects transcripts from most cable news outlets and breaks them into n-grams, recording screen time (seconds per day) for each term [132].

For the purpose of comparing n-gram usage across several disparate data sources, we take the weekly rate of usage for each term (case insensitive), and normalize each time series between 0 and 100 relative to the highest observed point within the given time window. A score of 100 represents the highest observed interest in the given term over time, while a value of 0 reflects low interest of that term and/or insufficient data. We display weekly interest over time for a set of 10 terms using Storywrangler for all tweets (AT, black), and originally authored tweets (OT, blue), Google Trends (orange), and cable news (green).

In Fig. 3.4.2A, we show how usage of the trigram 'working from home' peaks during March 2020 amid the COVID-19 pandemic. Although the term may be used in different contexts in each respective media source, we observe similar attention signals across all three data sources.

Similarly, Fig. 3.4.2B reveals increased mentions of 'unemployment' on all media platforms during the US national lockdown in April 2020. Individuals searching for unemployment claim forms could be responsible for the Google Trends spike, while news and social media usage of the term resulted from coverage of the economic crisis

[2] https://trends.google.com/trends/?geo=US

induced by the pandemic. The time series for 'unemployment' continues to fluctuate on Twitter, with distinct patterns across all tweets and originally authored tweets.

In Fig. 3.4.2C, we see the bigram 'fake news' roiling across social media and news outlets, reflecting the state of political discourse in 2020. Indeed, this period saw the most sustained usage of the term since its initial spike following the 2016 US election. The term was prominently searched for on Google in March 2020 during the early stages of the Coronavirus pandemic, but no corresponding spike is seen in cable news.

In Fig. 3.4.2D, the time series reveal attention to the 'QAnon' conspiracy theory on social media and Google Trends starting in mid 2020. Using Storywrangler, we note a spike of 'qanon' following Trump's remarks regarding violent far-right groups during the first presidential debate on September 29, 2020.[3] We see another spike of interest in October 2020 in response to the news about a kidnapping plot of the governor of Michigan by extremists.[4] Although the time series using both Storywrangler and Google Trends show sustained usage of the term in 2020, news outlets do not exhibit similar patterns until the US Capitol insurrection on January 6, 2021.

Fig. 3.4.2E shows mentions of 'antifa'—a political movement that drew attention in response to police violence during protests of the murder of George Floyd.[5] We note that mentions surged again in response to false flag allegations in the wake of the Capitol attack, most prominently on Twitter.

In Fig. 3.4.2F, we display interest over time of the term 'fandom'—a unigram that is widely used to refer to a group of people that share a common interest in creative genres, celebrities, fashion trends, modern tech, hobbies, etc. While this cultural

[3] https://www.nytimes.com/2020/09/30/us/politics/debate-takeaways.html
[4] https://www.nytimes.com/2020/10/08/us/gretchen-whitmer-michigan-militia.html
[5] https://www.nytimes.com/article/what-antifa-trump.html

phenomenon is rarely ever recorded by traditional news outlets, it dates back to the enormous fan base of Sherlock Holmes as one of the earliest signs of modern fandom, with public campaigners mourning the figurative death of their fictional character in 1893.[6] This cultural characteristic can not be easily captured with data sources such as Google books or search data. Nonetheless, it is intrinsic to non-mainstream media, illustrating the collective social attention of fans in various arenas as part of the ever changing digital pop culture.

Fig 3.4.2G shows a recent example where longtime fans of the pop music star, Britney Spears, organized and launched a social media support campaign in light of the controversy surrounding her conservatorship. Although the movement dates back to 2009, we see a surge of usage of the hashtag '#FreeBritney' in July 2020, after an interview with Britney's brother, revealing some personal details about her struggles and reigniting the movement on social media. The social movement has recently gained stronger cultural currency after the release of a documentary film by the New York Times in 2021.[7]

Moreover, Fig 3.4.2H shows interest over time of a popular South Korean pop band, 'Twice'. Although the official handle of the band on Twitter ('jypetwice') is virtually absent in other data sources, fans and followers use handles and hashtags regularly on Twitter to promote and share their comments for their musical bands.

In Figs. 3.4.2I and J, we see how communications and marketing campaigns of fitness trends such as 'keto diet', and fashion trends such as leggings, and athleisure receive sustained interest on Twitter while only occasionally popping up in news via commercial advertisements on some cable channels.

[6] https://www.bbc.com/culture/article/20160106-how-sherlock-holmes-changed-the-world
[7] https://www.nytimes.com/article/framing-britney-spears.html

3.4.3 Contagiograms

While rank time series for n-grams give us the bare temporal threads that make up the tapestries of major stories, our dataset offers more dimensions to explore. Per our introductory remarks on the limitations of text corpora, the most important enablement of our database is the ability to explore story amplification.

In Fig. 3.4.3, we present a set of six 'contagiograms'. With these expanded time series visualizations, we convey the degree to which an n-gram is retweeted both overall and relative to the background level of retweeting for a given language. We show both rates because retweet rates change strongly over time and variably so across languages [7].

Each contagiogram has three panels. The main panel at the bottom charts, as before, the rank time series for a given n-gram. For contagiograms running over a decade, we show rank time series in this main panel with month-scale smoothing (black line), and add a background shading in gray indicating the highest and lowest rank of each week.

The top two panels of each contagiogram capture the raw and relative social amplification for each n-gram. First, the top panel displays the raw RT/OT balance, the monthly relative volumes of each n-gram in retweets (RT, orange) and organic tweets (OT, blue):

$$R_{\tau,t,\ell} = f^{(\mathrm{RT})}_{\tau,t,\ell} \Big/ \left(f^{(\mathrm{RT})}_{\tau,t,\ell} + f^{(\mathrm{OT})}_{\tau,t,\ell} \right). \tag{3.5}$$

When the balance of appearances in retweets outweighs those in organic tweets, $R_{\tau,t,\ell} > 0.5$, we view the n-gram as nominally being amplified, and we add a solid background for emphasis.

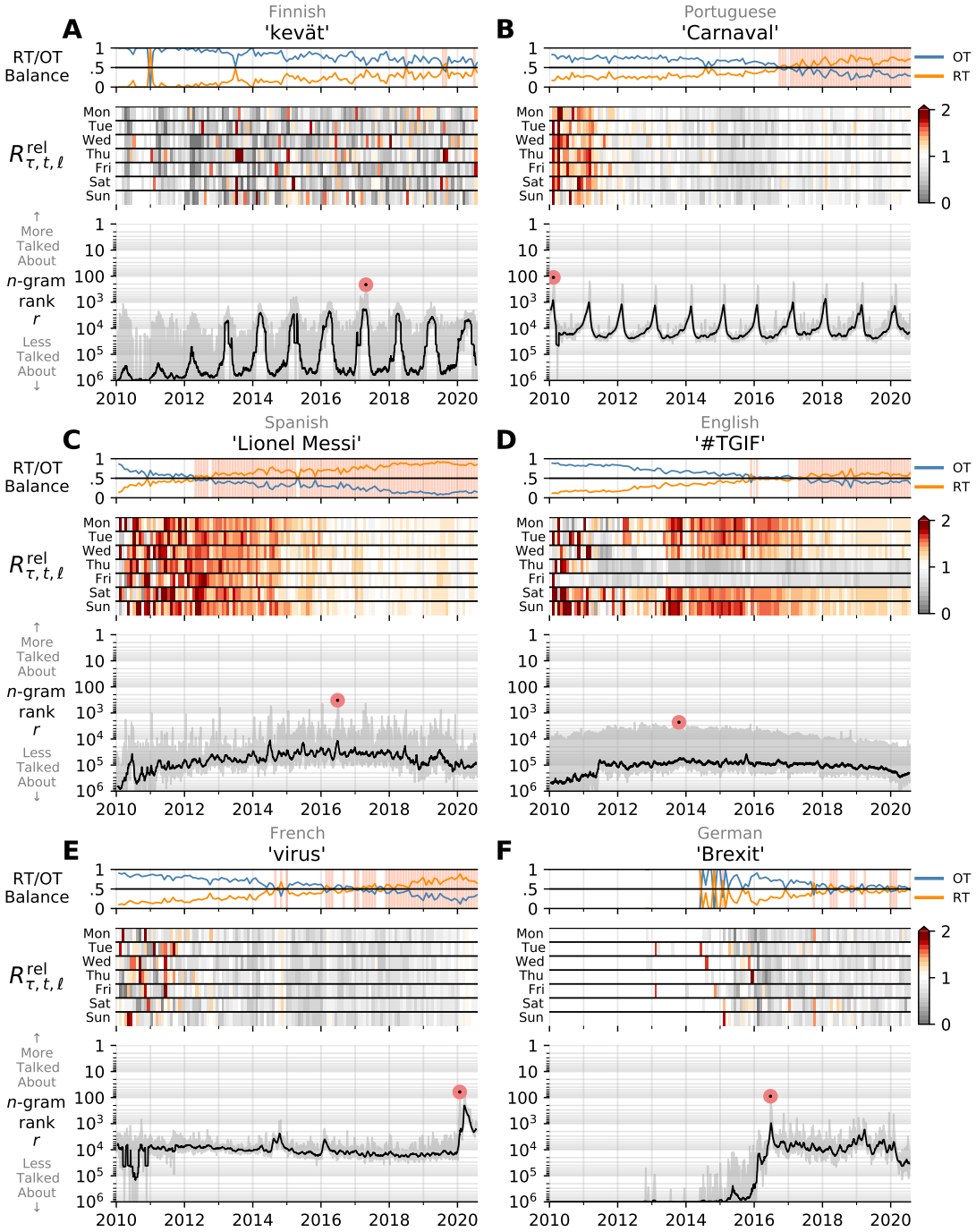

Figure 3.4.3: ***Contagiograms: Augmented time series charting the social amplification of*** n***-grams.***

Second, in the middle panel of each contagiogram, we display a heatmap of the values of the relative amplification rate for n-gram τ in language ℓ, $R^{\text{rel}}_{\tau,t,\ell}$. Building on from the RT/OT balance, we define $R^{\text{rel}}_{\tau,t,\ell}$ as:

$$R^{\text{rel}}_{\tau,t,\ell} = \frac{f^{(\text{RT})}_{\tau,t,\ell} \,/\, \left(f^{(\text{RT})}_{\tau,t,\ell} + f^{(\text{OT})}_{\tau,t,\ell}\right)}{\sum_{\tau'} f^{(\text{RT})}_{\tau',t,\ell} \,/\, \sum_{\tau'} \left(f^{(\text{RT})}_{\tau',t,\ell} + f^{(\text{OT})}_{\tau',t,\ell}\right)}, \tag{3.6}$$

where the denominator gives the overall fraction of n-grams that are found in retweets on day t for language ℓ. While still averaging at month scales, we now do so based on day of the week. Shades of red indicate that the relative volume of n-gram τ is being socially amplified over the baseline of retweets in language ℓ, $R^{\text{rel}}_{\tau,t,\ell} > 1$, while gray encodes the opposite, $R^{\text{rel}}_{\tau,t,\ell} < 1$.

The contagiogram in Fig. 3.4.3A for the word for 'kevät', 'spring' in Finnish, shows an expected annual periodicity. The word has a general tendency to appear in organic tweets more than retweets. But this is true of Finnish words in general, and we see that from the middle panel that kevät is in fact relatively, if patchily, amplified when compared to all Finnish words. For the anticipatory periodic times series in Fig. 3.4.3B, we track references to the 'Carnival of Madeira' festival—held forty days before Easter in Brazil. We see that 'Carnival' has become increasingly amplified over time, and has been relatively more amplified than Portuguese words except for 2015 and 2016.

By etymological definition, renowned individuals should feature strongly in retweets ('renown' derives from 'to name again'). Lionel Messi has been one of the most talked about sportspeople on Twitter over the past decade, and Fig. 3.4.3C shows his 2-gram is strongly retweeted, by both raw and relative measures. (See also Fig. 3.A.4F for

the K-pop band BTS's extreme levels of social amplification.)

Some n-grams exhibit a consistent weekly amplification signal. For example, '#TGIF' is organically tweeted on Thursdays and Fridays, but retweeted more often throughout the rest of the week (Fig. 3.4.3D). At least for those two days, individuals expressing relief for the coming weekend overwhelm any advertising from the eponymous restaurant chain.

Routinely, n-grams will take off in usage and amplification due to global events. In Fig. 3.4.3E, we see 'virus' in French tweets holding a stable rank throughout the 2010s before jumping in response to the COVID-19 pandemic, and showing mildly increased amplification levels. The word 'Brexit' in German has been prevalent from 2016 on, balanced in terms of organic tweet and retweet appearances, and generally more spread than German 1-grams.

The contagiograms in Fig. 3.4.3 give just a sample of the rich variety of social amplification patterns that appear on Twitter. We include some further examples in the supplementary material in Figs. 3.A.4 and 3.A.5. We provide Python package for generating arbitrary contagiograms along with further examples at `https://gitlab.com/compstorylab/contagiograms`. The figure-making scripts interact directly with the Storywrangler database, and offer a range of configurations.

3.4.4 Narratively trending storylines

Besides curating daily Zipf distributions, Storywrangler serves as an analytical tool to examine and explore the lexicon of emerging storylines in real-time. Using rank-turbulence divergence (RTD) [83], we examine the daily rate of usage of each n-gram assessing the subset of n-grams that have become most inflated in relative usage. For

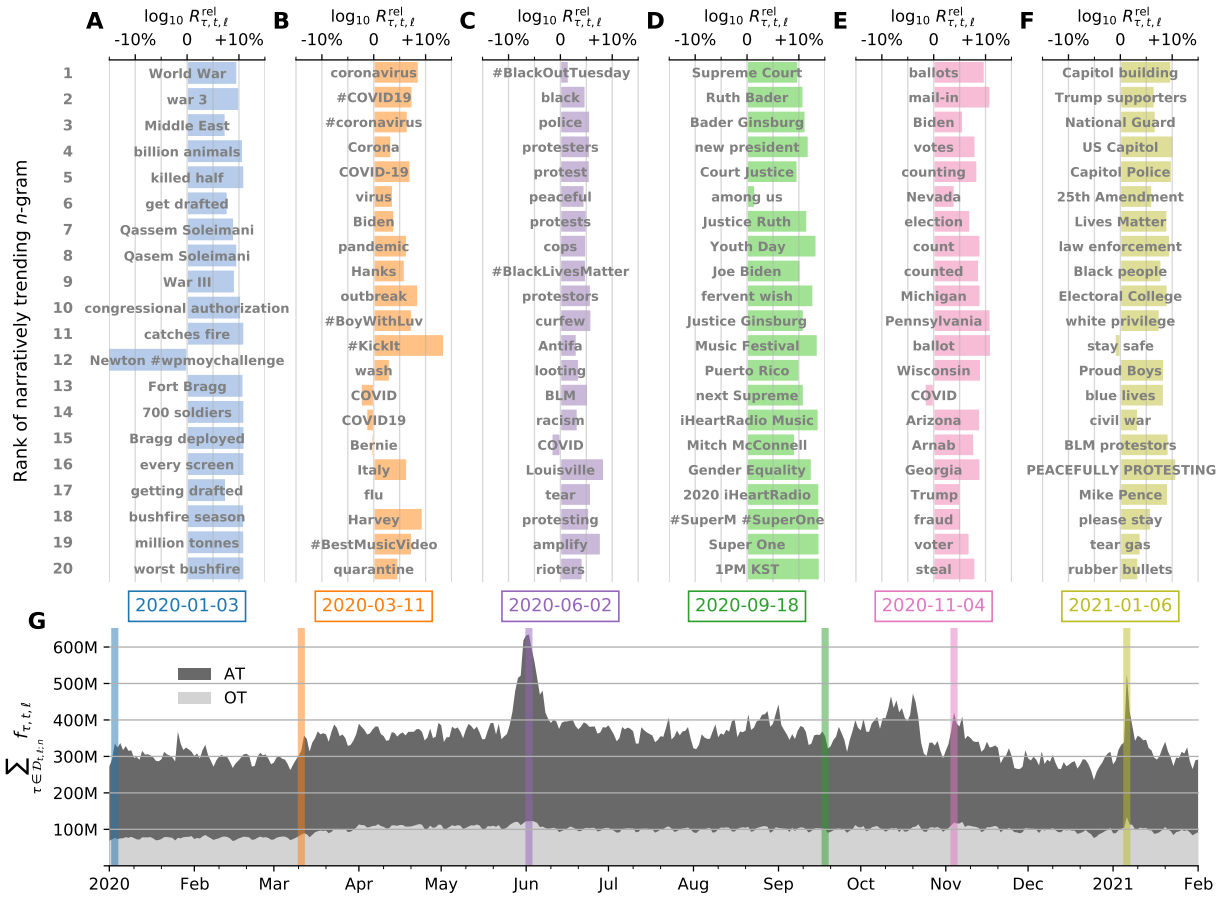

Figure 3.4.4: ***Narratively trending*** *n-grams. We use rank-turbulence divergence (RTD) [83] to find the most narratively trending n-grams of each day relative to the year before in English tweets. For each day, we display the top 20 n-grams sorted by their RTD value on that day. We also display the relative social amplification ratio* $R^{\mathrm{rel}}_{\tau,t,\ell}$ *for each n-gram on a logarithmic scale, whereby positive values indicate strong social amplification of that n-gram via retweets, and negative values imply that the given n-gram is often shared in originally authored tweets.* ***A.*** *The assassination of Iranian general Qasem Soleimani by a US drone strike on 2020-01-03 (blue).* ***B.*** *WHO declares COVID–19 a global Pandemic on 2020-03-11 (orange).* ***C.*** *Mass protests against racism and police brutality on 2020-06-02 (purple).* ***D.*** *Death of US Supreme Court justice Ruth Ginsburg from complications of pancreatic cancer on 2020-09-18 (green).* ***E.*** *The 2020 US presidential election held on 2020-11-04 (pink).* ***F.*** *The deadly insurrection of the US Capitol on 2021-01-06 (yellow).* ***G.*** *Daily n-gram volume (i.e., number of words) for all tweets (AT, grey), and organic tweets (OT, light-grey).*

each day t, we compute RTD for each n-gram τ relative to the year before t', setting the parameter α to 1/4 to examine the lexical turbulence of social media data such that:

$$\delta D_{\tau}^{\mathrm{R}} = \left| \frac{1}{r_{\tau,t,\ell}^{\alpha}} - \frac{1}{r_{\tau,t',\ell}^{\alpha}} \right|^{1/(\alpha+1)} ; (\alpha = 1/4). \tag{3.7}$$

Although our tool uses RTD to determine dramatic shifts in relative usage of n-grams, other divergence metrics will yield similar lists.

In Fig. 3.4.4, we show an example analysis of all English tweets for a few days of interest in 2020. First, we determine the top 20 narratively dominate n-grams of each day using RTD, leaving aside links, emojis, handles, and stop words but keeping hashtags. Second, we compute the relative social amplification ratio $R_{\tau,t,\ell}^{\mathrm{rel}}$ to examine whether a given n-gram τ is prevalent in originally authored tweets, or socially amplified via retweets on day t. For ease of plotting, we have further chosen to display $R_{\tau,t,\ell}^{\mathrm{rel}}$ at a logarithmic scale. Positive values of $\log_{10} R_{\tau,t,\ell}^{\mathrm{rel}}$ imply strong social amplification of τ, whereas negative values show that τ is relatively more predominant in organic tweets.

Fig. 3.4.4A gives us a sense of the growing discussions and fears of a global warfare following the assassination of Iranian general Qasem Soleimani by a US drone airstrike on 2020-01-03.[8] While most of the terms are socially amplified, we note that the bigram 'Newton #wpmoychallenge' was trending in organic tweets, reflecting the ongoing campaign and nomination of Cam Newton for Walter Payton NFL Man of the Year Award—an annual reward that is granted for an NFL player for their

[8] https://www.nytimes.com/2020/01/02/world/middleeast/qassem-soleimani-iraq-iran-attack.html

excellence and contributions.[9]

In Fig. 3.4.4B, we see how conversations of the Coronavirus disease becomes the most prevailing headline on Twitter with the World Health Organization (WHO) declaring COVID–19 a global Pandemic on 2020-03-11.

In light of the social unrest sparked by the murder of George Floyd in Minneapolis, we observe the growing rhetoric of the Black Lives Matter movement on Twitter driven by an enormous increase of retweets in Fig. 3.4.4C. The top narratively trending unigram is '#BlackOutTuesday'—a newborn movement that matured overnight on social media, leading to major music platforms such as Apple and Spotify to shut down their operations on 2020-06-02 in support of the nationwide protests against racism and police brutality.[10]

In Fig. 3.4.4D, we see the name of the US Supreme Court justice Ruth Bader Ginsburg amplified on Twitter, mourning her death from complications of pancreatic cancer on 2020-09-18. We also see names of politicians embodying the heated discourse on Twitter preceding the first US presidential debate. Emerging pop culture trends can also be observed in the anticipation of the first album by a K-pop South Korean band 'SuperM', entitled 'Super One'.[11]

In Fig. 3.4.4E, we see names of swing states and political candidates come to the fore during the US presidential election held on 2020-11-04. We observe another surge of retweets during the storming of the US Capitol by Trump supporters on 2021-01-06. Fig. 3.4.4F shows the top 20 prevalent bigrams emerging on Twitter in response to the deadly insurrection.

[9] https://www.panthers.com/news/cam-newton-named-walter-payton-nfl-man-of-the-year-nominee
[10] https://www.nytimes.com/2020/06/02/arts/music/what-blackout-tuesday.html
[11] https://en.wikipedia.org/wiki/Super_One_(album)

In Fig. 3.4.4G, we display the daily daily n-gram volume (i.e., number of words) throughout the year for all tweets (AT, grey), and organic tweets (OT, light-grey).

We provide more examples in Appendix 3.B and Figs. 3.B.2–3.B.3, demonstrating the wide variety of sociocultural and sociotechnical phenomena that can be identified and examined using Storywrangler.

3.4.5 Case studies

As a demonstration of our dataset's potential value to a diverse set of disciplines, we briefly present three case studies. We analyze (1) The dynamic behavior of famous individuals' full names and their association with the individuals' ages; (2) The relationship between movie revenue and anticipatory dynamics in title popularity; and (3) The potential of social unrest related words to predict future geopolitical risk.

We examine the dialog around celebrities by cross-referencing our English 2-grams corpus with names of famous personalities from the Pantheon dataset [317]. We searched through our English n-grams dataset and selected names that were found in the top million ranked 2-grams for at least one day between 2010-01-01 and 2020-06-01. In Fig. 3.4.5A, we display a monthly rolling average (centered) of the average rank for the top 5 individuals for each category $\langle r_{\min(5)} \rangle$ (see also Fig. 3.C.1). In Fig. 3.4.5B, we display a kernel density estimation of the top rank achieved by any of these individuals in each industry as a function of the number of years since the recorded year of birth. We note high density of individuals marking their best rankings between 40 and 60 years of age in the film and theatre industry. Different dynamics

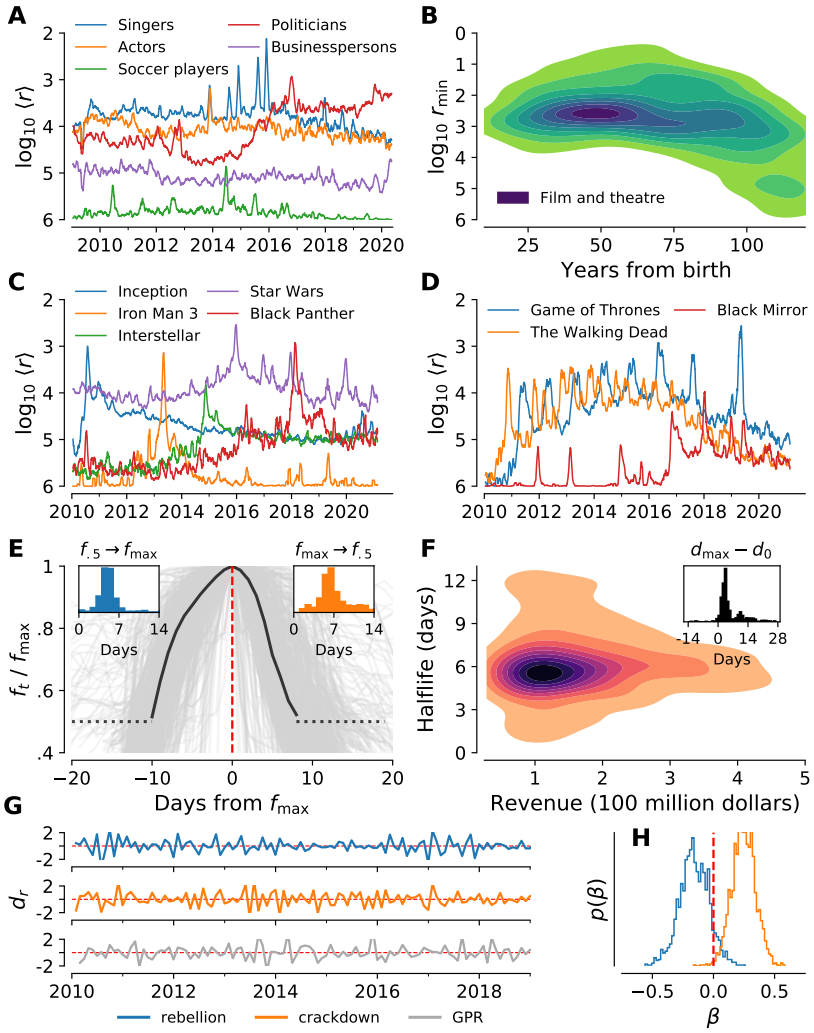

Figure 3.4.5: ***Three case studies joining Storywrangler with other data sources.*** ***A.*** *Monthly rolling average of rank $\langle r \rangle$ for the top-5 ranked Americans born in the past century in each category for a total of 960 individuals found in the Pantheon dataset [317].* ***B.*** *Kernel density estimation for the top rank $r_{\min}$ achieved by 751 personalities in the film and theater industry as a function of their age.* ***C.*** *Rank time series for example movie titles showing anticipation and decay.* ***D.*** *Contrasting with* ***C****, rank time series for TV series titles.* ***E–F.*** *Time series and half-life revenue comparison for 636 movie titles with gross revenue at or above the 95th percentile released between 2010-01-01 and 2017-07-31 [117].* ***G–H.*** *The Storywrangler dataset can also be used to potentially predict political and financial turmoil. Percent change in the words 'rebellion' and 'crackdown' in month m are significantly associated with percent change in a geopolitical risk index in month $m+1$ [44].* ***G.*** *Percent change time series.* ***H.*** *Distributions of coefficients of a fit linear model. See Appendix 3.C, 3.D, and 3.E for details of each study.*

can be observed in Fig. 3.C.1 for other industries.

We next investigate the conversation surrounding major film releases by tracking n-grams that appear in titles for 636 movies with gross revenue above the 95th percentile during the period ranging from 2010-01-01 to 2017-07-01 [117]. We find a median value of 3 days post-release for peak normalized frequency of usage for movie n-grams (Fig. 3.4.5F inset). Growth of n-gram usage from 50% ($f_{.5}$) to maximum normalized frequency ($f_{\max}$) has a median value of 5 days across our titles. The median value of time to return to $f_{.5}$ from $f_{\max}$ is 6 days. Looking at Fig. 3.4.5E we see the median shape of the spike around movie release dates tends to entail a gradual increase to peak usage, and a relatively more sudden decrease when returning to $f_{.5}$. There is also slightly more spread in the time to return to $f_{.5}$ than compared with the time to increase from $f_{.5}$ to $f_{\max}$ (Fig. 3.4.5E insets).

In Figs. 3.4.5G and H, we show that changes in word usage can be associated to future changes in geopolitical risk, which we define here as "a decline in real activity, lower stock returns, and movements in capital flows away from emerging economies", following the US Federal Reserve [44]. We chose a set of words that we *a priori* believed might be relevant to geopolitical risk as design variables and a geopolitical index created by the US Federal Reserve as the response. We fit a linear model using the values of the predictors at month m to predict the value of the geopolitical risk index at month m+1. Two of the words, 'rebellion' and 'crackdown', are statistically significantly associated with changes in the geopolitical risk index.

Although global events and breaking news are often recorded across conventional and modern social media platforms, Storywrangler uniquely tracks ephemeral day-to-day conversations and sociocultural trends. In creating Storywrangler, we sought

to develop and maintain a large-scale daily record of everyday dialogues that is complementary to existing data sources, but equally vital to identify and study emerging sociotechnical phenomena. For details about our methodology and further results, see Appendices 3.C to 3.E.

3.5 Discussion

With this initial effort, we aim to introduce Storywrangler as a platform enabling research in computational social science, data journalism, natural language processing, and the digital humanities. Along with phrases associated with important events, Storywrangler encodes casual daily conversation in a format unavailable through newspaper articles and books. While its utility is clear, there are many potential improvements to introduce to Storywrangler. For high volume languages, we would aim for higher temporal resolution—at the scale of minutes—and in such an implementation we would be limited by requiring n-gram counts to exceed some practical minimum. We would also want to expand the language parsing to cover continuous-script languages such as Japanese and Chinese.

Another large space of natural improvements would be to broadly categorize tweets in ways other than by language identification—while preserving privacy—such as geography, user type (e.g., people, institutions, or automated), and topic (e.g., tweets containing 'fake news'). We note that for Twitter, features such as location and user type are more difficult to establish with as much confidence as for language identification. Increasingly by design, geographic information is limited on Twitter as are user demographics, although some aspects may be gleaned indirectly [60, 179, 188,

235, 320]. Regardless, in this initial curation of Twitter n-grams, we purposefully do not attempt to incorporate any metadata beyond identified language into the n-gram database.

Topic-based subsets are particularly promising as they would allow for explorations of language use, ambient framings, narratives, and conspiracy theories. Parsing into 2-grams and 3-grams makes possible certain analyses of the temporal evolution of 1-grams adjacent to an anchor 1-gram or 2-gram. Future development will enable the use of wild cards so that linguists will in principle be able to track patterns of popular language use in a way that the Google Books n-gram corpus is unable to do [197, 221]. Similarly, journalists and political scientists could chart n-grams being used around, for example, '#BlackLivesMatter' or 'Trump' over time [85].

Looking outside of text, a major possible expansion of the instrument would be to incorporate image and video captions, a growing component of social media communications over the past decade. Moving away from Twitter, we could use Storywrangler for other platforms where social amplification is a recorded feature (e.g., Reddit, 4Chan, Weibo, and Parler).

There are substantive limitations to Twitter data, some of which are evident in many large-scale text corpora. Our n-gram dataset contends with popularity, allowing for the examination of story amplification, and we emphasize the importance of using contagiograms as visualization tools that go beyond presenting simple time series. Popularity, however, is notoriously difficult to measure. The main proxy we use for popularity is the relative rate of usage of a given n-gram across originally authored tweets, examining how each term or phrase is socially amplified via retweets. While Twitter attempts to measure popularity by counting impressions, is increasingly dif-

ficult to capture the number of people exposed to a tweet. Twitter's centralized trending feature is yet another dimension that alters the popularity of terms on the platform, personalizing each user timeline and inherently amplifying algorithmic bias. We have also observed a growing passive behavior across the platform leading to an increasing preference for retweets over original tweets for most languages on Twitter during the past few years [7].

Twitter's user base, while broad, is clearly not representative of the populace [193]; is moreover compounded by the mixing of voices from people, organizations, and bots, and has evolved over time as new users have joined. Still, modern social media provides an open platform for all people to carry out conversations that matter to their lives. Storywrangler serves as instrument to depict discourse on social media at a larger scale and finer time resolution than current existing resources. Sociocultural biases that are inherently intrinsic to these platforms will be exposed using Storywrangler, which can inspire developers to enhance their platforms accordingly [28, 155].

Social structures (e.g., news and social media platforms) form and reshape individual behavior, which evidently alters social structures in an algorithmic feedback loop fashion [107]. For instance, a trending hashtag can embody a social movement (e.g., #MeToo), such that an n-gram may become mutually constituted to a behavioral and sociocultural revolution. Social and political campaigns can leverage an n-gram in their organized marketing strategies, seeking sustained collective attention on social media platforms encoded through spikes in n-gram usage rates. There are many examples of this emerging sociotechnical phenomenon on Twitter ranging from civil rights (e.g., #WomensMarch) to gender identity (e.g., #LGBTQ) to political conspiracy theories (e.g., #QAnon) to academy awards promotions (e.g., #Oscar) to

movie advertisement (e.g., #Avengers), etc.

The Canadian awareness campaign 'Bell Let's Talk' is another example of an annual awareness campaign that subsidizes mental health institutions across Canada, donating 5 cents for every (re)tweet containing the hashtag '#BellLetsTalk' to reduce stigma surrounding mental illness. Marketing campaigns have also grasped the periodic feature of key trending n-grams and adjusted their language accordingly. Marketers and bots often exploit this periodicity by hijacking popular hashtags to broadcast their propaganda (e.g., including #FF and #TGIF as trending hashtags for Friday promotions).

In building Storywrangler, we have prioritized privacy by aggregating statistics to day-scale resolution for individual languages, truncating distributions, ignoring geography, and masking all metadata. We have also endeavored to make our work as transparent as possible by releasing all code associated with the API.

Although we frame Storywrangler as a research focused instrument akin to a microscope or telescope for the advancement of science, it does not have built-in ethical guardrails. There is potential for misinterpretation and mischaracterization of the data, whether purposeful or not. For example, we strongly caution against cherry picking isolated time series that might suggest a particular story or social trend. Words and phrases may drift in meaning and other terms take their place. For example, 'coronavirus' gave way to 'covid' as the dominant term of reference on Twitter for the COVID-19 pandemic in the first six months of 2020 [6]. To in part properly demonstrate a trend, researchers would need to at least marshal together thematically related n-grams, and do so in a data-driven way, as we have attempted to do for our case studies. Thoughtful consideration of overall and normalized frequency

of usage would also be needed to show whether a topic is changing in real volume.

In building Storywrangler, our primary goal has been to build an instrument to curate and share a rich, language-based ecology of interconnected n-gram time series derived from social media. We see some of the strongest potential for future work in the coupling of Storywrangler with other data streams to enable, for example, data-driven, computational versions of journalism, linguistics, history, economics, and political science.

3.6 Acknowledgments

The authors are grateful for the computing resources provided by the Vermont Advanced Computing Core and financial support from the Massachusetts Mutual Life Insurance Company and Google Open Source under the Open-Source Complex Ecosystems And Networks (OCEAN) project. The authors appreciate discussions and correspondence with Colin Van Oort, James Bagrow, and Randall Harp. We thank many of our colleagues at the Computational Story Lab for their feedback on this project. Computations were performed on the Vermont Advanced Computing Core supported in part by NSF award No. OAC-1827314.

Appendix

3.A Twitter Dataset

3.A.1 Language identification and detection

In previous work [7], we described how we re-identified the languages of all tweets in our collection using FastText [32, 143]. A uniform language re-identification was needed as Twitter's own real-time identification algorithm was introduced in late 2012 and then adjusted over time, resulting in temporal inconsistencies for long-term streaming collection of tweets [84].

While FastText is a language model that can be used for various text mining tasks, it requires an additional step of producing vector language representations to be used for LID. To accomplish that, we use an off-the-shelf language identification toolthat uses the word embeddings produced by the model.[12]

The word embeddings provided by FastText spans a wide set of languages, including some regional dialects. However, language detection of short text remains an outstanding challenge in NLP. While we hope to expand our language detection in

[12] https://fasttext.cc/docs/en/language-identification.html

Figure 3.A.1: ***Temporal summary statistics.*** ***A.*** *The grey bars show the daily unique number of n-grams, while the lines show a monthly rolling average for* 1*-grams (purple),* 2*-grams (yellow), and for* 3*-grams (pink).* ***B–D.*** *The growth of n-grams in our dataset by each category where n-grams captured from organic tweets (OT) are displayed in blue, retweets RT in green, and all tweets combined in grey.* ***E–G.*** *Normalized frequencies to illustrate the growth of retweets over time.*

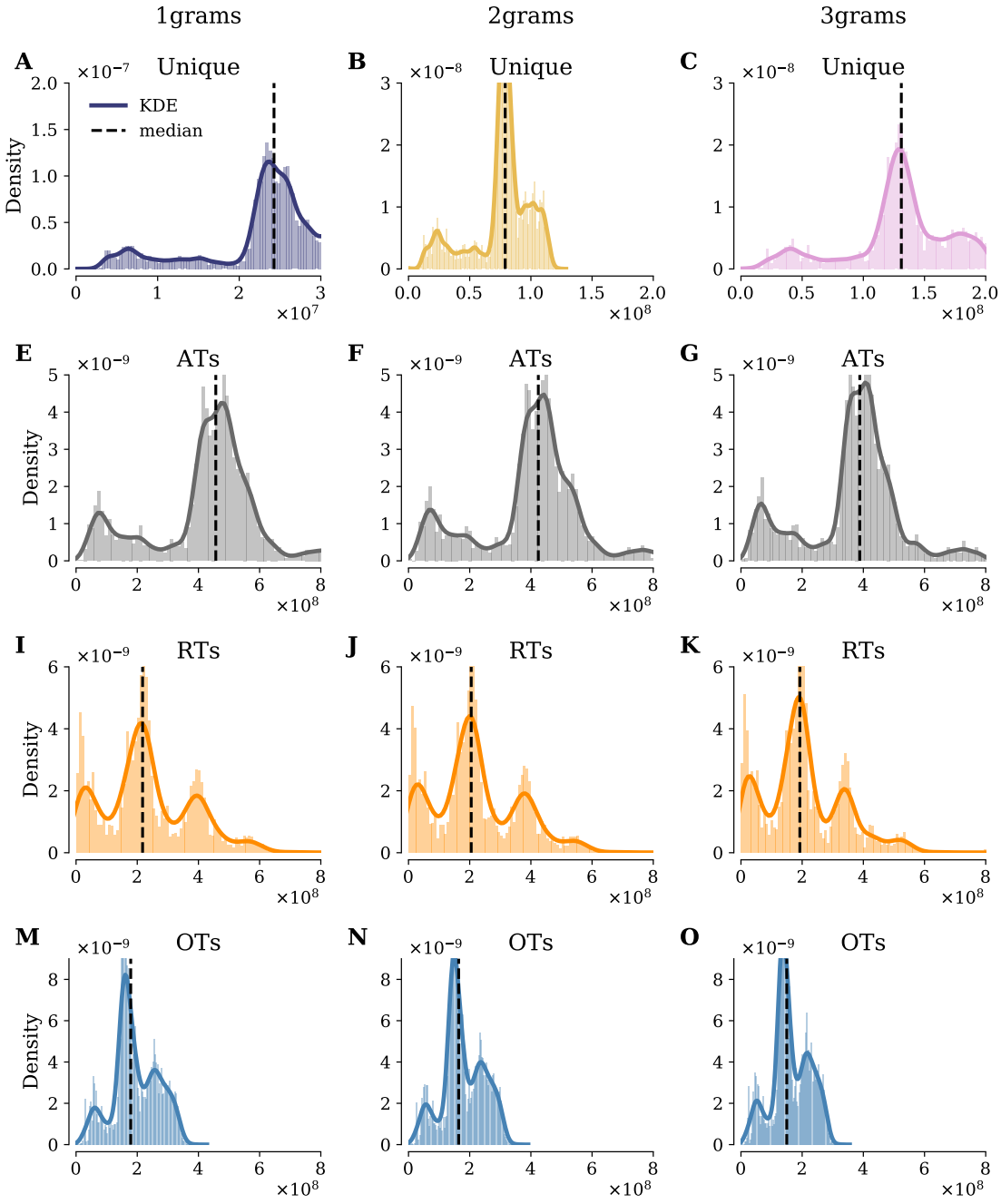

Figure 3.A.2: ***Kernel density estimations.*** ***A–C.*** *Distributions of unique unigrams, bigrams, and trigrams captured daily throughout the last decade.* ***E–G.*** *Distributions of n-grams occurrences in all tweets.* ***I–K.*** *Distributions of n-grams parsed from retweets (RT) only.* ***M–O.*** *Distributions of n-grams parsed from organic tweets (OT) only.*

future work, we still classify messages based on the languages identified by FastText-LID.

Importantly, in this work, we do not intend to reinvent FastText-LID, or improve upon existing LID tools. FastText-LID is on par with deep learning models in terms of accuracy and consistency, yet orders of magnitude faster in terms of inference and training time [32, 143]. They also show that FastText-LID outperforms previously introduced LID tools such as langid.[13] We use FastText-LID as a light, fast, and reasonably accurate language detection tool to overcome the challenge of missing language labels in our Twitter historical feed.

We group tweets by day according to Eastern Time (ET). Date and language are the only metadata we incorporate into our database. For user privacy in particular, we discard all other information associated with a tweet.

3.A.2 Social amplification and contagion

Twitter enables social amplification on the platform through the use of retweets and, from 2015 on, quote tweets. Users—including the general public, celebrities, scientists, decision-makers, and social bots [148]—can intervene in the information spread process and amplify the volume of any content being shared. We categorize tweets into two major classes: organic tweets (OT) and retweets (RT). Organic tweets represent the set of new information being shared on the platform, whereas retweets reflect information being socially amplified on Twitter. During our process, consistent with Twitter's original encoding of retweets, we enrich the text body of a retweet with (RT @userHandle: ...) to indicate the original user of the retweeted text. Our categoriza-

[13] https://fasttext.cc/blog/2017/10/02/blog-post.html

tion enables users of the Storywrangler data set to tune the amplification processes of the rich-get-richer mechanism [236, 265] by dialing the ratio of retweets added to the n-grams corpus.

3.A.3 Detailed dataset statistics

From 2008-09-09 on, we have been collecting a random subset of approximately 10% of all public messages using Twitter's Decahose API.[14] Every day, half a billion messages are shared on Twitter in hundreds of languages. By the end 2020, our data collection comprised around 150 billion messages, requiring over 100TB of storage.

It is worth noting again that this is an approximate daily leaderboard of language usage and word popularity. It is well established that n-gram frequency-of-usage (or Zipf) distributions are heavy-tailed [322]. Researchers have thoroughly investigated ways to study Zipf distributions and estimate the robustness and stability of their tails [33, 120, 229, 234]. Investigators have also examined various aspects of the Twitter's Sample API [227], and how that may affect the observed daily word distributions [303].

Our Twitter corpus contains an average of 23 million unique 1-grams every day with a maximum of a little over 36 million unique 1-grams captured on 2013-08-07. The numbers of unique bigrams and trigrams strongly outweigh the number of unique unigrams because of the combinatorial properties of language. On average, we extract around 76 million unique 2-grams and 128 million unique 3-grams for each day. On 2013-08-07, we recorded a high of 121 million unique 2-grams, and a high of 212

[14] https://developer.twitter.com/en/docs/twitter-ads-api/campaign-management/api-reference

Table 3.A.1: Average daily summary statistics for 1-grams.

	AT		OT	
	Volume	**Unique**	**Volume**	**Unique**
μ	4.25×10^8	2.27×10^7	1.93×10^8	1.68×10^7
25^{th}	3.87×10^8	2.20×10^7	1.52×10^8	1.43×10^7
50^{th}	4.56×10^8	2.42×10^7	1.78×10^8	1.74×10^7
75^{th}	5.16×10^8	2.67×10^7	2.53×10^8	2.05×10^7
max	1.13×10^9	3.61×10^7	3.83×10^8	2.90×10^7

million unique 3-grams.

We emphasize that these maxima for n-grams reflect only our data set, and not the entirety of Twitter. We are unable to make assertions about the size of Twitter's user base or message volume. Indeed, because we do not have knowledge of Twitter's overall volume (and do not seek to per Twitter's Terms of Service), we deliberately focus on ranks and relative usage rates for n-grams away from the tails of their distributions. Raw frequencies of exceedingly rare words are roughly one-tenth of the true values with regards to all of Twitter, however, rankings are likely to be subject to change.

In Tab. 3.A.1, we show daily summary statistics for 1-grams broken by each category in our data set. We demonstrate the same statistical information for 2-grams and 3-grams in Tab. 3.A.2 and Tab. 3.A.3 respectively. We show a time series of the unique number of n-grams captured daily in Fig. 3.A.1 and the statistical distributions of each category in Fig. 3.A.2.

Table 3.A.2: Average daily summary statistics for 2-grams.

	AT		**OT**	
	Volume	**Unique**	**Volume**	**Unique**
μ	3.98×10^8	7.60×10^7	1.77×10^8	5.41×10^7
25^{th}	3.59×10^8	7.29×10^7	1.41×10^8	4.71×10^7
50^{th}	4.23×10^8	7.90×10^7	1.63×10^8	5.24×10^7
75^{th}	4.83×10^8	8.79×10^7	2.33×10^8	6.56×10^7
max	1.09×10^9	1.21×10^8	3.51×10^8	9.34×10^7

Table 3.A.3: Average daily summary statistics for 3-grams.

	AT		**OT**	
	Volume	**Unique**	**Volume**	**Unique**
μ	3.66×10^8	1.28×10^8	1.62×10^8	9.04×10^7
25^{th}	3.30×10^8	1.17×10^8	1.30×10^8	7.66×10^7
50^{th}	3.88×10^8	1.31×10^8	1.50×10^8	8.65×10^7
75^{th}	4.42×10^8	1.52×10^8	2.13×10^8	1.12×10^8
max	1.03×10^9	2.12×10^8	3.19×10^8	1.61×10^8

3.A.4 Twitter n-grams

For the initial version of Storywrangler, we have extracted n-grams from tweets where $n \in \{1, 2, 3\}$. We record raw n-gram frequency (or count) at the day scale for each language (including unidentified), and for Twitter as a whole.

Although some older text tokenization toolkits followed different criteria (e.g., Google books [197]), we had to design a custom n-gram tokenizer to preserve all Unicode characters. In particular, to accommodate the rich lexicon of social media data, we have to preserve handles, hashtags, tickers, and emojis, which are irrelevant for books, but rather integral to social media platforms such as Twitter. We try to maintain the structures in each message as fully as possible, giving the user the option of filtering out our daily Zipf distributions in ways that would fulfill their needs. We display a screenshot of our regular expression pattern matching in Fig. 3.A.3. Our source code along with our documentation is publicly available online on a Gitlab repository.[15]

A 1-gram is a continuous string of characters bounded by either whitespace or punctuation marks. For example, the word 'the' is one of the prominent 1-grams in English. The 2-gram 'here?' consists of the 1-grams: 'here' and '?'. Numbers and emojis also count as 1-grams. Similarly, a word bounded by two quotes (e.g., "sentient") would be a 3-gram, and the expression 'see the light' is a 3-gram, and so forth.

We parse currency (e.g., $9.99), floating numbers (e.g., 3.14), and date/time strings (e.g., 2001/9/11, 2018-01-01, 11:59pm) all as 1-grams. We curate links (e.g., https://www.google.com/), handles (e.g., @NASA), stocks/tickers (e.g., $AAPL) and

[15] https://gitlab.com/compstorylab/storywrangler

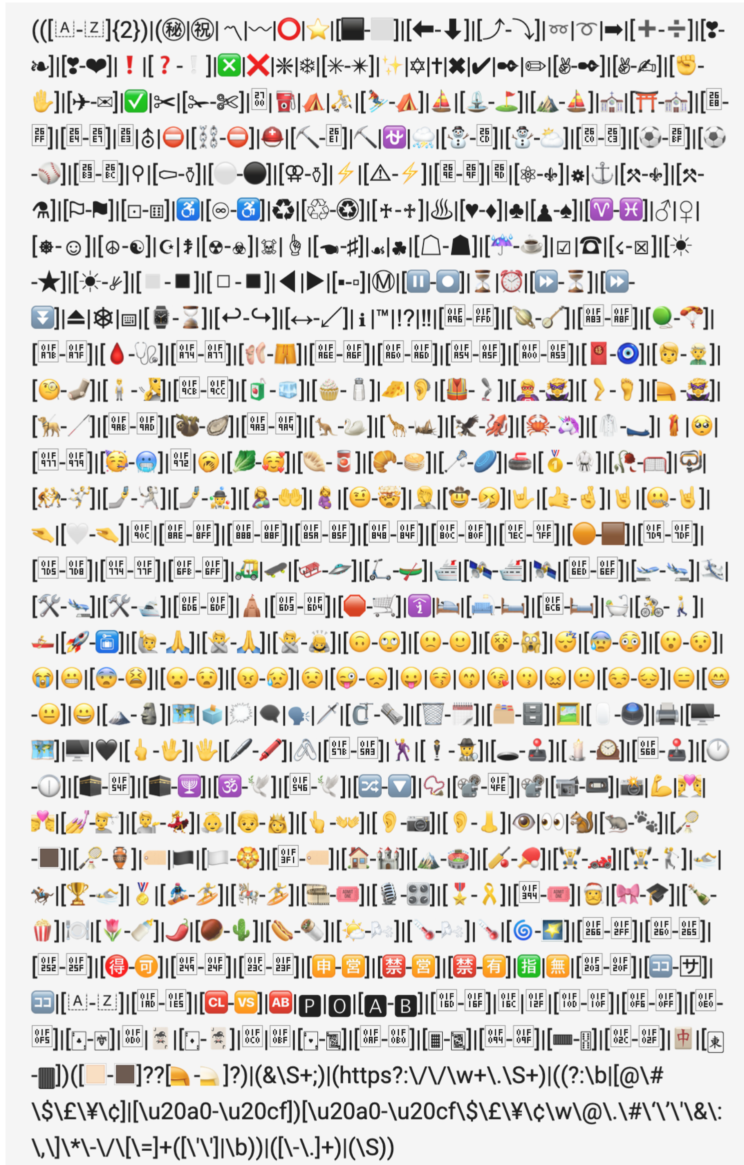

Figure 3.A.3: ***Screenshot of Storywrangler's n-gram regular expression pattern recognition.*** *For our application, we designed a custom n-gram tokenizer to accommodate all Unicode characters. Our n-gram parser is case sensitive. We preserve contractions, handles, hashtags, date/time strings, currency, HTML codes, and links (similar to the Tweet Tokenizer in the NLTK library [180]). We endeavor to combine contractions and acronyms as single objects and parse them out as 1-grams (e.g., 'It's', 'well-organized', and 'B&M'). In addition to text-based n-grams, we track all emojis as 1-grams. While we can identify tweets written in continuous-script-based languages (e.g., Japanese, Chinese, and Thai), our current parser does not support breaking them into n-grams. Although some older text tokenization toolkits followed different criteria, our protocol is consistent with modern computational linguistics for social media data and is adopted among researchers [26, 132].*

hashtags (e.g., #metoo) as 1-grams. We endeavor to combine contractions and acronyms as single objects and parse them out as 1-grams (e.g., 'It's', 'well-organized', and 'B&M').

Emojis are uniquely and interestingly complex entities.[16] People-centric emojis can be composed of skin-tone modifiers, hair-type modifiers, and family structures. Emoji flags are two component objects. The most elaborate emojis are encoded by seven or more unicode elements, rendering them difficult to extract as single entities. After contending with many emoji-parsing problems, we record all emojis as 1-grams. We consider repeated emojis with no intervening whitespace—a common feature in tweets—to be a series of 1-grams.

Our protocol is similar to the Tweet Tokenizer developed as part of the Natural Language Toolkit (NLTK) [180] and is adopted in recent applications of modern computational linguistics for social media [26, 132]. In Fig. 3.A.4, we show contagiograms for 12 example n-grams that involve punctuation, numbers, handles, hashtags, and emojis. A few more examples across various languages can be seen in Fig. 3.A.5. We provide a Python package for generating arbitrary contagiograms along with further examples at `https://gitlab.com/compstorylab/contagiograms`. The figure-making scripts interact directly with the Storywrangler database, and offer a range of configurations.

Our n-gram parser is case sensitive. For example, search queries made on storywrangling.org for 'New York City' and a search for 'new york city' would return different results. Normalized frequencies and rankings in our daily Zipf distributions are consequently for case-sensitive n-grams.

[16] `https://www.unicode.org/Public/emoji/12.0/emoji-data.txt`

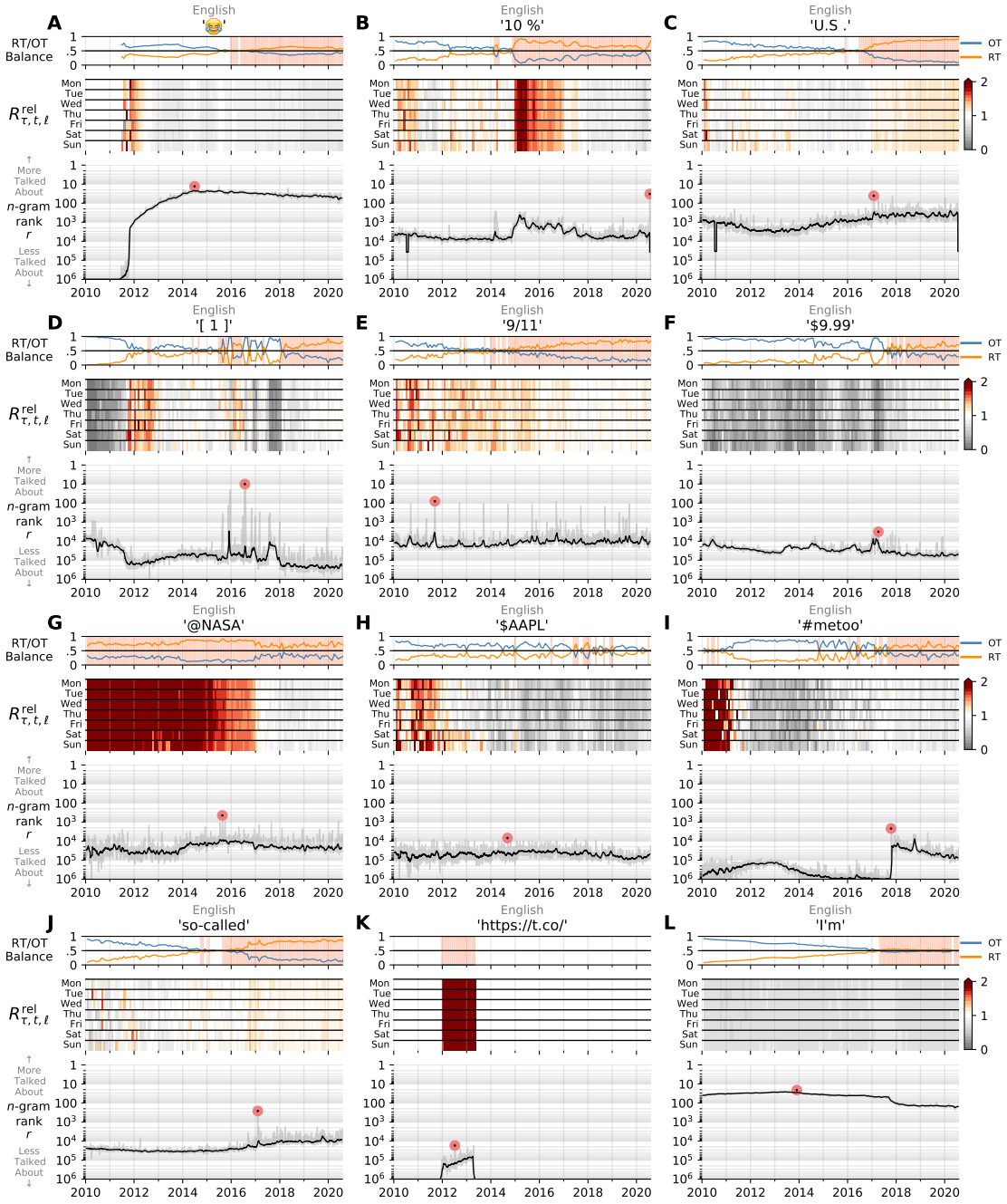

Figure 3.A.4: ***Contagiograms.*** *Example timeseries showing social amplification for Twitter n-grams involving emojis, punctuation, numerals, and so on.*

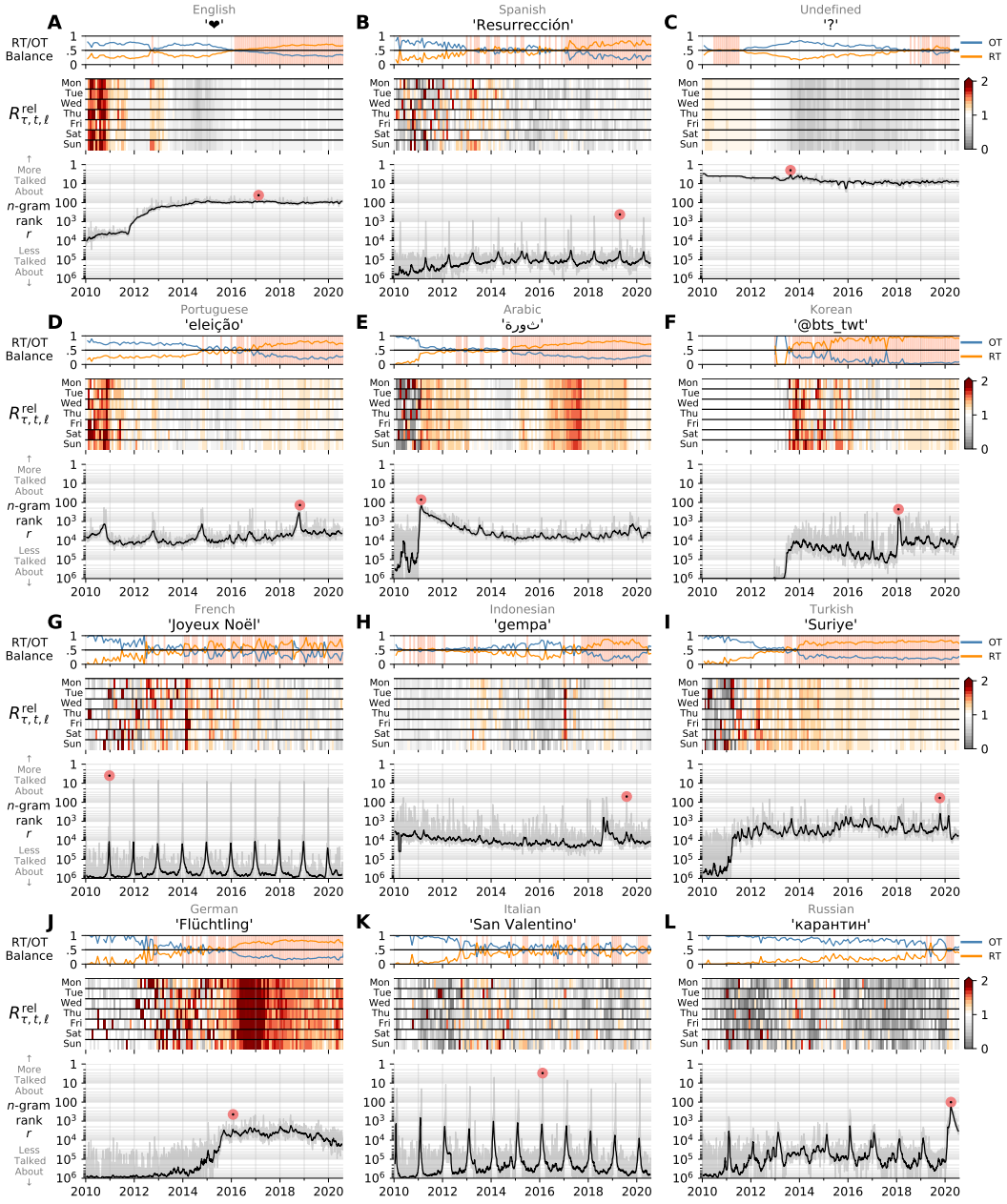

Figure 3.A.5: ***The interplay of social amplification across various languages.*** *We observe a wide range of sociotechnical dynamics starting with n-grams that are often mentioned within OTs and RTs equivalently to others that spread out after a geopolitical event and more extreme regimes whereby some n-grams are consistently amplified. English translations of n-grams:* ***A.*** *Heart emoji,* ***B.*** *'Resurrection',* ***C.*** *Question mark,* ***D.*** *'election',* ***E.*** *'revolution',* ***F.*** *Official handle for the South Korean boy band 'BTS',* ***G.*** *'Merry Christmas',* ***H.*** *'earthquake',* ***I.*** *'Syria',* ***J.*** *'Refugee',* ***K.*** *'Saint Valentine', and* ***L.*** *'quarantine'.*

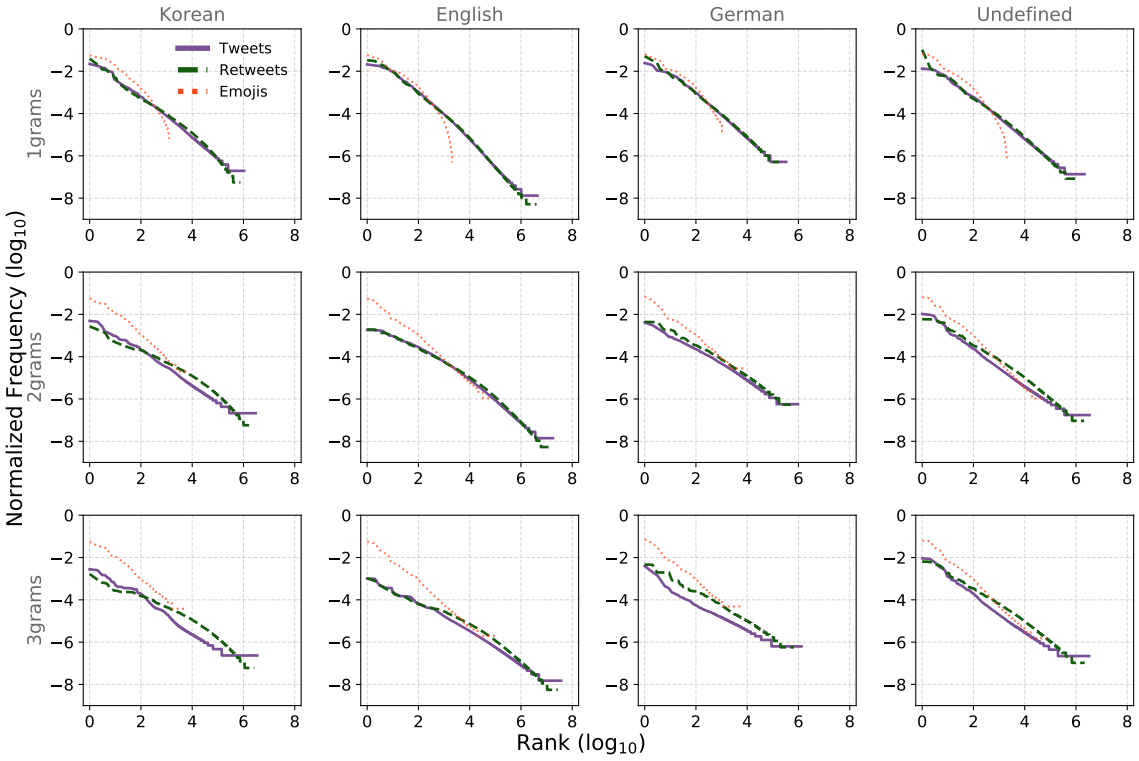

Figure 3.A.6: ***Zipf distributions for Korean, English, German and undefined language categories for October 16, 2019 on Twitter.*** *"Tweets" refer to organic content, "retweets" retweeted content, and "emojis" are n-grams comprised of strictly emojis (organic and retweets combined).*

Although we can identify tweets written in continuous-script-based languages (e.g., Japanese, Chinese, and Thai), our current parser does not support breaking them into n-grams. We label tweets as Undefined (und) to indicate tweets that we could not classify their language with a confidence score above 25%. The resulting n-grams are allocated to an "Undefined" category as well as to the overall Twitter n-gram data set.

To enable access to our dataset, we maintain a MongoDB database of the top million ranked n-grams on each day for each language. We index these collections by date, to allow efficient queries for all n-grams on a given day, as well as by n-gram, which allows for rapid time series queries. Data is typically inserted within two days (i.e., counts from Monday will be available by midnight Wednesday).

3.A.5 Constructing daily Zipf distributions

For ease of usability, we maintain two sets of daily measurements for each n-gram in our data set: raw frequency (count), normalized frequency (probability), and tied rank with and without retweets. We make the default ordering for the Zipf distribution files according to usage levels of n-grams for all of a given language on Twitter (i.e., including all retweets and quote tweets). Again, all daily distributions are made according to Eastern Time calendar days.

We denote an n-gram by τ and a day's lexicon for language ℓ—the set of distinct n-grams found in all tweets (AT) for a given date t—by $\mathcal{D}_{t,\ell;n}$. We further define the set of unique language ℓ n-grams found in organic tweets as $\mathcal{D}^{(\mathrm{OT})}_{t,\ell;n}$, and the set of

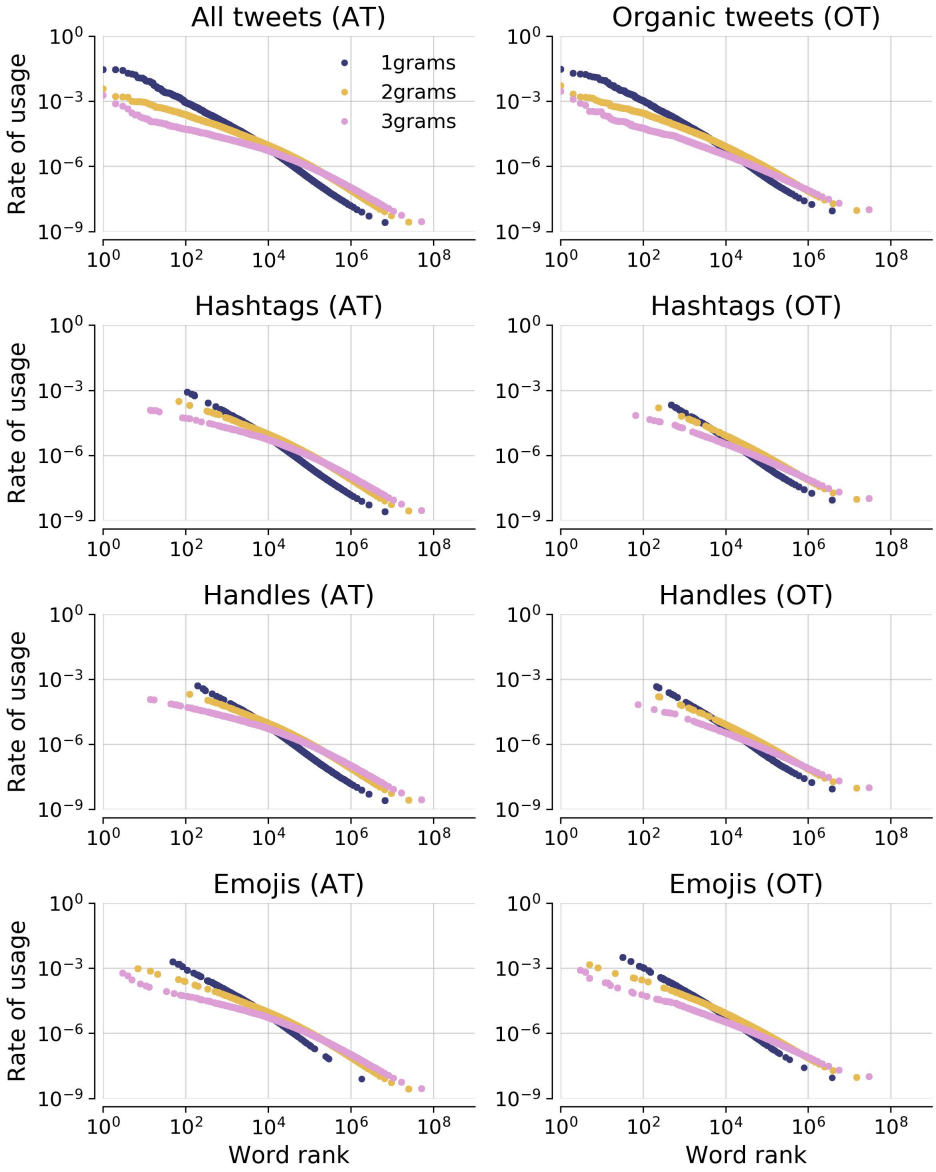

Figure 3.A.7: ***Daily Zipf distributions for English on May 1st, 2020.*** *We show a weighted 1% random sample of* 1*-grams (blue),* 2*-grams (yellow), and* 3*-grams (pink) in all tweets (AT) and organic tweets (OT) accordingly. On the vertical axis, we plot the relative rate of usage of each n-gram in our random sample whereas the horizontal axis displays the rank of that n-gram in the English corpus of that day. We first display Zipf distributions for all n-grams observed in our sample in the first row. We also demonstrate the equivalent distributions for hashtags (second row), handles (third row), and emojis (last row).*

unique n-grams found in retweets as $\mathcal{D}^{(\mathrm{RT})}_{t,\ell;n}$, such that

$$\mathcal{D}_{t,\ell;n} = \mathcal{D}^{(\mathrm{OT})}_{t,\ell;n} \cup \mathcal{D}^{(\mathrm{RT})}_{t,\ell;n}. \tag{3.8}$$

We compute relative daily rate of usage by dividing total number of occurrences of a given n-gram by the total number of n-grams for that day. We write n-gram raw frequency as $f_{\tau,t,\ell}$, and compute its usage rate in all tweets written in language ℓ as

$$p_{\tau,t,\ell} = \frac{f_{\tau,t,\ell}}{\sum_{\tau' \in \mathcal{D}_{t,\ell;n}} f_{\tau',t,\ell}}. \tag{3.9}$$

The corresponding normalized frequencies for n-grams in organic tweets and retweets are then defined as

$$p^{(\mathrm{OT})}_{\tau,t,\ell} = \frac{f^{(\mathrm{OT})}_{\tau,t,\ell}}{\sum_{\tau' \in \mathcal{D}^{(\mathrm{OT})}_{t,\ell;n}} f^{(\mathrm{OT})}_{\tau',t,\ell}}, \text{ and} \tag{3.10}$$

$$p^{(\mathrm{RT})}_{\tau,t,\ell} = \frac{f^{(\mathrm{RT})}_{\tau,t,\ell}}{\sum_{\tau' \in \mathcal{D}^{(\mathrm{RT})}_{t,\ell;n}} f^{(\mathrm{RT})}_{\tau',t,\ell}}. \tag{3.11}$$

We then rank all n-grams for a given day to create daily Zipf distribution [322] for all languages in our data set. If two or more distinct n-grams have the same number of instances (raw frequency), then we adjust their ranks by taking the average rank (i.e., tied-rank). The corresponding notation is:

$$r_{\tau,t,\ell},\ r^{(\mathrm{OT})}_{\tau,t,\ell},\ \text{and}\ r^{(\mathrm{RT})}_{\tau,t,\ell}. \tag{3.12}$$

We do not mix n-grams for different values of n, and leave this as an important future upgrade [303]. Users of the viewer storywrangling.org, will need to keep this

in mind when considering time series of n-grams for different n. In comparing, for example, 'NYC' (a 1-gram) to 'New York City' (a 3-gram), the shapes of the curves can be meaningfully compared while the ranks (or raw frequencies) of the 1-gram and 3-gram may not be.

We show complementary cumulative distribution functions (CCDFs) of organic tweets, retweets, and emojis for 1-, 2-, and 3-grams in Fig. 3.A.6 and Fig. 3.A.7.

3.B Narratively trending n-grams

In addition to curating daily Zipf distributions, we also examine lexical turbulence and emerging storylines in real-time. We do so using rank-turbulence divergence (RTD) [83], an instrument designed to examine the lexical turbulence of two heavy-tailed distributions.

For each day t and each language ℓ, we take the Zipf distribution $\Omega_{t,\ell}$ and compare it with the corresponding distribution of that day from the same day one year prior $\Omega_{t',\ell}$, identifying n-grams that have become most elevated in relative usage. We set the parameter α to $1/4$, which provides a reasonable fit for the lexical turbulence of social media data as recommended by Dodds et al. [83]. We compute rank divergence contribution such that:

$$\begin{aligned} D_{\alpha}^{\mathrm{R}}(\Omega_{t,\ell}||\Omega_{t',\ell}) &= \sum \delta D_{\tau,\ell}^{\mathrm{R}} \\ &= \frac{\alpha+1}{\alpha} \sum_{\tau} \left| \frac{1}{r_{\tau,t,\ell}^{\alpha}} - \frac{1}{r_{\tau,t',\ell}^{\alpha}} \right|^{1/(\alpha+1)}, \end{aligned} \tag{3.13}$$

where $r_{\tau,t,\ell}$ is the rank of n-gram τ on day t, and $r_{\tau,t',\ell}$ is the rank of the same n-gram on day t' (52 weeks earlier). Although we use rank-turbulence divergence to determine shifts in relative word rankings, other divergence metrics will provide similar lists.

We show an example Allotaxonograph using rank-turbulence divergence for English word usage on 2021-01-06 compared to 2020-01-06 (see Fig. 3.B.1). The main plot shows a 2D histogram of n-gram ranks on each day. Words located near the center vertical line are used equivalently on both days, while words on either side highlight higher usage of the given n-gram on either date.

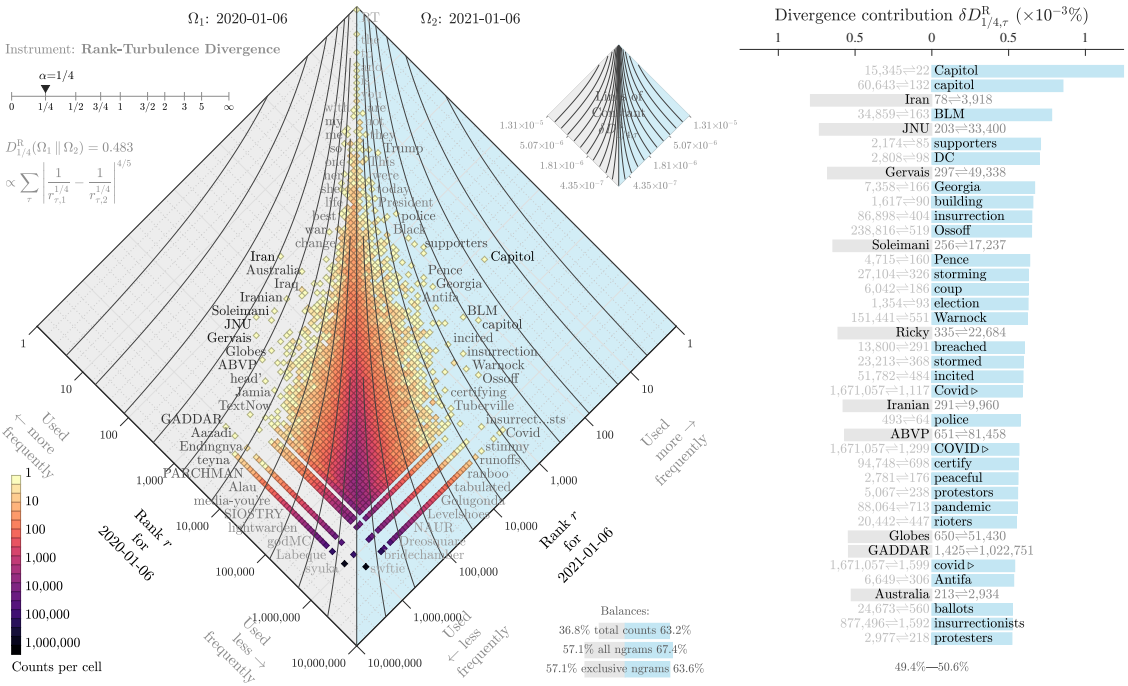

Figure 3.B.1: ***Allotaxonograph using rank-turbulence divergence for English word usage on 2021-01-06 compared to 2020-01-06****. The word 'Capitol' was the 22nd most common* 1*-gram on January 6, 2021, up from 15,345th most common one year earlier. Similarly, 'COVID' was the 1,117th most popular word January 6, 2021, and did not make the top million on January 6, 2020. See Dodds et al. [83] for further details on the allotaxonometric instrument.*

The right side of the plot displays the most narratively dominant n-grams ordered by their rank divergence contribution. For ease of plotting, we use the subset of words containing Latin characters only. Notably, words associated with the storming of the US Capitol by Trump supporters dominate the contributions from 2021-01-06 such as 'Capitol', 'supporters', 'DC', and 'insurrection'.

In Fig. 3.B.2 and Fig. 3.B.3, we show example analyses of English tweets highlighting the top 20 narratively trending 1-grams and 2-grams, respectively.

First, we compute RTD values to find the most narratively dominate n-grams for a few days of interest throughout the last decade. We filter out links, emojis, handles, and stop words, but keep hashtags to focus on the relevant linguistic terms on these days.

Second, we further compute the relative social amplification ratio $R^{\text{rel}}_{\tau,t,\ell}$ to measure whether a given n-gram τ is socially amplified, or rather shared organically in originally authored tweets.

For ease of plotting, we display $R^{\text{rel}}_{\tau,t,\ell}$ on a logarithmic scale. Positive values of $\log_{10} R^{\text{rel}}_{\tau,t,\ell}$ imply strong social amplification of τ, whereas negative values show that τ is rather predominant in organic tweets.

Figs. 3.B.2A and 3.B.3A show the top terms used to discuss the earthquake off the Pacific coast of Tokyo leading to a sequence of massive tsunami waves on 2011-03-11.[17] Although most terms are socially amplified, referring to the catastrophic event in Japan, other cultural references can be found in organic tweets such as '#np' (i.e., no problem), and '#FF' (i.e., follow Friday) where users recommend accounts to follow.

[17] https://www.britannica.com/event/Japan-earthquake-and-tsunami-of-2011

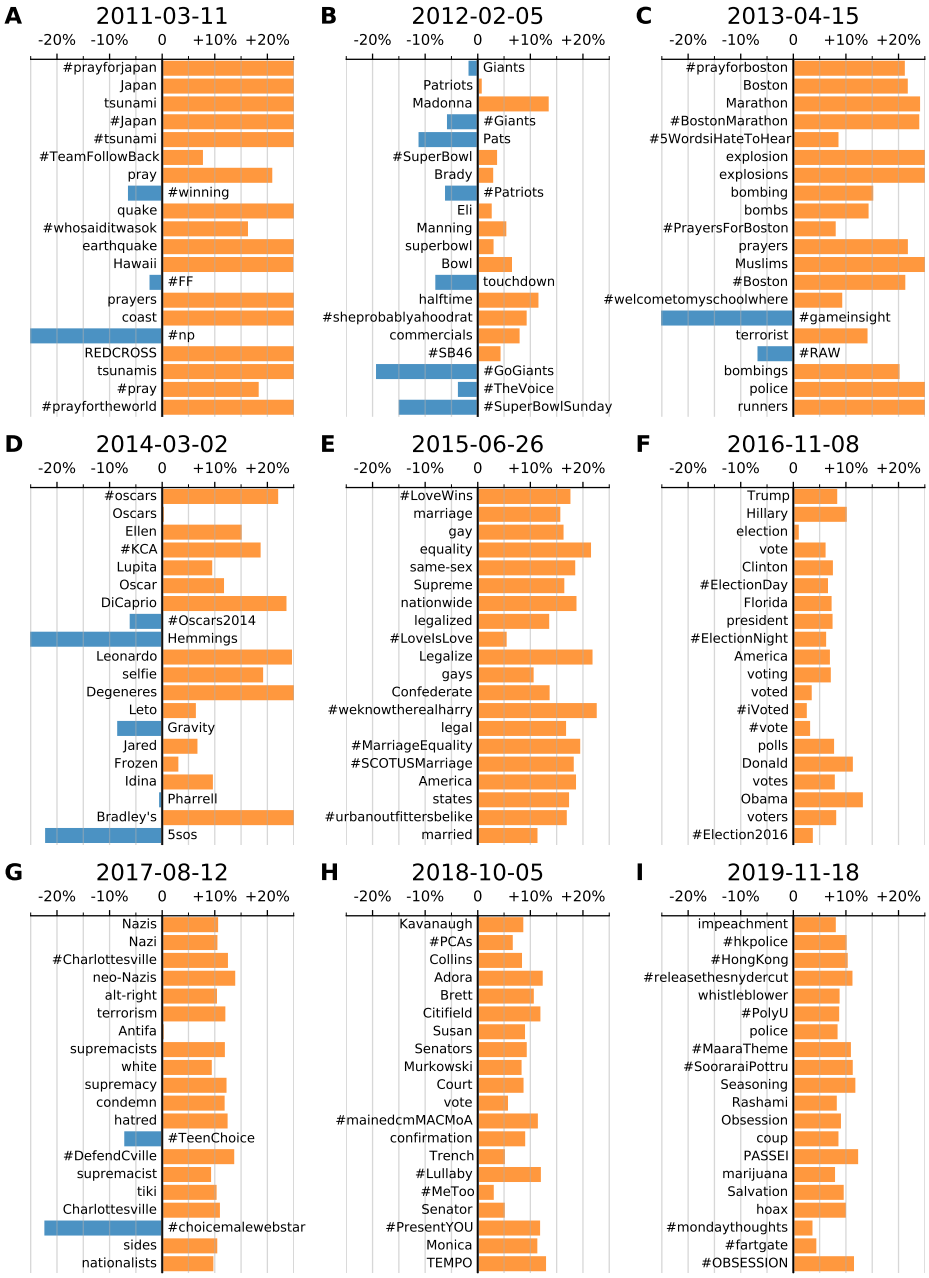

Figure 3.B.2: ***Narratively trending 1-grams****. Top 20 narratively dominate 1-grams for a few days of interest throughout the last decade (sorted by their rank-turbulence divergence contribution). Positive values (orange) indicate strong social amplification via retweets, whereas negative values (blue) show terms that are prevalent in originally authored tweets. See Supplementary text for details on each date.*

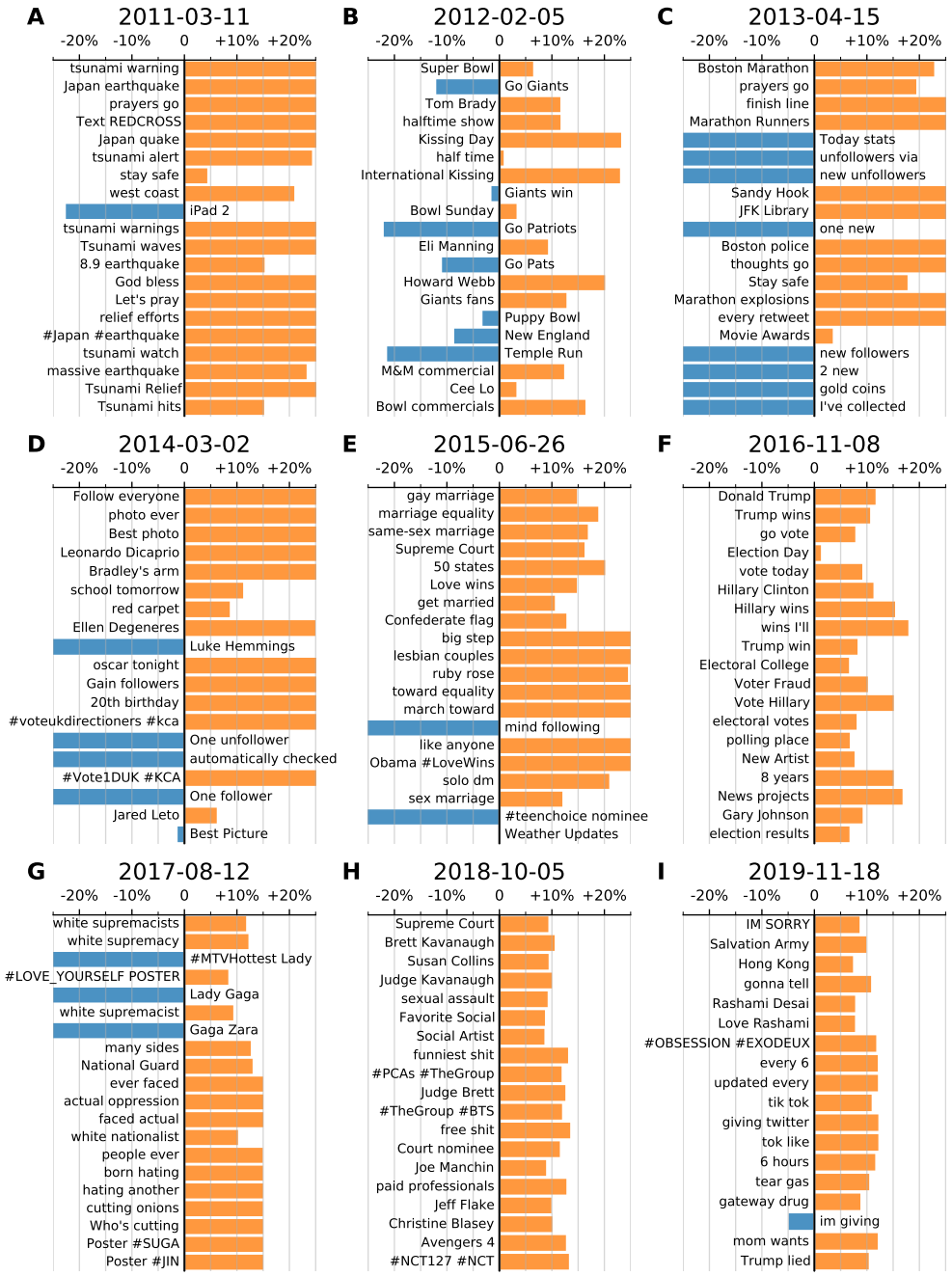

Figure 3.B.3: ***Narratively trending 2-grams****. Top 20 narratively dominate 2-grams for the same days shown in Fig. 3.B.2. See Supplementary text for details on each date.*

In Figs. 3.B.2B and 3.B.3B, we see trending terms discussing the Super Bowl held on 2012-02-05. Figs. 3.B.2C and 3.B.3C show salient n-grams in response to the terrorist attack during the annual Boston Marathon on 2013-04-15.[18] We also observe unusual bigrams in organic tweets generated by artificial Twitter bots for a new automated service that went viral in 2013, allowing users to keep track of their new followers and unfollowers by tweeting their stats daily.[19]

Figs. 3.B.2D and 3.B.3D show names of celebrities and movie titles that went viral on Twitter during the 86$^{\text{th}}$ Academy Awards hosted by Ellen DeGeneres on 2014-03-02.[20] Most socially amplified terms can be traced back to a single message with more than 3 million retweets. We do, however, see some n-grams trending in organic tweets such as '5sos', referring to an Australian pop rock band that was slowly taking off that year.

On 2015-06-26, the US Supreme Court ruled same-sex marriage is a legal right in the US[21], prompting a wave of reactions on social media as seen in Figs. 3.B.2E and 3.B.3E. The 2016 US presidential election had a similar chain of reactions on Twitter. In Figs. 3.B.2F and 3.B.3F, we see names of candidates and politicians across the aisle being amplified collectively on the platform.

As we have seen throughout the paper, cultural movements and social media are profoundly integrated. In Figs. 3.B.2G and 3.B.3G, we see the top 20 narratively trending terms in response to the deadly white supremacist rally that took place on 2017-08-12 in Charlottesville, Virginia.[22] In particular, we notice several bigrams

[18] https://www.history.com/topics/21st-century/boston-marathon-bombings
[19] https://who.unfollowed.me/
[20] https://en.wikipedia.org/wiki/86th_Academy_Awards
[21] https://www.usatoday.com/story/news/politics/2020/06/25/lgbtq-rights-five-years-after-gay-marriage-ruling-battles-continue/3242992001/
[22] https://time.com/charlottesville-white-nationalist-rally-clashes/

from Obama's tweet that went viral on that day, quoting Nelson Mandela's remarks on racism.

In Figs. 3.B.2H and 3.B.3H, we see some headlines and narratively amplified n-grams, in light of Kavanaugh's testimony before the Senate Judiciary Committee for his nomination to the US Supreme Court 2018-10-05.

Figs. 3.B.2I and 3.B.3I display the top n-grams on 2019-11-18 whereby we see several terms and hashtags amplified on Twitter triggered by the Siege of the Hong Kong Polytechnic University amid the nationwide protests in Hong Kong.[23] Moreover, we also see notable references to the first impeachment of Trump that took place between September 2019 and February 2020.

[23] https://www.theguardian.com/world/2019/nov/18/hong-kong-university-siege-a-visual-guide

3.C Pantheon case study

We examine the dialog around celebrities by cross-referencing our English 2-grams corpus with names of famous personalities from the Pantheon data set [317]. The data set has over 10 thousand biographies. We use the place and date of birth to select Americans born in the last century.

We searched through our English n-grams data set and selected names that were found in the top million ranked 2-grams for at least a day between 2010-01-01 and 2020-06-01. Our list contains 1010 individuals. We show the average best rank $\bar{r}_{\min}$, median best rank $\tilde{r}_{\min}$, and best rank $r^*_{\min}$ for all individuals in each occupation in Tab. 3.C.1. In Figs. 3.C.1A and B, we display a monthly rolling average (centered) of the average rank for the top 5 individuals for each category $\langle r_{\min(5)} \rangle$.

Table 3.C.1: Celebrities by occupation

Occupation	n	$\bar{r}_{\min}$	$\tilde{r}_{\min}$	$r^*_{\min}$
Actors	674	40,632	9,255	2
Singers	162	59,713	3,479	4
Politicians	59	6,365	1,376	6
Film directors	57	75,783	10,580	13
Business-persons	26	35,737	4195	15
Soccer players	12	20,868	8,507	25
Social activists	10	20,302	1,781	841
Extremists	10	104,621	20,129	117

Additionally, we select a total of 1162 celebrities that were also found in the top million ranked 2-grams for at least a day between 2010-01-01 and 2020-06-01 in a few

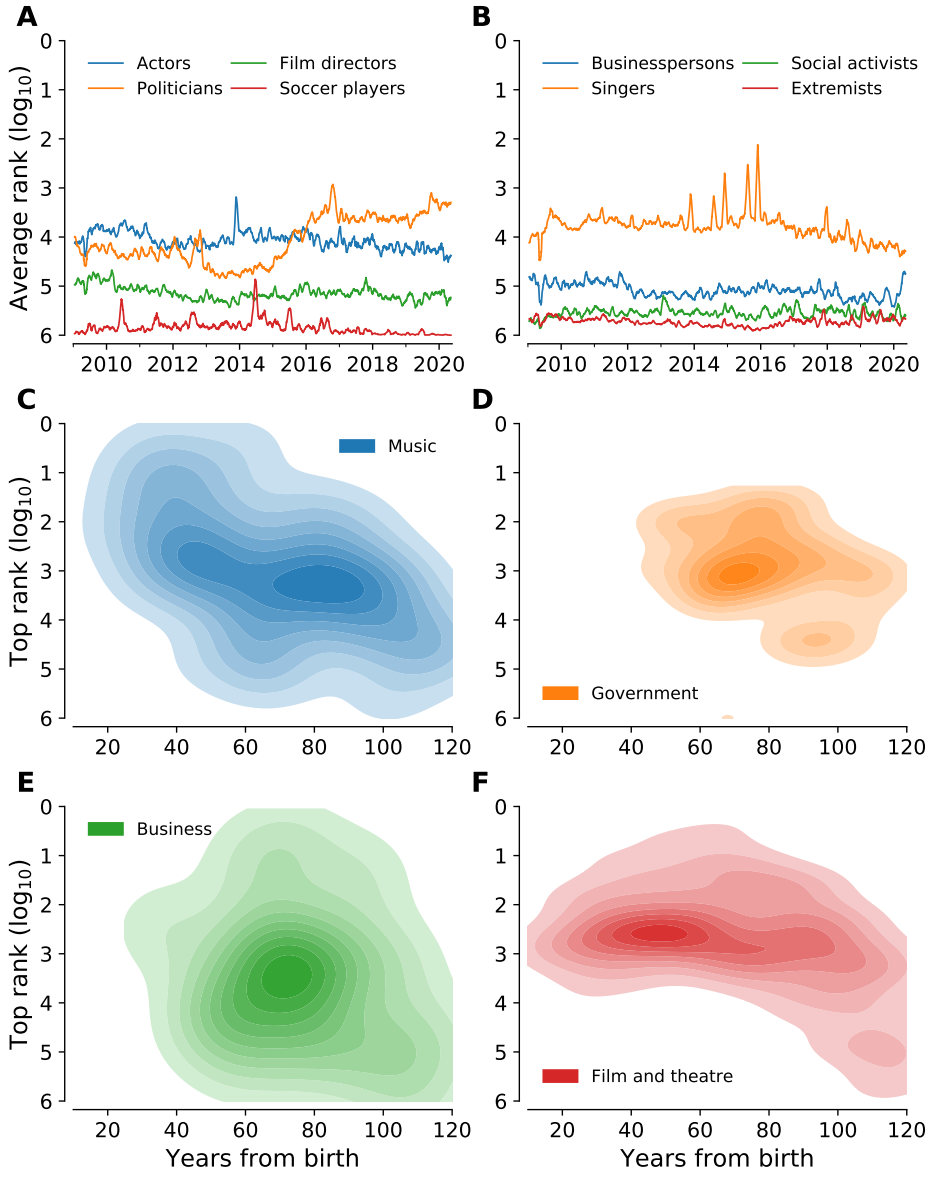

Figure 3.C.1: ***Rankings of celebrities on Twitter****. We take a closer look at rankings of famous figures by cross-referencing our English corpus with names of celebrities from the Pantheon dataset [317]. We use their first and last name to search through our* 2*-grams data set. We select names of Americans who were born in the last century and can be found in the top million ranked* 2*-grams for at least a day between 2010-01-01 and 2020-06-01. In panels* ***A*** *and* ***B****, we display a centered monthly rolling average of the average rank for the top 5 individuals for each category* $\langle r_{\min(5)} \rangle$*. We also plot the kernel density estimation of the best rank achieved by another 1162 famous characters in each of the following industries:* ***C.*** *music,* ***D.*** *government,* ***E.*** *business, and* ***F.*** *film.*

selected industries (see Tab. 3.C.2)

Table 3.C.2: Celebrities filtered through Pantheon and Twitter rank, by industry

Industry	Individuals
Film and theater	751
Music	324
Government	59
Business	28

For each of these individuals, we track their age and top daily rank of their names (first and last). In Figs. 3.C.1A, B, C, and D, we display kernel density estimation of the top rank achieved by any of these individuals in each industry as a function of the number of years since the recorded year of birth (age of the cohort).

3.D Movies case study

We investigate the conversation surrounding major film releases by tracking n-grams that appear in movie titles. From the MovieLens dataset [117], we selected 636 movies with gross revenue above the 95th percentile during the period ranging from 2010-01 to 2017-07.

We then retrieved the normalized frequency time series for up to the first 3-grams of a movie's title (e.g., "Prince of Persia: The Sands of Time" would correspond to the 3-gram time series "Prince of Persia").

From there, we look for the maximum daily normalized frequency. To disambiguate between movies within the same franchise and/or titles with common n-grams, we restrain this search to the release year of the given movie. With the peak usage in the year of a movie's release, we then search backward for the date on which the n-gram usage first breaks 50% of the peak usage normalized frequency, $f_{.5}$. Similarly we search forward, from peak usage, for the date on which the time series first declines below $f_{.5}$.

Peak conversation surrounding major movies tends to occur a few days after the release data of the title. We find a median value of 3 days post-release for peak normalized frequency of usage for movie n-grams (Fig. 3.4.5F inset). Growth of n-gram usage from 50% to maximum normalized frequency has a median value of 5 days across our 636 titles.

The median value of time to return to 50% from maximum normalized frequency is 6 days. Looking at Fig. 3.4.5E we see the median shape of the spike around movie release dates tend to entail a gradual increase to peak usage, and a more sudden

decrease when returning to 50% of maximum normalized frequency. There is also slightly more spread in the time to return to 50% normalized frequency of usage than compared with the time to increase from 50% to maximum usage (Fig. 3.4.5E insets).

3.E Geopolitical risk case study

Twitter sentiment has already been shown to provide a useful signal in monitoring public opinion [134, 316]. The aggregation process in which individual tweet documents are turned into popularity time series reduces the time and computation power required to use this data in models of political sentiment and public opinion. As a case study, we sought to predict values of a geopolitical risk (GPR) index using popularity of words that we heuristically associated with (inter)national unrest and popular discontent. We conduct both an exploratory Bayesian analysis and a more rigorous frequentist analysis of the relationship between word time series and the GPR index.

3.E.1 Data description

The geopolitical risk index is developed by the U.S. Federal Reserve (central bank) [44]. We chose the words "revolution", "rebellion", "uprising", "coup", "overthrow", "unrest", "protests", and "crackdown" to include as predictors. We chose these words arbitrarily due to denotations and connotations of conflict associated with their common English meaning. Since we could not reject the null hypothesis that at least one of the logarithm of normalized frequency time series associated with these words contained a unit root (ADF("overthrow") $= -1.61$, $p = 0.474$), we computed the difference of the logarithm of normalized frequency time series and used these observations as features. We could also not reject the null hypothesis that the geopolitical risk time series contained a unit root (ADF $= -0.65$, $p = 0.858$) and therefore sought to predict the log difference of the GPR. Because GPR is computed at monthly frequency, we resample normalized word frequencies to monthly normalized frequency

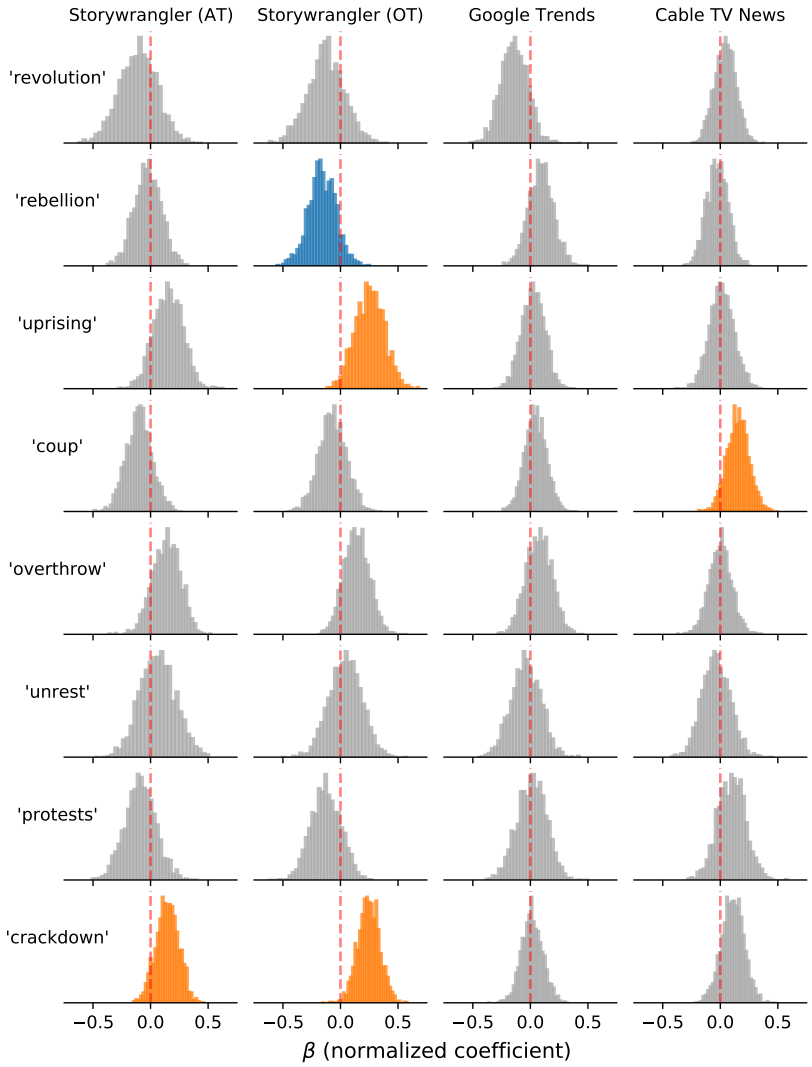

Figure 3.E.1: ***Empirical distributions of the β coefficient of each word.*** *Percent change in word popularity is significantly associated with percent change of future geopolitical risk (GPR) index level for a few words out of a panel of eight words: "revolution", "rebellion", "uprising", "coup", "overthrow", "unrest", "crackdown", and "protests". We assess significance of model coefficients using centered 80% credible intervals (CI). The sign of the coefficient differs between the words, with positive associations shown in orange and negative associations shown in blue. Using Storywrangler, we note the words "crackdown", and "uprising" are positively associated with GPR, whereas "rebellion" is negatively associated. We see some overlap between Storywrangler and other data streams. Percent change of the word "rebellion" is also negatively associated with GPR, using Google trends search data [54], but not statistically significant. By contrast, mentions of the word "coup" in cable news is positively associated with GPR using the Stanford cable TV news analyzer [132].*

by taking the average of the lagged month's values. For example, the normalized frequency of the word "crackdown" sampled at month level timestamped at 2010-03-31 is taken to be the average daily normalized frequency of the word "crackdown" from 2010-03-01 through 2010-03-31.

We compute the monthly normalized frequencies for each word in all tweets (AT), and originally authored tweets (OT), separately, for the last decade starting from 2010 to the end of 2019. We also collect monthly n-gram usage time series for the same set of words from two other data sources—namely, Google trends [54], and the Stanford cable TV news analyzer [132] to examine the utility of the predictors derived from Storywrangler compared to existing data sources. We conduct out of sample analyses on word frequency and GPR data from 2020-01-01 through 2020-05-01, which is the last day on which the GPR data was publicly accessible.

3.E.2 Exploratory analysis

We fit a linear model for each data source that hypothesized a linear relationship between the log difference in normalized word frequencies for each of the words listed in the previous paragraph and the log difference in GPR. The likelihood function of this model took the form

$$p(y|\beta,\sigma) = \prod_{t=0}^{T-1} \text{Normal}(y_{t+1}|X_t\beta,\sigma^2). \tag{3.14}$$

We denote $y \equiv (y_1,...,y_T)$ and $X = (X_0,...,X_{T-1})$. Each y_t is the difference of log GPR measured on the first day of month t, while each X_{t-1} is the p-dimensional row vector of difference in log word frequencies averaged over the entirety of month

$t-1$. Thus, y is a T-dimensional column vector and X is a $T \times p$-dimensional matrix. Note that this model design respects temporal causality. We regularize the model toward the null hypothesis of no relationship between X and y by placing a zero-mean normal prior on β, the p-dimensional column vector of coefficients, as $\beta \sim \text{MultivariateNormal}(0, I)$. We place a weakly informative prior on the noise scale, $\sigma \sim \text{LogNormal}(0, 1)$, as we are *a priori* unsure of what noise level the data would exhibit.

We sample from the model using the No U-Turn Sampler (NUTS) algorithm [127] for 500 warmup iterations and 2000 iterations of sampling. There is strong evidence to suggest that the sampler converged since the maximum Gelman-Rubin statistic [104] for all priors was less than 1.01 ($\max \hat{R} = 1.0009$).

We compute centered 80% credible intervals for each of the model coefficients β_k, $k = 0, ..., p$. (A centered Q% credible interval for the univariate random variable $Y \sim p(y)$ is an interval (a, b) such that $\frac{1}{2}(1 - \frac{Q}{100}) = \int_{-\infty}^{a} dy\ p(y) = \int_{b}^{\infty} dy\ p(y)$. For example, a centered 80% credible interval is an interval such that $0.1 = \int_{-\infty}^{a} dy\ p(y) = \int_{b}^{\infty} dy\ p(y)$.) We termed a relationship significant if the 80% credible interval did not contain zero.

In Fig. 3.E.1, we display the empirical distributions of the β coefficient of each word for each model. The sign of the coefficient differs between the words, with positive significant associations highlighted in orange and negative significant associations highlighted in blue.

Using the predictors derived from Storywrangler, we recognize that the log difference in normalized usage frequency of "rebellion" is negatively associated with future log difference of GPR in originally authored tweets (OT). Similarly, "uprising" is

positively associated with the percent change of GPR in organic tweets (OT), but not statistically significant in all tweets (AT). The word "crackdown" is positively associated with GPR in both organic tweets (OT), and all tweets (AT).

We speculate that increases in the usage of "crackdown" may imply that a popular revolt is already in the process of being crushed, a realization of increased geopolitical uncertainty. Conversely, usage of "uprising" could be driven by a growing collective attention of a newborn social of movement with increased tension and uncertainty, while usage of 'rebellion' might be referring to past events [53]. Although our results using all tweets and organic tweets are fairly similar, future work can further investigate how social amplifications via retweets can influence our perception of geopolitical risk uncertainty.

Furthermore, the log difference in normalized usage frequency of "revolution" and "rebellion" in Google trends search data [54] are associated with future log difference of the GPR index level but not statistically significant. We believe that is likely associated with an increased number of searches for an ongoing revolt. We observe a similar signal for data derived from the Stanford cable TV news analyzer [132], whereby growing mentions of "coup" in news outlets can be linked to higher levels of uncertainty in GPR.

Although we see some overlap across the data streams examined here, Storywrangler provides a unique signal derived from everyday conversations on Twitter that is sometimes similarly portrayed across platforms, but often orthogonal and complementary to existing data sources. We foresee stronger potential for future work cross-referencing Storywrangler with other data repositories to enrich current data sources (e.g., search data and news).

3.E.3 Qualitative comparison

We conduct an additional analysis to differentiate between intra-GPR dynamics and potential predictive power of difference in log word frequency on log difference in GPR. Our notation is identical to that used in Appendix 3.E.2.

We first fit a null autoregressive model, denoted by M_0, that assessed the effect of lagged GPR values on the current GPR value. We fit an autoregressive model of order $p = 4$ to the difference in log GPR data from 2010-01-01 to 2019-12-01, where $p = 4$ was chosen by minimizing Akaike information criterion (AIC). This model takes the form

$$y_t = \theta_0 + \sum_{p=1}^{4} \theta_p y_{t-p} + u_t, \tag{3.15}$$

with $u_t \sim \text{Normal}(0, \sigma^2)$. Here, y_t is the difference of log GPR measured on the first day of month t. We fit the model using conditional MLE. Each of θ_p for $p \in \{1, 2, 3\}$ were significantly different from zero at the $p = 0.05$ confidence level (mean $\pm$ sd) ($\theta_1 = -0.4028 \pm 0.093$, $p < 0.001$; $\theta_2 = -0.3858 \pm 0.098$, $p < 0.001$; $\theta_3 = -0.2650 \pm 0.098$, $p = 0.007$). The model is stable because all roots of the polynomial $P(z) = \sum_{p=1}^{4} \theta_p z^{-p}$ are outside of the unit circle. The model exhibited AIC $= -2.349$. For out-of-sample comparison with other models, we computed the mean square forecast error (MSFE), defined as

$$\text{MSFE}(M) = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} (y_t - \hat{y}_t)^2, \tag{3.16}$$

for $t_1, t_2 \in \mathbb{N}$ and $t_2 - t_1 > 0$, where $\hat{y}$ are the out-of-sample predictions by model M and y are the true out-of-sample data. The null model M_0 had $\text{MSFE}(M_0) = 0.6734$.

We then assessed the additional predictive power of each dataset, using an autoregressive with exogenous regressors (AR-X) model. For this model we used the same number of lags as the null AR model, $p = 4$. This model takes the form

$$y_t = \theta_0 + \sum_{p=1}^{4} \theta_p y_{t-p} + X_{t-1}\beta + u_t, \tag{3.17}$$

where y_t has the same interpretation as in Eq. 3.15, X_{t-1} is the $N \times p$-dimensional of exogenous regressors (Storywrangler OT, Storywrangler AT, cable news, or Google trends) aggregated over month $t-1$, as described in Appendix 3.E.1, and β is the p-dimensional vector of regression coefficients. We fit each model using conditional MLE. Each estimated model was stable as indicated by the roots of the polynomial $P(z) = \sum_{p=1}^{4} \theta_p z^{-p}$ being outside of the unit circle for each model. No regression coefficient of any exogenous regressor for any of {Storywrangler AT, Google trends, News} was significant at the Bonferroni corrected $p_c = p/N = 0.05/4 = 0.0125$ level. However, one exogenous regressor—the difference of log frequency for the word "crackdown"— was significant at the $p_c = 0.0125$ level $\beta_{\text{crackdown}} = 0.0712 \pm 0.028$, $p_c = 0.010$, indicating that an increase in the average difference in log frequency of usage of the word "crackdown" in month $t-1$ is significantly associated with an increase in the difference of log GPR in month t, even after controlling for autoregressive effects. We display summaries of conditional MLE inference results of the null AR model and AR-X model for each data source in Tables 3.E.2—3.E.6.

We then compare the MSFE results of each AR-X model to the null AR model and display the results in Table S6. The AR-X model using Storywrangler OT has the lowest MSFE (0.578), followed by Storywrangler AT (0.650). However, AIC for each model (measured on training model) are not significantly different and do not

differ significantly from AIC for the null AR model.

Table 3.E.1: MSFE results of null AR and AR-X model with each dataset.

	Storywrangler		**Cable**	**Google**	**Null**
	OT	**AT**	**News**	**Trends**	M_0
MSEF	0.578	0.650	0.659	0.679	0.673
R^2	0.288	0.226	0.233	0.212	0.186

To assess significance of associations between lagged difference in log word frequency from each dataset and difference in log GPR without taking autoregressive behavior of GPR into account, we also estimate ordinary least squares models (OLS) models of the form

$$y_t = X_{t-1}\beta + u_t, \tag{3.18}$$

where all terms in the equation have their same meaning as in the previous paragraphs. We fit these models using exact least squares. No regression coefficient in any design matrix was significant at the $p_c = 0.0125$ level except for "crackdown" in the Storywrangler OT dataset ($\beta_{\text{crackdown}} = 0.0776 \pm 0.030,\ p = 0.011$). We summarize results of model fits in Tables 3.E.7—3.E.10.

The results of this analysis are mixed and warrant in-depth further investigation. We arbitrarily chose the set of words to analyze; this is not appropriate for an analysis of any rigor greater than a simple case study such as this. Rather, further analysis should choose 1-, 2-, and 3-grams to analyze using a more principled analysis, such as a subject matter expert interview that we now describe.

A group of acknowledged experts in geopolitical risk (e.g., military strategic planners, emerging markets funds managers, diplomats, and senior humanitarian aid

workers) should be systematically interviewed and asked to provide n-grams that they would (1) use, or expect other experts to use, to describe both impactful and non-impactful events in developing countries, (2) expect to hear non-experts use to describe these events. The Storywrangler OT and AT datasets, as well as the Google trends, News, and similar datasets, can then be queried for these words and an association analysis performed. This methodology will help to eliminate the selection bias that we have probably introduced via our word selection.

Model	AutoReg(4)	**AIC**	-2.349
Method	Conditional MLE	**BIC**	-2.205
Data	GPR Only	**HQIC**	-2.290

	coef	**std err**	**z**	**P> |z|**	v[**0.025**	**0.975**]
intercept	0.0188	0.028	0.682	0.495	-0.035	0.073
GPR.L1	-0.4028	0.093	-4.341	0.000	-0.585	-0.221
GPR.L2	-0.3858	0.098	-3.926	0.000	-0.578	-0.193
GPR.L3	-0.2650	0.098	-2.709	0.007	-0.457	-0.073
GPR.L4	-0.1702	0.094	-1.815	0.070	-0.354	0.014

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	0.5127	-1.3379j	1.4328	-0.1918
AR.2	0.5127	+1.3379j	1.4328	0.1918
AR.3	-1.2913	-1.0930j	1.6918	-0.3882
AR.4	-1.2913	+1.0930j	1.6918	0.3882

Table 3.E.2: Summary of conditional MLE inference results, fitting null AR model to GPR data only.

Model	AutoReg-X	**AIC**	-2.332
Method	Conditional MLE	**BIC**	-1.996
Data	GPR + Storywrangler OT	**HQIC**	-2.195

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	0.0208	0.026	0.799	0.424	-0.030	0.072
GPR.L1	-0.2999	0.092	-3.253	0.001	-0.481	-0.119
GPR.L2	-0.3872	0.093	-4.155	0.000	-0.570	-0.205
GPR.L3	-0.2371	0.095	-2.501	0.012	-0.423	-0.051
GPR.L4	-0.2056	0.091	-2.249	0.025	-0.385	-0.026
revolution	-0.0461	0.042	-1.097	0.273	-0.128	0.036
rebellion	-0.0315	0.032	-0.973	0.331	-0.095	0.032
uprising	0.0628	0.035	1.782	0.075	-0.006	0.132
coup	-0.0296	0.030	-0.996	0.319	-0.088	0.029
overthrow	0.0426	0.031	1.369	0.171	-0.018	0.104
unrest	0.0363	0.045	0.802	0.423	-0.052	0.125
protests	-0.0486	0.040	-1.212	0.226	-0.127	0.030
crackdown	0.0712	0.028	2.587	0.010	0.017	0.125

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	0.5643	-1.2589j	1.3796	-0.1829
AR.2	0.5643	+1.2589j	1.3796	0.1829
AR.3	-1.1411	-1.1198j	1.5988	-0.3765
AR.4	-1.1411	+1.1198j	1.5988	0.3765

Table 3.E.3: Summary of conditional MLE inference results, fitting AR-X model to GPR data with Storywrangler OT.

Model	AutoReg-X	**AIC**	-2.269
Method	Conditional MLE	**BIC**	-1.933
Data	GPR + Storywrangler AT	**HQIC**	-2.133

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	0.0199	0.027	0.743	0.458	-0.033	0.072
GPR.L1	-0.3365	0.094	-3.575	0.000	-0.521	-0.152
GPR.L2	-0.3954	0.098	-4.030	0.000	-0.588	-0.203
GPR.L3	-0.2547	0.096	-2.641	0.008	-0.444	-0.066
GPR.L4	-0.1954	0.092	-2.114	0.034	-0.377	-0.014
revolution	-0.0366	0.045	-0.809	0.418	-0.125	0.052
rebellion	-0.0085	0.035	-0.244	0.807	-0.077	0.060
uprising	0.0381	0.037	1.022	0.307	-0.035	0.111
coup	-0.0417	0.032	-1.304	0.192	-0.104	0.021
overthrow	0.0431	0.036	1.197	0.231	-0.027	0.114
unrest	0.0347	0.049	0.701	0.483	-0.062	0.132
protests	-0.0382	0.041	-0.932	0.351	-0.118	0.042
crackdown	0.0408	0.030	1.382	0.167	-0.017	0.099

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	0.5379	-1.2792j	1.3877	-0.1866
AR.2	0.5379	+1.2792j	1.3877	0.1866
AR.3	-1.1896	-1.1147j	1.6302	-0.3802
AR.4	-1.1896	+1.1147j	1.6302	0.3802

Table 3.E.4: Summary of conditional MLE inference results, fitting AR-X model to GPR data with Storywrangler AT.

Model	AutoReg-X	**AIC**	-2.236
Method	Conditional MLE	**BIC**	-1.900
Data	GPR + Google Trends	**HQIC**	-2.099

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	0.0192	0.027	0.705	0.481	-0.034	0.073
GPR.L1	-0.4184	0.094	-4.442	0.000	-0.603	-0.234
GPR.L2	-0.3944	0.101	-3.890	0.000	-0.593	-0.196
GPR.L3	-0.2607	0.097	-2.694	0.007	-0.450	-0.071
GPR.L4	-0.1809	0.094	-1.925	0.054	-0.365	0.003
revolution	-0.0262	0.034	-0.774	0.439	-0.092	0.040
rebellion	0.0212	0.032	0.670	0.503	-0.041	0.083
uprising	0.0265	0.027	0.987	0.324	-0.026	0.079
coup	0.0198	0.029	0.685	0.493	-0.037	0.076
overthrow	0.0067	0.033	0.204	0.838	-0.058	0.071
unrest	7.62e-5	0.038	0.002	0.998	-0.074	0.074
protests	-0.0025	0.035	-0.070	0.944	-0.072	0.067
crackdown	0.0172	0.028	0.609	0.543	-0.038	0.072

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	0.5177	-1.3376j	1.4342	-0.1912
AR.2	0.5177	+1.3376j	1.4342	0.1912
AR.3	-1.2383	-1.0741j	1.6392	-0.3863
AR.4	-1.2383	+1.0741j	1.6392	0.3863

Table 3.E.5: Summary of conditional MLE inference results, fitting AR-X model to GPR data with Google trends.

Model	AutoReg-X	**AIC**	-2.241
Method	Conditional MLE	**BIC**	-1.905
Data	GPR + Cable News	**HQIC**	-2.104

	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
intercept	0.0196	0.027	0.720	0.471	-0.034	0.073
GPR.L1	-0.3702	0.095	-3.880	0.000	-0.557	-0.183
GPR.L2	-0.4010	0.100	-3.994	0.000	-0.598	-0.204
GPR.L3	-0.2528	0.100	-2.518	0.012	-0.450	-0.056
GPR.L4	-0.2095	0.100	-2.104	0.035	-0.405	-0.014
revolution	0.0099	0.030	0.328	0.743	-0.049	0.069
rebellion	-0.0159	0.029	-0.539	0.590	-0.074	0.042
uprising	-0.0279	0.036	-0.778	0.437	-0.098	0.042
coup	0.0367	0.032	1.155	0.248	-0.026	0.099
overthrow	0.0235	0.030	0.779	0.436	-0.036	0.083
unrest	0.0174	0.037	0.465	0.642	-0.056	0.091
protests	0.0059	0.038	0.157	0.875	-0.068	0.079
crackdown	0.0317	0.030	1.043	0.297	-0.028	0.091

	Real	**Imaginary**	**Modulus**	**Frequency**
AR.1	0.5425	-1.2819j	1.3920	-0.1863
AR.2	0.5425	+1.2819j	1.3920	0.1863
AR.3	-1.1461	-1.0726j	1.5697	-0.3803
AR.4	-1.1461	+1.0726j	1.5697	0.3803

Table 3.E.6: Summary of conditional MLE inference results, fitting AR-X model to GPR data with cable news.

Model	OLS	**Method**	Least Squares
R-squared	0.126	**Adj. R-squared**	0.062
Log-Likelihood	-25.449	**Durbin-Watson**	2.328
AIC	68.90	**BIC**	93.83
F-statistic	1.973	**Prob (F-statistic)**	0.0565
Omnibus	6.186	**Prob(Omnibus)**	0.045
Jarque-Bera (JB)	6.251	**Prob(JB)**	0.0439
Kurtosis	3.826	**Skew**	0.384

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.0064	0.029	0.223	0.824	-0.051	0.063
revolution	-0.0443	0.045	-0.989	0.325	-0.133	0.044
rebellion	-0.0295	0.035	-0.834	0.406	-0.100	0.041
uprising	0.0800	0.038	2.093	0.039	0.004	0.156
coup	-0.0260	0.033	-0.781	0.437	-0.092	0.040
overthrow	0.0469	0.035	1.354	0.179	-0.022	0.116
unrest	0.0006	0.046	0.013	0.989	-0.091	0.092
protests	-0.0394	0.042	-0.931	0.354	-0.123	0.045
crackdown	0.0776	0.030	2.578	0.011	0.018	0.137

Table 3.E.7: OLS summary: GPR and Storywrangler OT.

Model	OLS	**Method**	Least Squares
R-squared	0.071	**Adj. R-squared**	0.003
Log-Likelihood	-29.094	**Durbin-Watson**	2.345
AIC	76.19	**BIC**	101.1
F-statistic	1.039	**Prob (F-statistic)**	0.412
Omnibus	8.949	**Prob(Omnibus)**	0.011
Jarque-Bera (JB)	9.355	**Prob(JB)**	0.00930
Kurtosis	3.885	**Skew**	0.529

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.0058	0.030	0.195	0.846	-0.053	0.065
revolution	-0.0457	0.050	-0.916	0.362	-0.145	0.053
rebellion	0.0004	0.038	0.011	0.991	-0.075	0.076
uprising	0.0600	0.041	1.473	0.144	-0.021	0.141
coup	-0.0377	0.035	-1.077	0.284	-0.107	0.032
overthrow	0.0598	0.040	1.486	0.140	-0.020	0.139
unrest	-0.0043	0.052	-0.084	0.933	-0.107	0.098
protests	-0.0294	0.043	-0.680	0.498	-0.115	0.056
crackdown	0.0423	0.032	1.316	0.191	-0.021	0.106

Table 3.E.8: OLS summary: GPR and Storywrangler AT.

Model	OLS	**Method**	Least Squares
R-squared	0.013	**Adj. R-squared**	-0.059
Log-Likelihood	-32.636	**Durbin-Watson**	2.479
AIC	83.27	**BIC**	108.2
F-statistic	0.1843	**Prob (F-statistic)**	0.993
Omnibus	12.770	**Prob(Omnibus)**	0.002
Jarque-Bera (JB)	15.141	**Prob(JB)**	0.000515
Kurtosis	4.195	**Skew**	0.643

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.0058	0.031	0.188	0.851	-0.055	0.066
revolution	-0.0204	0.038	-0.537	0.592	-0.096	0.055
rebellion	0.0206	0.036	0.575	0.567	-0.050	0.091
uprising	0.0118	0.030	0.394	0.694	-0.048	0.071
coup	0.0161	0.033	0.493	0.623	-0.049	0.081
overthrow	0.0010	0.035	0.028	0.978	-0.069	0.071
unrest	0.0079	0.041	0.191	0.849	-0.074	0.090
protests	-0.0040	0.040	-0.100	0.920	-0.082	0.074
crackdown	0.0192	0.032	0.598	0.551	-0.045	0.083

Table 3.E.9: OLS summary: GPR and Google trends.

Model	OLS	**Method**	Least Squares
R-squared	0.039	**Adj. R-squared**	-0.032
Log-Likelihood	-31.092	**Durbin-Watson**	2.462
AIC	80.18	**BIC**	105.1
F-statistic	0.5504	**Prob (F-statistic)**	0.816
Omnibus	14.135	**Prob(Omnibus)**	0.001
Jarque-Bera (JB)	17.910	**Prob(JB)**	0.000129
Kurtosis	4.366	**Skew**	0.666

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.0068	0.030	0.226	0.821	-0.053	0.067
revolution	0.0217	0.033	0.659	0.511	-0.044	0.087
rebellion	-0.0156	0.032	-0.486	0.628	-0.079	0.048
uprising	-0.0187	0.035	-0.530	0.597	-0.089	0.051
coup	0.0460	0.033	1.395	0.166	-0.019	0.111
overthrow	0.0082	0.033	0.250	0.803	-0.057	0.074
unrest	0.0104	0.041	0.252	0.801	-0.071	0.092
protests	0.0210	0.041	0.518	0.606	-0.059	0.101
crackdown	0.0326	0.033	0.978	0.330	-0.033	0.099

Table 3.E.10: OLS summary: GPR and cable news.

Chapter 4

Examining the world's collective attention on social media to the COVID-19 pandemic

4.1 Abstract

In confronting the global spread of the coronavirus disease COVID-19 pandemic we must have coordinated medical, operational, and political responses. In all efforts, data is crucial. Fundamentally, and in the possible absence of a vaccine for 12 to 18 months, we need universal, well-documented testing for both the presence of the disease as well as confirmed recovery through serological tests for antibodies, and we need to track major socioeconomic indices. But we also need auxiliary data of all kinds, including data related to how populations are talking about the unfolding pandemic through news and stories. To in part help on the social media side, we curate a set of 2000 day-scale time series of 1- and 2-grams across 24 languages on Twitter that are most 'important' for April 2020 with respect to April 2019. We determine importance through our allotaxonometric instrument, rank-turbulence divergence. We make some basic observations about some of the time series, including a comparison to numbers of confirmed deaths due to COVID-19 over time. We broadly observe across all languages a peak for the language-specific word for 'virus' in January 2020 followed by a decline through February and then a surge through March and April. The world's collective attention dropped away while the virus spread out from China. We host the time series on Gitlab, updating them on a daily basis while relevant. Our main intent is for other researchers to use these time series to enhance whatever analyses that may be of use during the pandemic as well as for retrospective investigations.

4.2 Introduction

Understanding how major disasters affect the wellbeing of populations both in real time and historically is of paramount importance. We especially need real-time measurement to enable policy makers in health systems and government to gauge the immediate situation and evaluate scenarios, and for researchers to model possible future trajectories of social systems. Researchers have demonstrated how characterizing and tracking public discourse of the COVID-19 spread on social media [52, 71, 175] can support local authorities' efforts in response to the global pandemic [27, 296]. Recent studies have also investigated the impact of pre-existing political polarization on discussions related to COVID-19 throughout Twitter's ecosystem [141], as well as the extent of misinformation on social media [42, 140, 224]. Our primary aim here is to generate a particular data stream that may be of help to other researchers: A principled set of n-gram time series across major languages used on Twitter and news-relevant for April, 2020. Our work is complementary to extant efforts to enable research on the COVID-19 pandemic [158, 177] by gathering and sharing epidemiological data [31, 41, 86, 131, 138, 181, 311, 312], economic data, and internet and social media data [22, 49, 50, 58, 169].

In this short piece, we describe how we select languages and n-grams relevant to the time period of the present COVID-19 pandemic; show example time series plots for the word 'virus' (and its translations), including a visual comparison with COVID-19 confirmed case and death numbers; and describe the data sets, figures, and visualizations for 24 languages that we share online.

Rank	Language	Code	Rank	Language	Code
1	English	en	13	Hindi	hi
2	Spanish	es	14	Persian	fa
3	Portuguese	pt	15	Urdu	ur
4	Arabic	ar	16	Polish	pl
5	Korean	ko	17	Catalan	ca
6	French	fr	18	Dutch	nl
7	Indonesian	id	19	Tamil	ta
8	Turkish	tr	20	Greek	el
9	German	de	21	Swedish	sv
10	Italian	it	22	Serbian	sr
11	Russian	ru	23	Finnish	fi
12	Tagalog	tl	24	Ukrainian	uk

Table 4.2.1: The 24 languages for which we provide COVID-19 related Twitter time series.

4.3 Data and methods

4.3.1 Selection of languages and n-grams

We base our curation on our work in two of our previous papers [7, 83], and we draw from a database of approximately 10% of all tweets from 2008-09-09 to present. Our process of obtaining salient n-grams for April 2020 comprises two steps. First, we used the language identification and detection tool FastText-LID [32, 143] to evaluate all tweets in our historical archive, finding over 100 languages [7]. Besides analyzing all tweets (AT), we also separately process what we call organic tweets (OT): All Twitter messages which are original. Organic tweets exclude retweets while including all added text for quote tweets. In doing so, we are able to carry through a measure of spreadability for all n-grams. The key threshold we use for spreading is the naive one from biological and social contagion models: When an n-gram appears in more retweeted than organic material, we view it as being socially amplified. We subsequently extracted day-scale Zipf distributions for 1-, 2-, and 3-grams along with day-scale n-gram time series [5]. We preserve case where applicable, do not apply any stemming. We note that the top 10 languages on Twitter comprise 85% of all tweets. Here, we take 24 of the most commonly used languages on Twitter in 2019, with the provision that we are able to parse them into n-grams. For the time being, we are unable to reliably parse continuous-based script languages such as Japanese, Thai, and Chinese, the 2nd, 6th, and 13th most common languages. The selected languages comprise two thirds of the daily tweets on the platform. We exclude all tweets not assigned a language with sufficient confidence (an effective 4th ranked collection). In

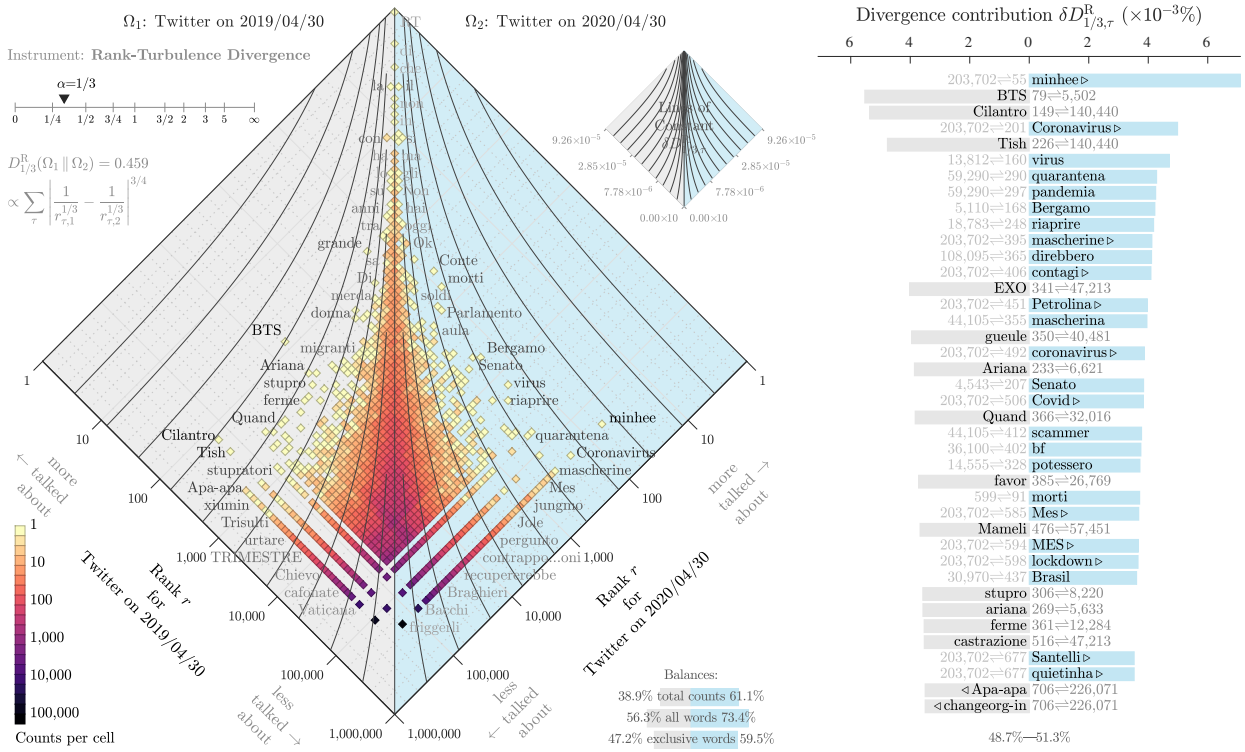

Figure 4.3.1: Allotaxonograph using rank-turbulence divergence for Italian word usage on April 30, 2019 versus April 30, 2020. For this visualization, we consider the subset of 1-grams that are formed from latin characters. The right hand sides of the rank-rank histogram and the rank-turbulence contribution list are dominated by COVID-19 related terms. See Dodds et al. [83] for a full explanation of our allotaxonometric instrument.

other words, we select the predicted language with the highest confidence score. If the confidence score of our FastText-LID model is less than 25% for a given tweet, then we label that tweet as Undefined (und). We also choose to include Ukranian (29th) over Cebuano (28th) due to a marginal degree of uncertainty for detecting messages written in Cebuano [7]. We list the 24 languages by overall usage frequency in Tab. 4.2.1.

Second, we compare usage of n-grams in April of 2020 with April 2019 to determine which n-grams have become most elevated in relative usage. We do so by using rank-

turbulence divergence [83], an instrument for comparing any pair of heavy-tailed size distributions of categorical data. Other well-considered divergences will produce similar lists. For each language, we take Zipf distributions for each day of April 2020, and compare them with the Zipf distributions of 52 weeks earlier. For an example, we show in Fig. 4.3.1 an allotaxonograph for Italian comparing 2019-04-30 and 2020-04-30. The main plot displays a rotated 2D-histogram to avoid misinterpretation of causality. We bin n-grams logarithmically such that bins located near the center vertical line indicate n-grams that are used equivalently on both days, whereas bins on either side highlight n-grams that are used more often on the corresponding date. We use rank-turbulence divergence with the parameter α set to $1/3$ as this provides a reasonable fit to the lexical turbulence we observe [83, 222]. Up to a normalization factor [83], we compute rank-turbulence divergence for each n-gram τ as follows:

$$\delta D^{\mathrm{R}}_{\alpha,\tau} \propto \left| \frac{1}{r^{\alpha}_{\tau,t_1}} - \frac{1}{r^{\alpha}_{\tau,t_2}} \right|^{1/(\alpha+1)} = \left| \frac{1}{r^{1/3}_{\tau,t_1}} - \frac{1}{r^{1/3}_{\tau,t_2}} \right|^{3/4},$$

where r_{τ,t_1} and r_{τ,t_2} indicate the rank of usage for τ at time step t_1 and t_2 respectively. We plot contour lines to demonstrate the scale of rank-turbulence divergence and use divergence contributions of each n-gram to compile an ordered set of relevant n-grams for each day (see right panel of Fig. 4.3.1). For ease of plotting, we have further chosen to compare the subset of words containing Latin characters only. Words associated with the pandemic dominate the contributions from 2020-04-30. On the right side of the allotaxonograph, we see 'Coronavirus', 'virus', 'quarantina', 'pandemia', 'Bergamo', and 'morti'. We repeat this process for every day in April, and combine divergence contributions for all n-grams across these days, and rank n-grams in descending order indicating the most narratively dominate n-grams for the month

of April.

4.3.2 Data, visualizations, and sites

For each language, and for each of the top n-grams we have identified, we extract three day-scale time series starting on 2019-09-01: Daily counts, ranks, and normalized frequencies based on the Eastern Time Zone (ET). Understandably, as the pandemic was unfolding in early 2020, most regional health organizations could not confirm the roots or exact initial date of the first COVID-19 case within their population of charge, with speculations that the virus may have started spreading in late 2019. Therefore, we started our data collection on September of 2019 to cover the last quarter of 2019 and the few months leading to the pandemic spreading worldwide.

The degree to which the pandemic is being discussed on Twitter is of great interest in itself, and our data set will allow for such examination.

For the n-grams our method surfaces, we observe variations in punctuation and grammatical structures. These variants as well as non-pandemic-related elements may be filtered out for individual languages by hand as may suit interested researchers. We provide a cleaned version of the data set whereby we omit links, handles, hashtags, emojis, and punctuation. We also note that our decision to respect capitalization leads to n-grams that some researchers may wish to collapse, and we also provide a case-insensitive version of our data set. We repeat all of the above steps for n-grams derived from organic tweets (OT).

We share and maintain all data on Gitlab at: `https://gitlab.com/compstorylab/covid19ngrams`. We also provide a connected website associated with our paper at: `http://compstorylab.org/covid19ngrams/`. We show tables of the leading n-

English	Spanish	Portuguese	Arabic	Korean	French
coronavirus	cuarentena	quarentena	خصم	코로나	confinement
pandemic	pandemia	Babu	كود	한승우	masques
virus	coronavirus	live	كوبون	승우	Coronavirus
lockdown	virus	babu	نمشي	n번방	virus
quarantine	confinamiento	Manu	تخفيض	마스크	coronavirus
Coronavirus	mascarillas	Thelma	إستخدم	스위치	masque
deaths	Coronavirus	pandemia	نسناس	해찬	pandémie
masks	casos	Rafa	فوغا	스밍	sanitaire
cases	salud	coronavírus	كورونا	N번방	crise
distancing	sanitaria	vírus	كلوسيت	수호	tests
China	fallecidos	manu	اند	정우	soignants
testing	test	Gizelly	قسيمة	그	déconfinement
workers	medidas	Mari	ستايلي	안	décès
tested	crisis	paredão	الكود	온라인	Sco
PPE	médicos	isolamento	اوناس	크래비티	patients
crisis	contagios	rafa	اتش	사회적	manaa
mask	aislamiento	Ivy	رهيب	그냥	Raoult
COVID	sanitarios	bbb	الوباء	한	période
Fauci	Gobierno	gizelly	فيروس	이	Confinement
Corona	contagio	corona	يخصم	닌텐도	confiné

Indonesian	Turkish	German	Italian	Russian	Tagalog
corona	CenkKaraçay	Corona	Coronavirus	коронавируса	quarantine
Corona	maske	Masken	quarantena	коронавирусом	SB19
PKP	NedimKaraçay	Virus	virus	карантина	na
virus	CemNed	GT	MES	самоизоляции	lockdown
pandemi	virüs	Krise	mascherine	карантин	ECQ
masker	çıkma	Coronavirus	Lombardia	коронавирус	tiktok
ak	sağlık	Pandemie	coronavirus	пандемии	covid
wabah	BerkerGüven	Maske	pandemia	карантине	frontliners
pasien	vaka	Abstand	Mes	маски	virus
PSBB	Sağlık	bgt	Conte	эпидемии	ecq
covid	koronavirüs	Quarantäne	2020	масок	Alab
bgt	evde	Lockdown	contagi	вирус	ghorl
aku	2020	Maßnahmen	mascherina	ИВЛ	gobyerno
online	Koronavirüs	Coronakrise	Covid	врачей	relief
mutualan	yardım	hyung	tamponi	случаев	ayuda
positif	yasağı	Mundschutz	SIGA	заболевших	pandemic
ni	Korona	Feb	FAV	заражения	series
mudik	hasta	Zeiten	contagio	вируса	kalat
pkp	SeraKutlubey	ak	lockdown	Коронавирус	DDS
hyung	korona	Lockerungen	positivi	защиты	workout

Table 4.3.1: Top 20 (of 1,000) 1-grams for our top 12 languages for the first three weeks of April 2020 relative to a year earlier. Our intent is to capture 1-grams that are topically and culturally important during the COVID-19 pandemic. While overall, we see pandemic-related words dominate the lists across languages, we also find considerable specific variation. Words for virus, quarantine, protective equipment, and testing show different orderings (note that we do not employ stemming). Unrelated 1-grams but important to the time of April 2020 are in evidence; the balance of these are important for our understanding of how much the pandemic is being talked about. To generate these lists we use the allotaxonometric method of rank-turbulence divergence to find the most distinguishing 1-grams (see Sec. 4.3.1, Fig. 4.3.1, and Dodds et al. [83]).

Hindi	Persian	Urdu	Polish	Catalan	Dutch
यरस	کرونا	کرونا	koronawirusa	confinament	Bomboclaat
तबल	ویروس	کورونا	epidemii	coronavirus	corona
मरकज	قرنطینه	وائرس	wyborów	crisi	How
Asharamji	ماسک	لاک	pandemii	mascaretes	Corona
तरजन	چین	راشن	testów	pandèmia	virus
Lockdown	شیوع	ڈاؤن	koronawirusem	virus	mondkapjes
Corona	بهداشت	وبا	maseczki	residències	coronacrisis
lockdown	مبتلا	ماسک	wybory	sanitaris	coronavirus
शट	ساعات	امداد	maseczek	tests	lockdown
PPE	بیماری	ڈاون	głosowania	sanitari	crisis
टहन	کروناپی	آٹا	zdrowia	sanitària	how
ऊन	آمار	فورس	kwarantanny	mesures	RIVM
Sadhna	زندانیان	ٹائیگر	wirusa	desconfinament	quarantaine
औरवह	بیماران	کٹس	mniez	Gobierno	testen
कड	فولو	مساجد	zakażonych	hospitals	mondmaskers
corona	۹۹	سندھ	SIM	gestió	IC
आपद	بورس	وباء	zgonów	ไหว	getest
FB	سناب	چینی	wirus	salut	besmet
लघर	رائفی	ریورٹ	korespondencyjne	material	vs
नलश	ابتلا	ٹیسٹ	przypadków	นน	pandemie

Tamil	Greek	Swedish	Serbian	Finnish	Ukrainian
ஊரடங	καραντίνα	Tegnell	virusa	amg	карантину
வழங	κρούσματα	Corona	virus	yh	коронавірус
கப	πανδημίας	corona	korona	ak	коронавірусу
கவசம	πανδημία	Ak	mere	yhh	карантин
உணவ	καραντίνας	FHM	korone	koronan	ak
ணமட	Κορωνοϊός	ak	amei	manaa	коронавірусом
கமன	μάσκες	makasii	mera	Sco	маски
உபகரணங	κορονοϊό	tidur	policijski	simm	медиків
ஞரண	κορωνοϊό	smittade	pandemije	obg	хворих
கவசங	μέτρα	viruset	Kon	korona	випадків
RMM	τεστ	coronakrisen	vanrednog	sim	Єрмака
உதவ	κορονοϊού	bgt	вируса	old	Єрмак
ககவசம	κορωνοϊού	äldreboenden	maske	hyyy	пандемії
ஆதரவற	Κορονοϊός	skyddsutrustning	stanja	obrigada	лікарні
Corona	μέτρων	repp	virusom	kriisin	карантині
Back	ιό	dödsfall	čas	aamiin	масок
ரணப	Τσιόδρας	krisen	struka	paling	EU
Comment	καραντινα	munskydd	корона	syg	МОЗ
ID	μάσκα	döda	karantin	muk	Dub
பரவல	ΜΕΘ	virus	epidemije	simmm	добу

Table 4.3.2: Continuing on from Fig. 4.3.1: Top 20 1-grams for the second 12 of 24 languages we study for April 2020 relative to April 2019.

grams in our data set, as well as example "bar chart races" for the dominant COVID-19 n-grams in major languages. Our intention is to automatically update the data set on Gitlab, as soon as we have processed all tweets for a day.

We show the resulting top 20 April-2020-specific 1-grams for the 24 languages in Tabs. 4.3.1 and 4.3.2. For display, we use the cleaned version, omitting hashtags, handles, emojis, numbers, and punctuation. We also removed all variations of 'Bomboclaat' from Dutch. Overall, we see that the lists are dominated by language specific words for coronavirus virus, quarantine, pandemic, testing, and spreading.

In the full, unfiltered data set, some 1-grams such as punctuation represent functional changes in the use of Twitter across languages. The white heart emoji makes the top 20 in a few languages such as English, Arabic, Korean and German. By contrast, and according to the measurements we have used here, the worried face emoji, has become important across many languages in April 2020 relative to April 2019. It would be natural to see this emoji as being pandemic-related but in fact, we see from time series that the worried emoji has slowly being increasing in usage over time for several years (determining the reasons for which we will leave for a separate line of inquiry). All 1-grams are included in the shared raw version of the data sets.

We emphasize that with our approach, we do not explicitly determine whether or not an n-gram is relevant to COVID-19. While the pandemic was one of the top stories of 2020 for the majority of countries, there have of course been other major events and moments in popular culture around the world. For example, in March 2020 for the United States, the democratic primary leads to the 1-gram in English Twitter of 'Biden' being prominent. Similarly, we see many n-grams related to the Big Brother Brazil show in Portuguese, and K-pop in Korean. Further, most languages

have a strong degree of geographic specificity (e.g., Finnish for Finland, Portuguese for Brazil), and we have not filtered for precise geo-location. English, Spanish, Arabic, and French are some of the more geographically distributed languages.

4.4 Results

We briefly consider two sets of sample time series based on our data set. Across Figs. 4.4.1 and 4.4.2, we plot contagiograms [5] for the word 'virus' translated as appropriate in to each of the 24 languages. For each language, we display the daily (Zipfian) rank for 'virus' in the main panel of each plot. We add a grey background indicating the best and worst rank of each week overlaid by a centered weekly rolling average (black). The pale disk highlights the date of maximum observed rate. In the secondary time series at the top of each panel, we show the relative fraction of 1-gram contained in retweets (RT) versus organic tweets (OT). When the RT/OT balance exceeds 50%, we shade the background to indicate that the 1-gram is being spread (e.g., retweeted) more than organically tweeted. For each contagiogram, we also display a heatmap of the relative amplification of each 1-gram compared to the fraction of 1-grams that are found in RTs on that day. For each day of the week, shades of red indicate higher social amplification, whereas gray shows that the volume of that 1-gram is often shared organically. See Alshaabi et al. [5] for technical details of contagiograms.

Alone, the highest ranks for 'virus' show the enormity of the pandemic. While a common enough word in normal times, 'virus' has reached into the top 100 ranks across many languages, a region that we have elsewhere referred to as the realm of

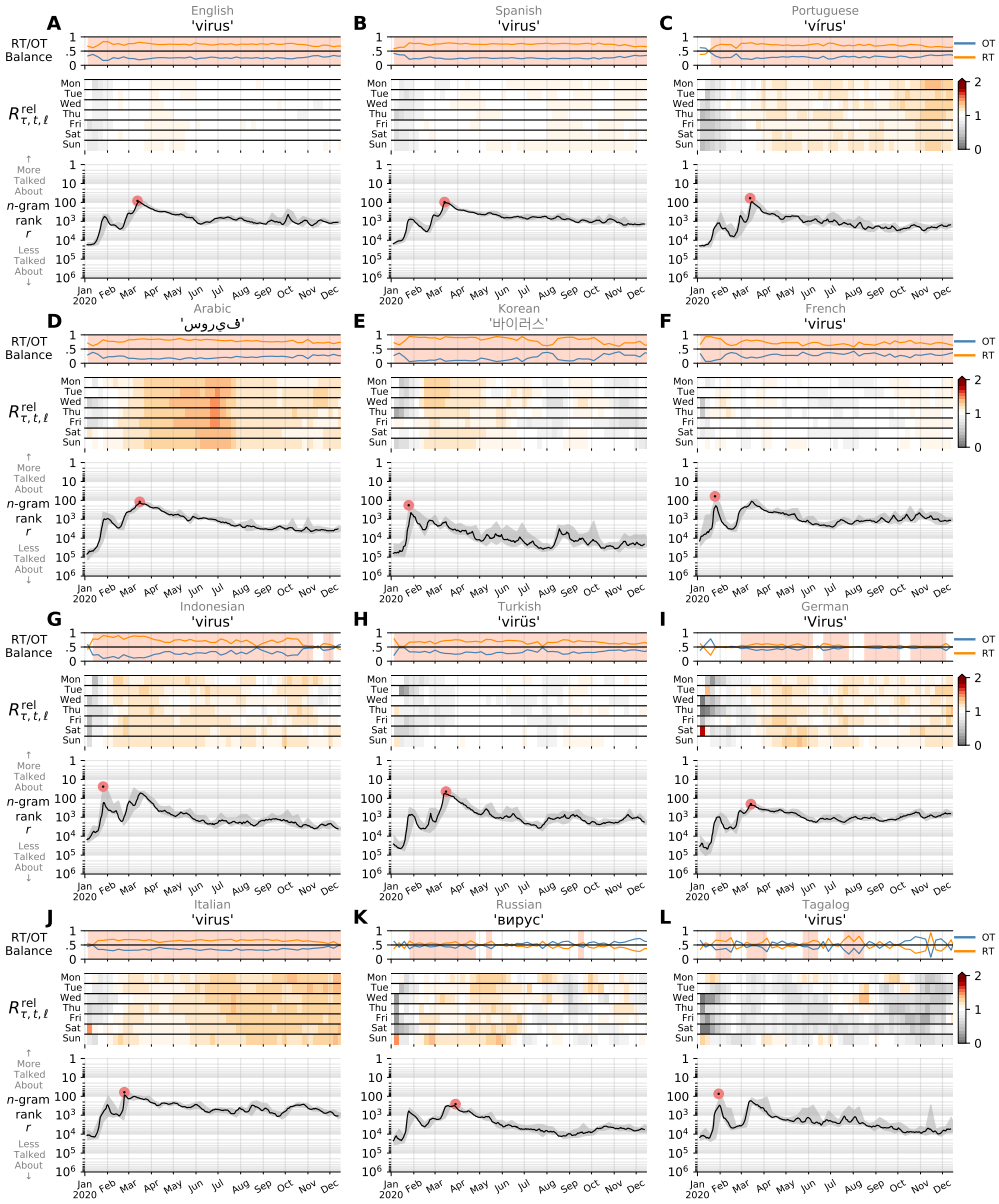

Figure 4.4.1: Contagiograms for the word 'virus' in the top 12 of the 24 languages we study here. The major observation is that the world's attention peaked early in late January around the news of an outbreak of a new infectious disease in Wuhan, declining through well into February before waking back up. The main plots in each panel show usage ranks at the day scale (ET). The solid lines indicating smoothing with a one week average (centered). The plots along the top of each panel show the relative fractions of each 1-gram's daily counts indicating as to whether they appear in retweets (RT, spreading) or organic tweets (OT, new material). The background shading shows when the balance favors spreading—story contagion.

lexical ultrafame [82]. Normally only the most basic function words of a language will populate the top 100 ranks. In the last few months, we have seen 'virus' rise as high as r=24 in Indonesian (2020-01-26), r=27 in Polish (2020-03-11), r=29 in Urdu (2020-03-22), r=44 in German (2020-03-14), and r=83 in English (2020-03-13). In terms of the shapes of the time series for 'virus', most languages show a late January peak consistent with the news from China of a novel coronavirus disease spreading in Wuhan. The subsequent drop in usage rate across most of the 24 languages reflects a global decline in attention being paid to the outbreak. The Italian time series for 'virus' in Fig. 4.4.1J shows an abrupt jump about three quarters of the way through February, strikingly just after a drop in RT/OT balance. Persian has a similar shock jump just after midway of February (Fig. 4.4.1B). We see in Fig 4.4.2E that 'virus' in Catalan shows no early January peak like most of the other 23 languages, suggesting that even the initial news from China did not have great impact.

One of the major problems we face with the COVID-19 pandemic is the unevenness of testing across the world. South Korea and Iceland have tested early and extensively while the United States's testing has been uncoordinated and slow to expand. Urdu's heightened time series for 'virus' (Fig 4.4.2C) would seem especially concerning given low numbers coming out of Pakistan which, as of 2020-03-24, had reported 1,063 cases and 8 deaths [86]. For Indonesia, where testing has also been limited [86] and with peak attention on Twitter coming in January and early focus on economic issues and evacuation of nationals from Wuhan, a dip in the rank of 'virus' in the second half of February is also worrying (Fig 4.4.1G).

Countries around the world have adopted different strategies and policies in response to the coronavirus pandemic. While most languages have COVID-19 related

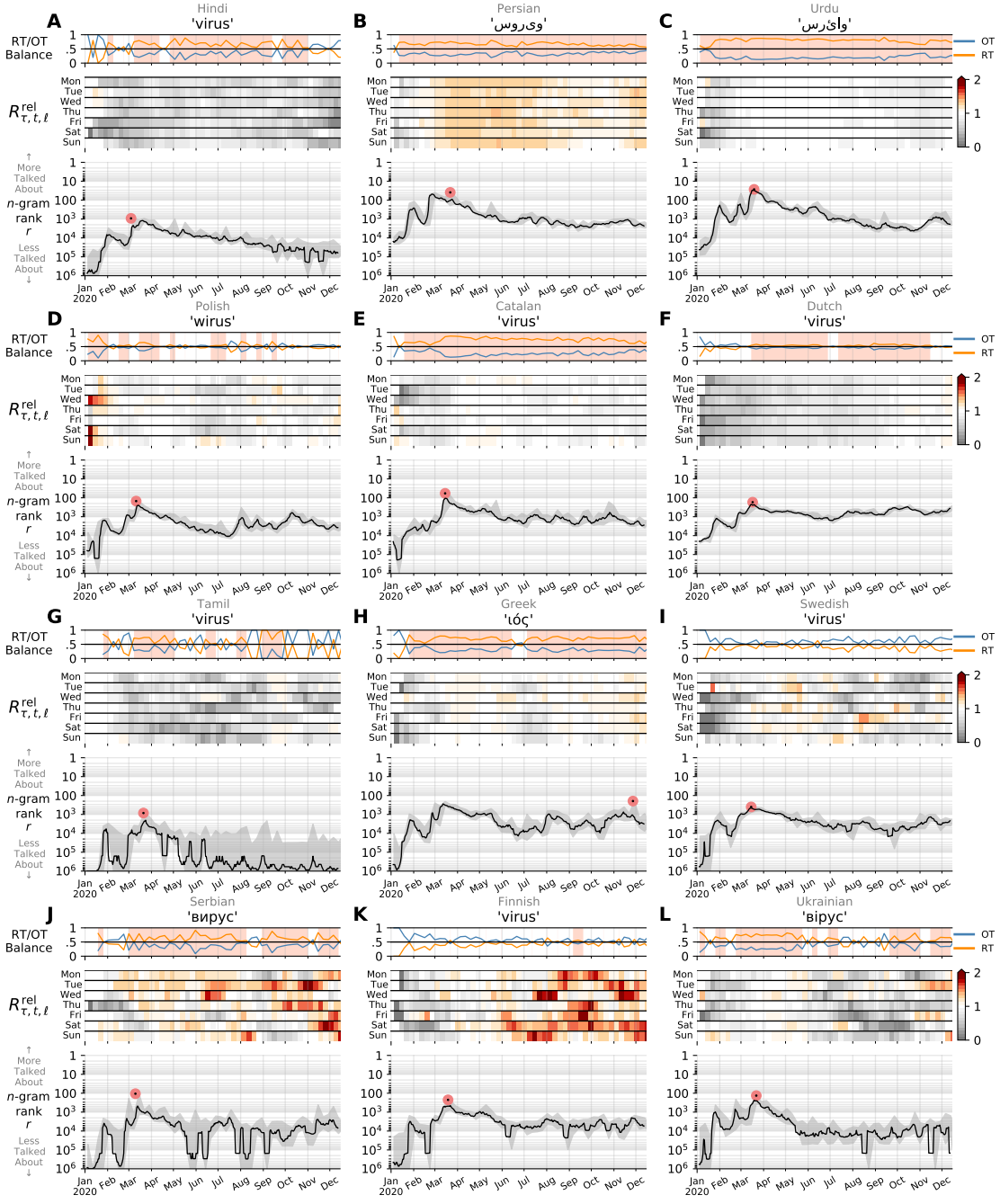

Figure 4.4.2: Following on from Fig. 4.4.1, contagiograms for the word 'virus' in the second 12 of the 24 languages. We note that some of these 1-grams are socially amplified over time, while others often shared organically.

terms across the top n-grams, some languages also have terms related to other big events happening simultaneously. For example, we see many n-grams discussing the democratic primary election in the US. We also find n-grams connected to the Big Brother Brazil show in Portuguese, while Korean has many K-pop references. This in part shows that the collective attention of different populations will, indeed, vary depending on the spread of the virus across countries all over the globe for the time period considered in this study. We note, however, that n-grams related to the pandemic can still be found in Portuguese showing the initial response to the news about the COVID-19 outbreak as the virus started slowly spreading in Brazil.

As one very simple example of comparing our Twitter times series with pandemic-related data, in Fig. 4.4.3, we present plots of daily reported cases and deaths over time for 12 countries, along with time series for 10 salient 1-grams in the top spoken language for each country. We note that the reported number of cases and deaths are subject to under-reportings. For each country, we use the left vertical axis to plot a weekly rolling average of usage ranks at the day scale for 10 1-grams (gray lines) translated in the top spoken language for each country, while the black solid line shows an average of all these 1-grams. We selected 10 1-grams from the top of each list that are directly related to the coronavirus pandemic to highlight the collective attention around the COVID-19 outbreak. The set of 1-grams we use for each language can be found online at: https://gitlab.com/compstorylab/covid19ngrams/-/blob/master/src/consts.py. Using the right vertical axis, we display a weekly rolling average of daily new cases (red solid-line), and reported new deaths (orange dashed-line).

We see a global surge of attention on Twitter starting mid March through April

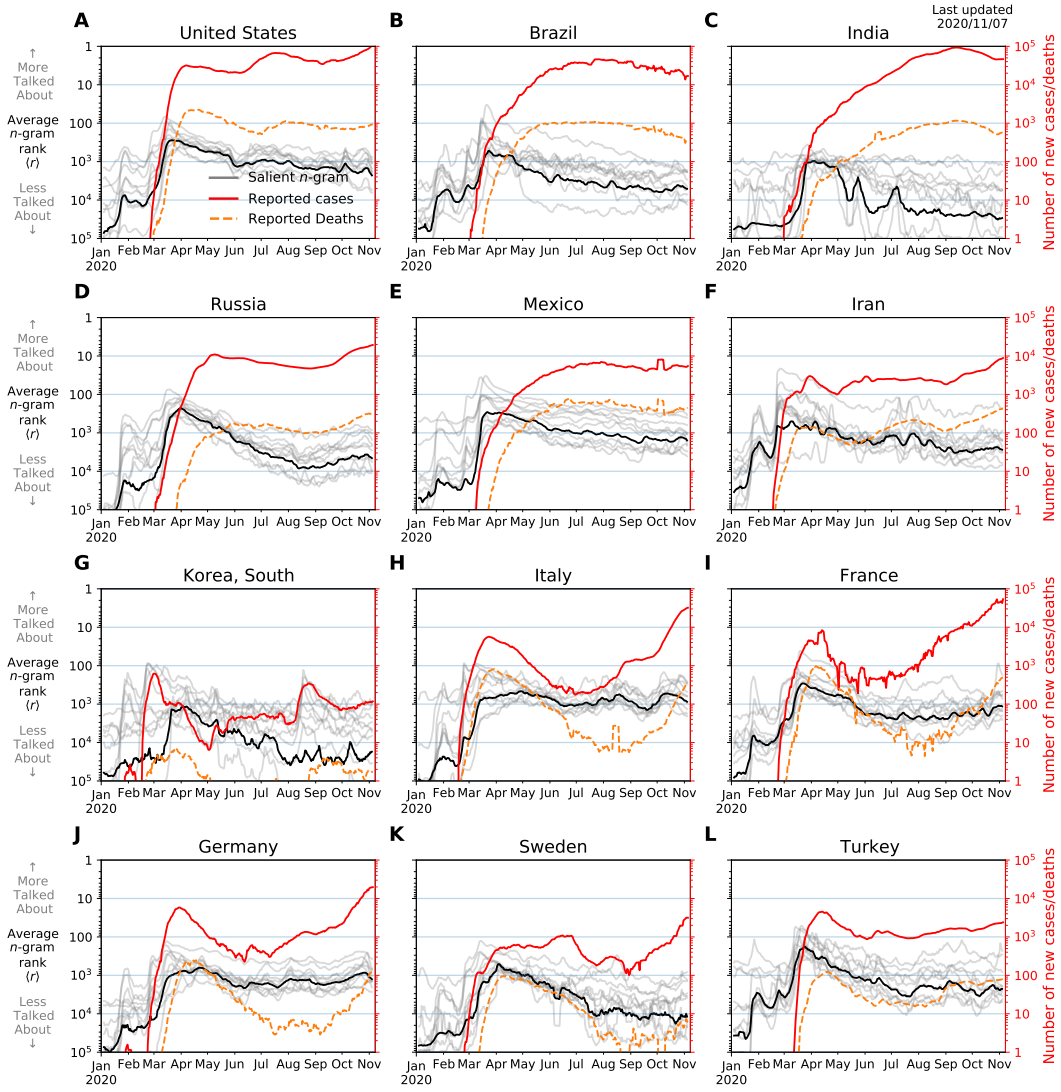

Figure 4.4.3: Time series for daily reported case loads and death compared with a list of 10 salient 1-grams for the top language spoken in each country. For each n-gram, we display a weekly rolling average of usage ranks at the day scale in gray overlaid by an average of all these 1-grams in black marking their corresponding ranks using the left vertical axis. Similarly, we use the right vertical axis to display a weekly rolling average of daily new cases (red solid-line), and reported new deaths (orange dashed-line). We note that the reported counts are underestimates, more so for cases than deaths, and errors are unknown. We sourced data for confirmed cases and fatalities from JHU's COVID-19 project [86]. Starting on 2020-01-22, the project's data has been collected from national and regional health authorities across the world. The data is augmented by case reports from medical associations and social media posts—these later sources are validated against official records before publication.

following the state-wide lockdowns in most countries. Some languages such as Italian and German display a fairly steady level of attention paid to the pandemic. However, the average rank of usage of the selected 1-grams slows down and starts to decay across many languages in April through the summer. In fact, the average rank of usage have dropped an order of magnitude in Indian, Russian, Korean, and Swedish. While the number of new daily cases and deaths are climbing up again, we do not observe the same level of attention reciprocated on Twitter.

4.5 Discussion

We echo our main general observation of how COVID-19 has been discussed through late April 2020: After reacting strongly in late January to the news that a coronavirus-based disease was spreading in China, attention across all but 2 of the 24 languages we survey dropped through February before resurging in late February and through March. We see abrupt shocks in time series as populations shifted rapidly to heightened levels of awareness, particularly in the Italian time series.

In the time series for 'virus', we see two and sometimes three peaks of attention in the space of just a few months. Our hope is that our collection of Twitter n-gram time series that are especially relevant to April 2020 will be of benefit to other researchers. The time series we share will, in part, reflect many other aspects beyond mentions of 'virus', which we have only briefly explored here. Possible topics to investigate include washing (including the soap and microbe emojis), testing, serology, vaccine, masks and protection equipment, social and physical distancing, terms of community support versus loneliness and isolation, closures of schools and universities, economic

problems, job loss, and food concerns.

We repeat that the lists we provide are meant to represent the important n-grams of April 2020, and we urge a degree of caution in the use of the data set. As we have indicated above, our lists of n-grams contain some peculiarities that will not be directly relevant to COVID-19. Entertainment (e.g., movies, celebrities, and K-pop) and sports (football along with sports in the United States) are standard fare on Twitter when no major events are taking place in the world. The extent to which these aspects of Twitter are submerged as pandemic related n-grams rise is of interest.

Finally, while we have been able to identify languages well, geolocation is coarse and at best will be at the level of countries. The strength of geolocation for our time series will depend on the degree of localization of a given language as well as Twitter user demographics. We leave producing n-grams with serviceable physical location as a separate project.

4.6 Acknowledgments

The authors are grateful for support furnished by MassMutual and Google, and the computational facilities provided by the Vermont Advanced Computing Core. Computations were performed on the Vermont Advanced Computing Core supported in part by NSF award No. OAC-1827314. The authors appreciate discussions and correspondence with Aaron Schwartz, Todd DeLuca, Nina Safavi, and Nicholas Danforth.

Chapter 5

Augmenting semantic lexicons using word embeddings and transfer learning

5.1 Abstract

Sentiment-aware intelligent systems are essential to an enormous set of applications from marketing and political campaigns to recommender systems to behavioral and social psychology to intelligence and national security. These sentiment-aware intelligent systems are driven by language models, which fall into two paradigms: lexicon-based or contextual. Although recent contextual models are relevant to a wide range of applications, we still see a tremendous and constant demand for lexicon-based models because of their interpretability, and ease of use. In particular, researchers are often interested in both the linguistic and sociotechnical details of a given sentiment shift—which words contribute most to either a positive or negative sentiment trend. Lexicon dictionaries, however, need to be updated periodically to support new words and expressions that were not examined when the dictionaries were outsourced. Crowdsourcing annotations for semantic dictionaries is an arduous task. Here, we propose two models for predicting sentiment scores to augment semantic lexicons at a relatively low cost using word embeddings and transfer learning. Our first model establishes a baseline employing a simple and shallow neural network initialized with pre-trained word embeddings using a non-contextual approach. Our second model improves upon our baseline, featuring a deep Transformer-based network that brings to bear word definitions to estimate their lexical polarity. Our evaluation shows that both models are able to score new words with a similar accuracy to reviewers from Amazon Mechanical Turk, but at a fraction of the cost.

5.2 Introduction

Sentiment analysis is the process of extracting lexical and emotional semantics from text archives. In computational linguistics, it is often used to study the lexical polarity of words and phrases via natural language processing (NLP). There is an increasing demand for sentiment-aware intelligent systems. Indeed, the synergy of sentiment-aware frameworks and online services can be seen across a vast, multidisciplinary set of applications [12, 192, 209].

With the advent of big data that is time-consuming to analyze, automated sentiment analysis has been used by businesses in evaluating customer feedback to make informed decisions regarding product development and risk management [43, 291]. Combined with recommender systems, it has also been useful in improving consumer experience by providing the consumer with aggregated and curated feedback from other consumers, as in the case of the retail [279**? ?**], e-commerce [**? ?**], and entertainment [218, 281] sectors.

Aside from having applications in industry, sentiment analysis has been widely applied in academic research, particularly in the social and political sciences. Public support for or opposition to pending policies can be gauged from online political discourse, giving policymakers an important window into the public's awareness and attitude [171, 283]. Beyond that, these tools can also forecast elections [290], and monitor inflammatory discourse on social media, with vital relevance to national security [217]. Sentiment analysis has also been used in the public health domain [64, 109, 313], with recent studies analyzing discussions on social media regarding mental health [16, 276], especially in light of the latest advances in NLP and

emerging ethical concerns [63]. The growing number of applications of sentiment-aware systems has prompted tremendous efforts among the NLP community in the past decade to develop end-to-end models to examine short- and medium-length text documents [92, 304], particularly for social media [3, 157, 215].

Sentiment analysis tools differ in two main ways: the definition of the sentiment measure and the model for computing sentiment. The probability of belonging to a discrete class (e.g., positive, negative) is a common way of defining sentiment for a given piece of text. When edge cases are common, adding a neutral class has been reported to improve overall performance [241]. However, sometimes a cardinal measure of sentiment is desired, resulting in a sentiment score rather than a sentiment class [282]. The sentiment scoring paradigm is widely adopted in e-commerce, movies, and restaurant reviews [266].

Sentiment analysis models can be classified into two major paradigms: lexicon-based models and contextual models. Lexicon-based models compute sentiment scores based on sentiment dictionaries that are usually constructed by human annotators [10, 81, 278]. Contextual models, on the other hand, extrapolate semantics by converting words to vectors in an embedding space and learning from large-scale annotated datasets to predict sentiment based on this learned relationship between words [3, 92, 215, 267, 304]. Contextual models have the advantage in differentiating multiple meanings, as in the case of "The dog is *lying* on the beach" vs. "I never said that, you are *lying*", while lexicon-based models usually only have one score for the same word regardless of usage. Despite the flexibility of contextual models, it suffers from less interpretability, explaining the consistent demand for lexicon-based models because of their ease of use [81, 217, 278]. As an example, while a change in sentiment may

be hard to explain with word embeddings, one can use lexicon scores to study how words contribute to sentiment shifts [80, 100, 240].

A big challenge in lexicon-based models is the time and financial investment associated with maintaining them. Lexicon dictionaries need to get updated regularly to mitigate the out-of-vocabulary (OOV) problem—words/phrases that were not considered when the dictionaries were originally constructed [243]. While researchers show general sentiment trends are observable unless the lexicon dictionary does not have enough words, having a versatile dictionary with specialized and rarely used words improves the signal [76]. Notably, language is an evolving sociotechnical phenomenon. New words are often created in real-time, especially on social media [5]. Words occasionally substitute others or drift in meaning over time. For example, the word 'covid' slowly became the most narratively trending n-gram in reference to the global Coronavirus outbreak during 2020 [6].

In this work, we propose an automated framework extending semantic lexicons to OOV words, reducing the need for crowdsourcing scores from human annotators, a process that can be time-consuming and expensive. Although our framework can be used in a more general sense, we focus on predicting *happiness scores* based on the labMT dataset [81]. This dataset rates the "happiness" of words on a continuous scale, averaging scores from multiple human annotators for more than 10,000 words. We discuss this dataset in detail in Sec. 5.4.1. In Sec. 5.3, we discuss recent developments in using deep learning in natural language processing and how they relate to our work. We introduce two models, demonstrating accuracy on par with human performance (see Sec. 5.4 for technical details). We first introduce a baseline model—a small neural network initialized with pre-trained word embeddings—to gauge happiness

scores. Second, we present a deep Transformer-based model that uses word definitions to estimate their lexical polarity. We will refer to our models as the 'Token' and 'Dictionary' models, respectively. We present our results and model evaluation in Sec. 5.5, highlighting how the models perform compared with reviewers from Amazon Mechanical Turk. Finally, we highlight key limitations of our approach, and we outline some potential future developments in our concluding remarks.

5.3 Related work

Word embeddings are numerical representations of words, whereby words with similar semantics have similar representations [21]. Researchers have shown that efficient representations of words can both express meanings and preserve context [129, 130, 176, 187]. While there are many ways to construct such embedding (e.g., matrix factorization), we often use the term to refer to a specific class of word embeddings that are learnable via neural networks.

Word2Vec is one of the key breakthroughs in NLP introducing an efficient way for learning word embeddings from a given text corpus [199, 200]. At its core, it builds off a simple idea borrowed from linguistics and formally known as the 'distributional hypothesis'—words that are semantically similar are also used in similar ways and likely to appear with similar context words [119].

Starting from a fixed vocabulary, we can learn a vector representation for each word via a shallow network with a single hidden layer trained in one of two fashions [199, 200]. Both approaches formalize the task as a unsupervised prediction problem, whereby an embedding is learned jointly with a network that is trained

to either predict an anchor word given the words around it (i.e., continuous bag-of-words (CBOW)), or by predicting context words for an anchor word (i.e., skip-gram (SG)) [199].

Both approaches, however, are limited to local context bounded by the size of the context window. Global Vectors (GloVe) addresses that problem by capturing corpus global statistics with a word co-occurrence probability matrix [223].

While Word2Vec and GloVe offer substantial improvements over previous methods, they both fail to encode unfamiliar words—tokens that were not processed in the training corpora. FastText refines word embeddings by supplementing the learned embedding matrix with subwords to overcome the challenge of OOV tokens [32, 144]. This is achieved by training the network with character-level n-grams ($n \in \{3, 4, 5, 6\}$), then taking the sum of all subwords to construct a vector representation for any given word. Although the idea behind FastText is rather simple, it presents an elegant solution to account for rare words, allowing the model to learn more general word representations.

A major shortcoming of the earlier models is their inability to capture contextual descriptions of words as they all produce a fixed vector representation for each word. In building context-aware models, researchers often use fundamental building blocks such as recurrent neural networks (RNN) [252]—particularly long short-term memory (LSTM) [124]—that are designed to process sequential data. Many methods have provided incremental improvements over time [51, 172, 225]. ELMo is one of the key milestones towards efficient contextualized models, using deep bi-directional LSTM language representations [226].

In late 2017, the advent of Transformers [298] rapidly changed the landscape in

the NLP community. The encoder-decoder framework, powered by attention blocks, enables faster processing of the input sequence while also preserving context [298]. Recent adaptations of the building blocks of Transformers continue to break records, improving the state-of-the-art across all NLP benchmarks with recent applications to computer vision and pattern recognition [87].

Exploiting the versatile nature of Transformers, we observe the emergence of a new family of language models widely known as self-supervised models such as bidirectional encoders (e.g., BERT) [73], and left-to-right decoders (e.g., GPT) [238]. Self-supervised language models are pre-trained by masking random tokens in the unlabeled input data and training the model to predict these tokens. Researchers leverage recent subword tokenization techniques, such as WordPiece [310], SentencePiece [162], and Byte Pair Encoding (BPE) [261], to overcome the challenge of rare and OOV words. Subtle contextualized representations of words can be learned by predicting whether sentence B follows sentence A [73]. Pre-trained language models can then be fine-tuned using labeled data for downstream NLP tasks, such as NER, question answering, text summarization, and sentiment analysis [73, 238].

Recent advances in NLP continue to improve the language facility of Transformer-based models. The introduction of XLNet [314] is another remarkable breakthrough, that combines the bidirectionality of BERT [73] and the autoregressive pre-training scheme from Transformer-XL [68].

While the current trend of making ever-larger and deeper language models shows an impressive track record, it is arguably unfruitful to maintain unreasonably large models that only giant corporations can afford to use due to hardware limitations [284]. Vitally, some language models need to be both computationally efficient and on par

with larger models in terms of performance. Addressing that challenge, researchers proposed clever techniques of leveraging knowledge distillation [121] to train smaller and faster models (e.g, DistilBERT [258]). Similarly, efficient parameterization strategies via sharing weights across layers can also reduce the size of the model while maintaining state-of-the-art results (e.g., ALBERT [170]).

5.4 Data and methods

In this paper, we propose two models for predicting happiness scores for the labMT lexicon [81]—a general-purpose lexicon used to measure happiness in text corpora (see Sec. 5.4.1 for more details).

Our first model is a small neural network initialized with pre-trained FastText word embeddings. The model uses fixed word representations to gauge the happiness score for a given expression, enabling us to augment the labMT dataset at a low cost. For simplicity, we will refer to this model as the Token model.

Bridging the link between lexicon-based and contextualized models, we also propose a deep Transformer-based model that uses word definitions to estimate their happiness scores—namely, the Dictionary model. The contextualized nature of the input data allows our model to accurately estimate the expressed happiness score for a given word based on its lexical meaning.

We implement our models using Tensorflow [1] and Transformers [306]. See Sec. 5.4.2 and Sec. 5.4.3 for further technical details of our Token and Dictionary models, respectively. Our source code, along with pre-trained models, are publicly available via our GitLab repository.

5.4.1 Data

In this study, we use the labMT dataset as an example semantic dictionary to test and evaluate our models [81]. The labMT lexicon has a little over ten thousand unique words—combining the five thousand most frequently used words from New York Times articles, Google Books, Twitter messages, and music lyrics [81]. It is a lexicon designed to gauge the happiness (i.e., valence or hedonic tone) of text archives. Happiness is defined on a continuous scale $h \in \{1 \rightarrow 9\}$, where 1 bounds the most negative (sad) side of the spectrum, and 9 is the most positive (happy). Ratings for each word are crowdsourced via Amazon Mechanical Turk (AMT), taking the average score h_{avg} from 50 reviewers to set a happiness score for any given word.

The labMT dataset also powers the Hedonometer, an instrument to measure the daily rate of happiness on Twitter [80]. Over the past few years, the dictionary was updated a few times to include new words that were not found in the original survey (e.g, terms related to the COVID19 pandemic [6]).

We are particularly interested in this dataset because it also provides the standard deviation of human ratings for each word, which we use to evaluate our models. In this work, we propose two models to estimate h_{avg} using word embeddings, and thus provide an automated tool to augment the labMT dataset both reliably and efficiently.

In Fig. 5.4.1, we display a 2D histogram of the happiness scores in the labMT dataset compared with the standard deviation of human ratings for each given word. The figure highlights the degree of uncertainty in human ratings of the emotional valence of words. While the majority of words are neutral, with a score between 4 and 6, we still observe a human positivity bias in the English language [81].

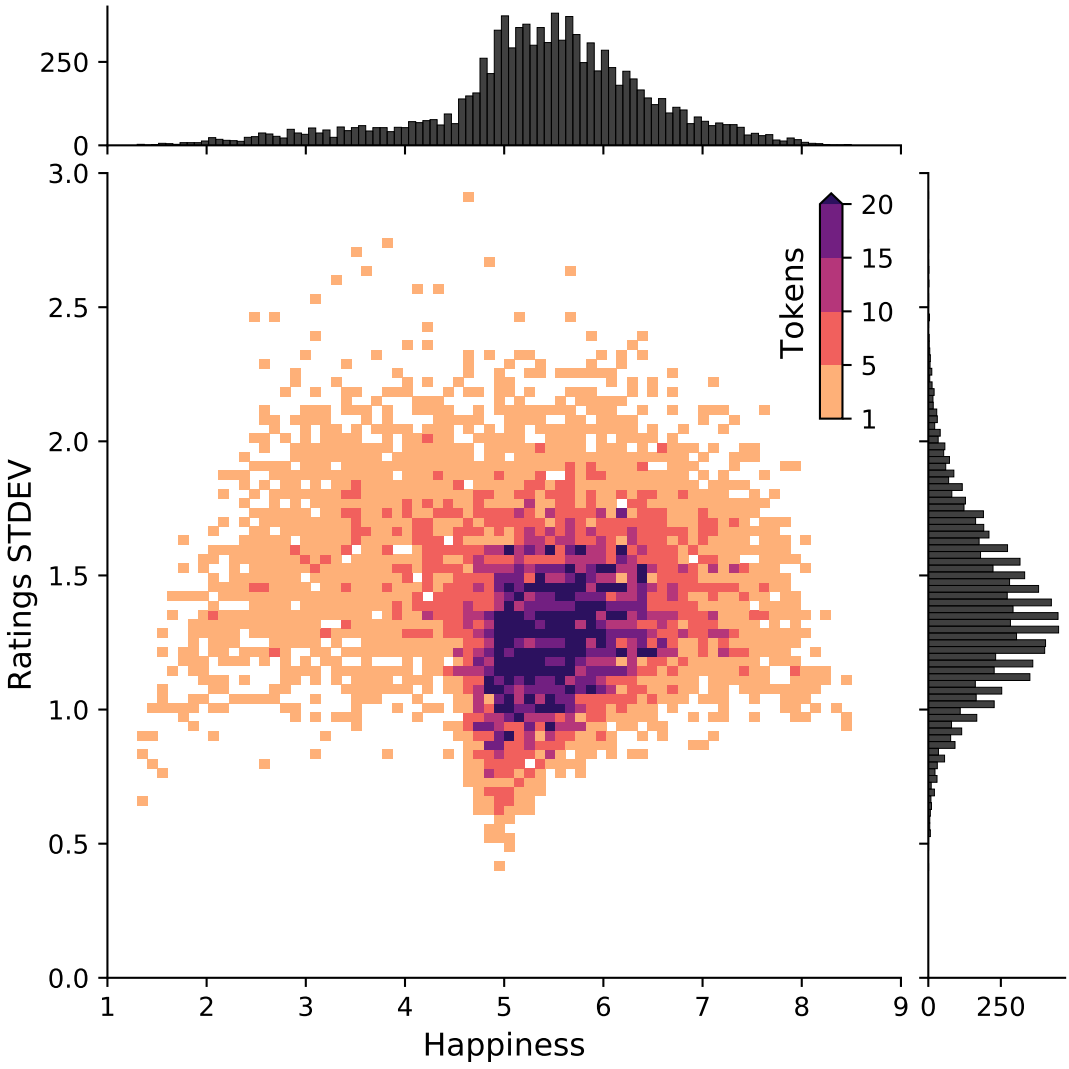

Figure 5.4.1: ***Emotional valence of words and uncertainty in human ratings of lexical polarity.*** *A 2D histogram of happiness h_{avg} and standard deviation of human ratings for each word in the labMT dataset. Happiness is defined on a continuous scale from 1 to 9, where 1 is the least happy and 9 is the most. Words with a score between 4 and 6 are considered neutral. While the vast majority of words are neutral, we still note a positive bias in human language [81]. The average standard deviation of human ratings for estimating the emotional valence of words in the labMT dataset is 1.38.*

On average, the standard deviation of human ratings is 1.38. In our evaluation (Sec. 5.5), we show how our models perform relative to the uncertainty we observe in human ratings.

5.4.2 Token Model

Our first model is a small neural network that learns to map words from the labMT lexicon to their sentiment scores. While still being able to learn a non-linear mapping between the words and their happiness scores, the model only considers the individual words as input—enriching its internal utility function with subword representations to gauge the happiness score.

The input word is first processed into a token embedding—sequentially breaking each word into its equivalent character-level n-grams whereby $n \in \{3, 4, 5\}$ (see Fig. 5.4.2 for a simple illustration of that).

English words have an average length of 5 characters [190, 202], which would yield 5 unique character-level n-grams given our tokenization scheme. While we did try shorter and longer sequences, we fix the length of the input sequence to a size of 50 and pad shorter sequences to ensure a universal input size. We choose a longer sequence length to allow us to encode longer n-grams and rare words.

We then pass the token embeddings to a 300-dimensional embedding layer. We initialize the embedding layer with weights trained with subword information on Common Crawl and Wikipedia using FastText [32]. In particular, we use weights from a pre-trained model using CBOW with character-level n-grams of length 5 and a window size of 5 and 10 (`https://fasttext.cc/docs/en/english-vectors.html`).

The output of the embedding layer is pooled down and passed to a sequence of three dense layers of decreasing sizes: 128, 64, and 32, respectively. We use a rectified linear activation function (ReLU) for all dense layers. We also add a dropout layer after each dense layer, with a 50% dropout rate to encode stochasticity into the model as a simple way for estimating uncertainty and standard deviation of the network's predictions [269].

We experimented with a few different layout configurations, finding that making the network either wider or deeper has minimal effect on the network performance. Therefore, we choose to keep our model rather simple with roughly 10 million trainable parameters.

The output of the last dense layer is finally passed over to a single output layer with a linear activation function to regress a sentiment score between 1 and 9. See Fig. 5.4.3 for a simple diagram of the model architecture.

5.4.3 Dictionary Model

Historically, lexicon-based models have only considered simple statistical methods to estimate the emotional valance of words. Here, we try to bridge the connection between the conventional techniques among the community and recent advances in NLP.

For our second model, we employ a contextualized Transformer-based language model to gauge a sentiment score for a given word based on its dictionary definition. While still predicting scores for individual words, we now do so by augmenting each word with its expressed meaning(s) from a general dictionary.

Given an input word, we look up its lexical definition via a free online dictionary

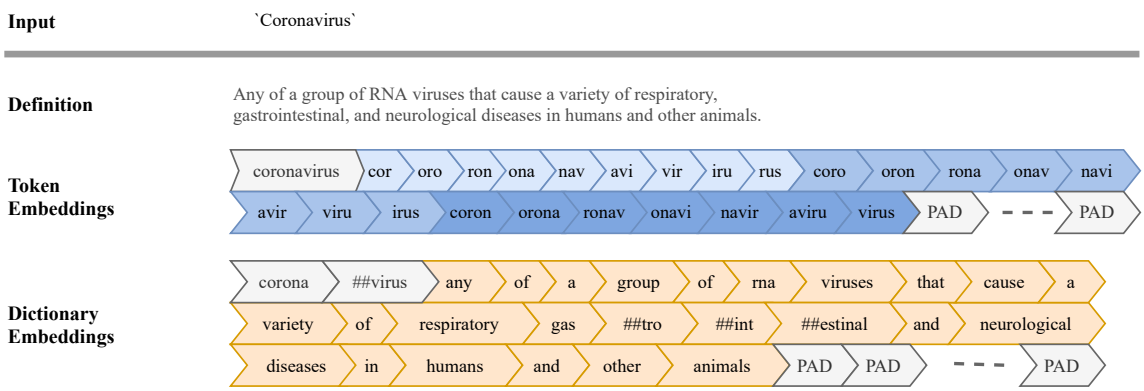

Figure 5.4.2: ***Input sequence embeddings.*** *We use two encoding schemes to prepare input sequences for our models: token embeddings (blue) and dictionary embeddings (orange) for our Token and Dictionary models, respectively. Given an input word (e.g., 'coronavirus'), we first break the input token into character-level n-grams ($n \in \{3, 4, 5\}$). The resulting sequence of n-grams along with the original word at the beginning of the embeddings are used in our Token model. For our Dictionary model, we first look up a dictionary definition for the given input. We then process the input word along with its definition into subwords using WordPiece [310]. Uncommon and novel words are broken into subwords, with double hashtags indicating that the given token is not a full word.*

API available at `https://dictionaryapi.dev`. The average length of definitions for the words found in labMT is roughly 38. We choose 50—which covers the 75th percentile of that distribution—to ensure that words with multiple definitions are adequately represented. While increasing the sequence length beyond 50 did not improve our accuracy, it increases the model complexity slowing our training and inference time substantially. Therefore, we fix the length of word definitions to a maximum of 50 words and pad shorter sequences to ensure a fixed input size.

The word, along with its definition, is processed into dictionary embeddings by breaking each word into subwords based on their frequency of usage using WordPiece [310]. This is a widely adopted tokenization technique that breaks uncommon and novel words into subwords, which reduces the vocabulary size of language models and enables them to handle OOV tokens. Other tokenization models will give

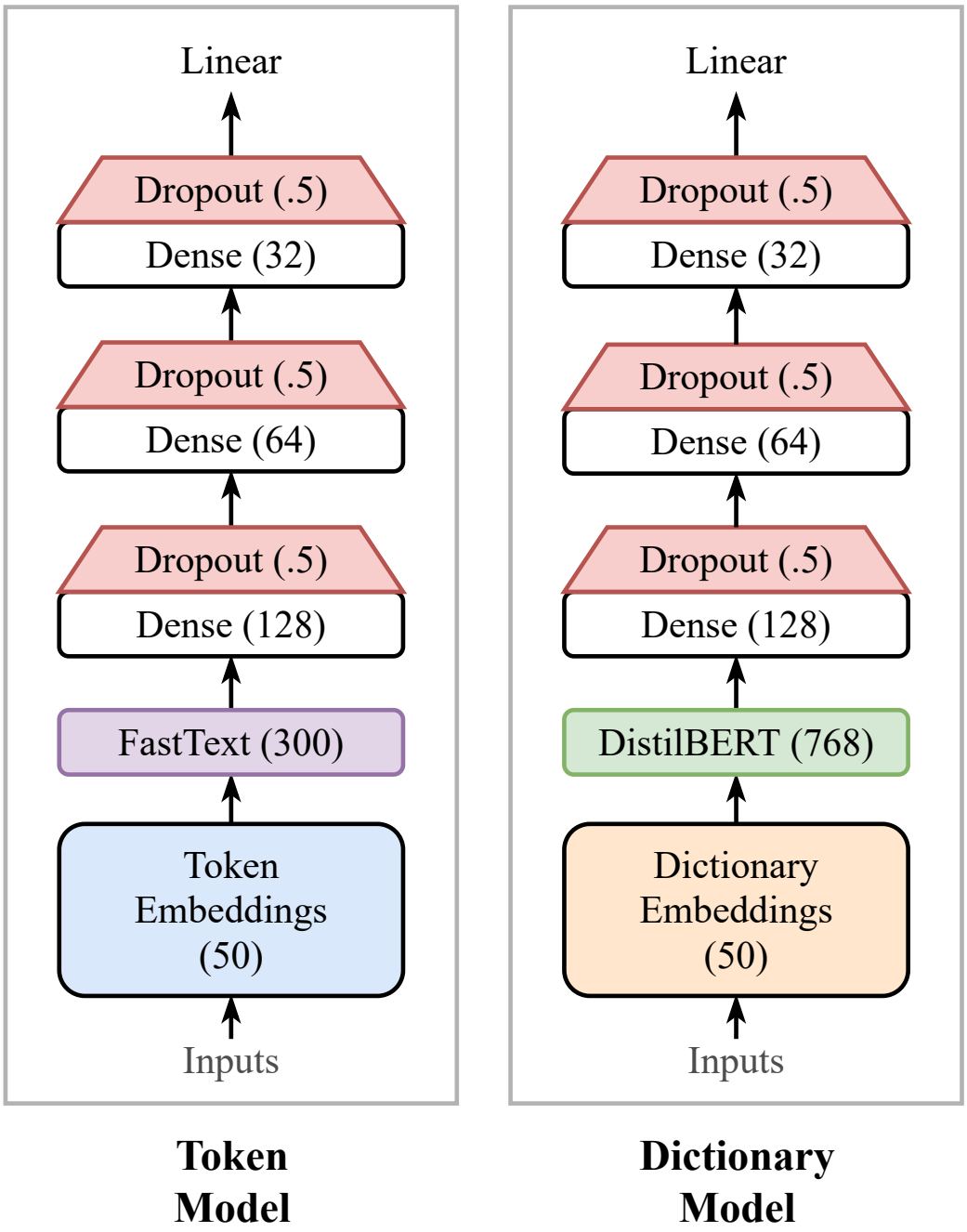

Figure 5.4.3: ***Model architectures.*** *Our first model is a small neural network initialized with pre-trained word embeddings to gauge happiness scores. Our second model, is a deep Transformer-based model that uses word definitions to estimate their sentiment scores. See Sec. 5.4.2 and Sec. 5.4.3 for further technical details of each model, respectively. Note the Token model is considerably smaller with roughly 10 million trainable parameters compared with the Dictionary model that has a little over 66 million parameters.*

similar results [162]. We only use the word as input to our model for terms without definitions.

In principle, the dictionary embeddings can be passed to a vanilla Transformer model (e.g., BERT [73], XLNet [314]). However, we prefer more manageable models (i.e., smaller and faster) due to their efficiency while maintaining state-of-the-art results. We tried both ALBERT [170] and DistilBERT [258]. Both models have equivalent performance on our task.

The output of the model's pooling layer is passed to a sequence of three dense layers of decreasing sizes with dropout applied after each layer—similar to our approach in the Token model. Finally, the output of the last dense layer is projected down to a single output value that servers as the sentiment score prediction.

The Token model is substantially smaller than the Dictionary model. Hence, the Token model is considerably lighter in terms of memory usage, and faster in terms of training and inference time. Our current configuration of the Token model results in roughly 10 million trainable parameters compared with the Dictionary model that has over 66 million parameters.

5.5 Results

5.5.1 Ensemble learning and k-fold cross-validation

Given that our dataset is relatively small, we use k-fold cross-validation rather than a fixed testing subset to set an upper limit on our margin of error and mitigate any risk of overfitting [20, 160]. Using 80/20 split for training/validation, we train our

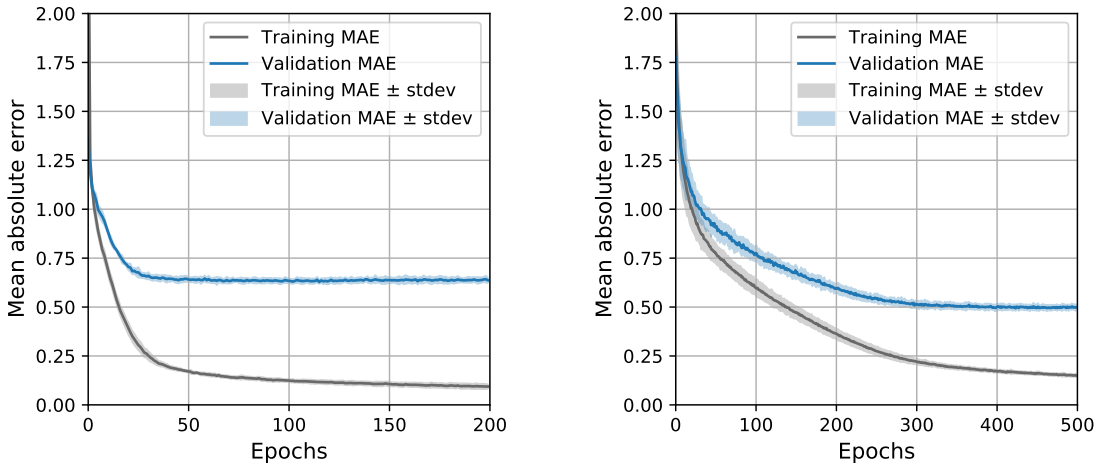

Figure 5.5.1: ***Learning curves for the Token model (left), and Dictionary model (right).*** *We train our models using 5-fold cross-validation, with a maximum of 500 epochs per fold. The left panel shows the learning curves for the Token model (see Sec. 5.4.2), while the right panel shows the Dictionary model (see Sec. 5.4.3). We display our average mean absolute error (MAE) as well as standard deviation across all folds for training (grey) and validation (blue).*

models for a maximum of 500 epochs per fold for a total of 5 folds. While there are many gradient descent optimization algorithms, we use Adam [154] as a popular and well-established optimizer, keeping its default configuration and setting our initial learning rate to 0.001. In Fig. 5.5.1, we display learning curves, showing that both models have converged successfully.

Ensemble learning is a widely known and adopted family of methods in which the average performance of an ensemble is shown to be both less biased and better than the individual models [116, 160]. Capitalizing on our 5-fold cross-validation strategy, we use the model trained from each fold to build an ensemble (see Fig. 5.5.2). To get a happiness score for a given word, we aggregate over 100 predictions per model and report the average and standard deviation of predictions from all models as our final

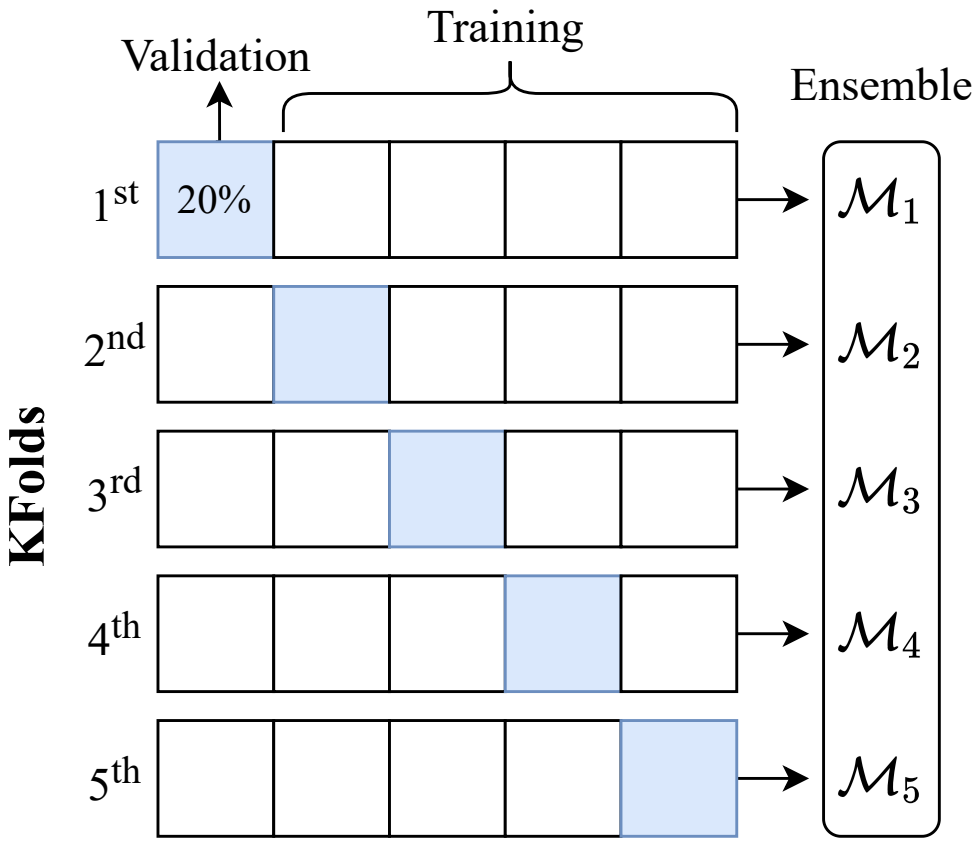

Figure 5.5.2: ***Ensemble learning and k-fold cross-validation.*** *Using an 80/20 split for training/validation, we train our models for a maximum of 500 epochs per fold for a total of 5 folds. We use the model trained from each fold to build an ensemble because the average performance of an ensemble is less biased and better than the individual models.*

prediction for a given ensemble.

5.5.2 Comparing predictions to human ratings

While both strategies—using character-level n-grams and word definitions—perform well, we note that the Dictionary model outperforms the Token model. Our evaluation shows the Token model has an average cross-validation MAE of 0.64 ± 0.01, trailing behind our Dictionary model which has an average cross-validation MAE of 0.50 ± 0.01.

As discussed, our cross-validation defines an upper limit on our margin of error for predicting happiness scores in the labMT dictionary. We further examine our error distributions to investigate if the models have a bias towards high or low happiness scores.

We rerun our models on all words recorded in the labMT dataset. In Figs. 5.5.3 and 5.5.4, we display a breakdown of our MAE distributions for the Token and Dictionary models, respectively. We categorize the happiness scores into three groups: negative ($h_{avg} \in [1, 4)$), neutral ($h_{avg} \in [4, 6]$), and positive ($h_{avg} \in (6, 9]$). While the distributions show our models operate well on all words, particularly neutral expressions, we note a relatively higher MAE for negative words, whereby our predictions to these terms are more positive than the annotations.

We further compare our predictions to the ground-truth ratings, examining the degree to which the models either overshoot or undershoot the happiness scores crowdsourced via AMT. Words in the labMT lexicon were scored by taking the average happiness score of distinct evaluations from 50 different individuals (see Table S2 [81]). Since the variance of human ratings and our model MAEs are on the same scale, we

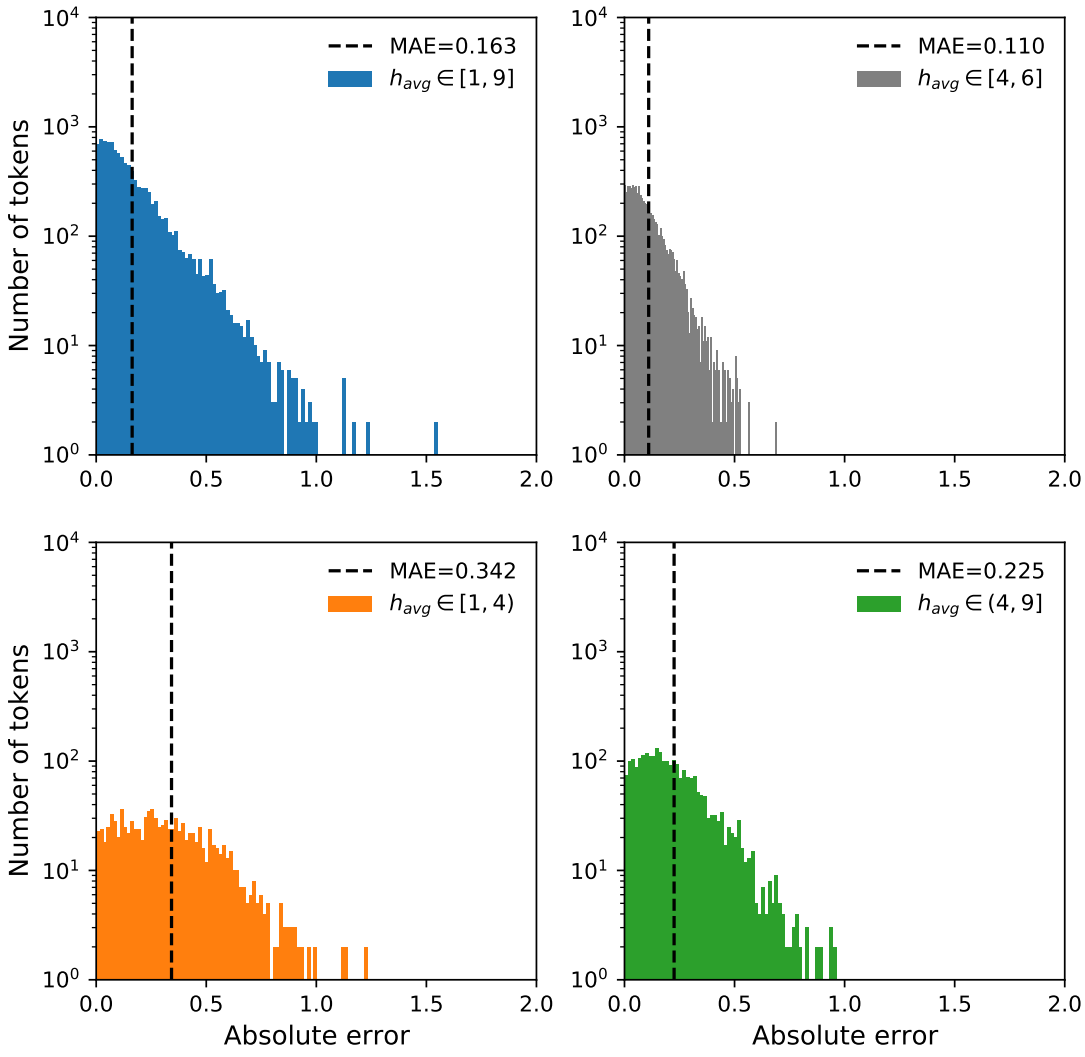

Figure 5.5.3: ***Error distributions for the Token model.*** *We display mean absolute errors for predictions using the Token model on all words in labMT. We arrange the happiness scores into three groups: negative ($h_{avg} \in [1, 4)$, orange), neutral ($h_{avg} \in [4, 6]$, grey), and positive ($h_{avg} \in (6, 9]$, green). Most words have an MAE less than 1 with the exception of a few outliers. We see a relatively higher MAE for negative and positive terms compared to neutral expressions.*

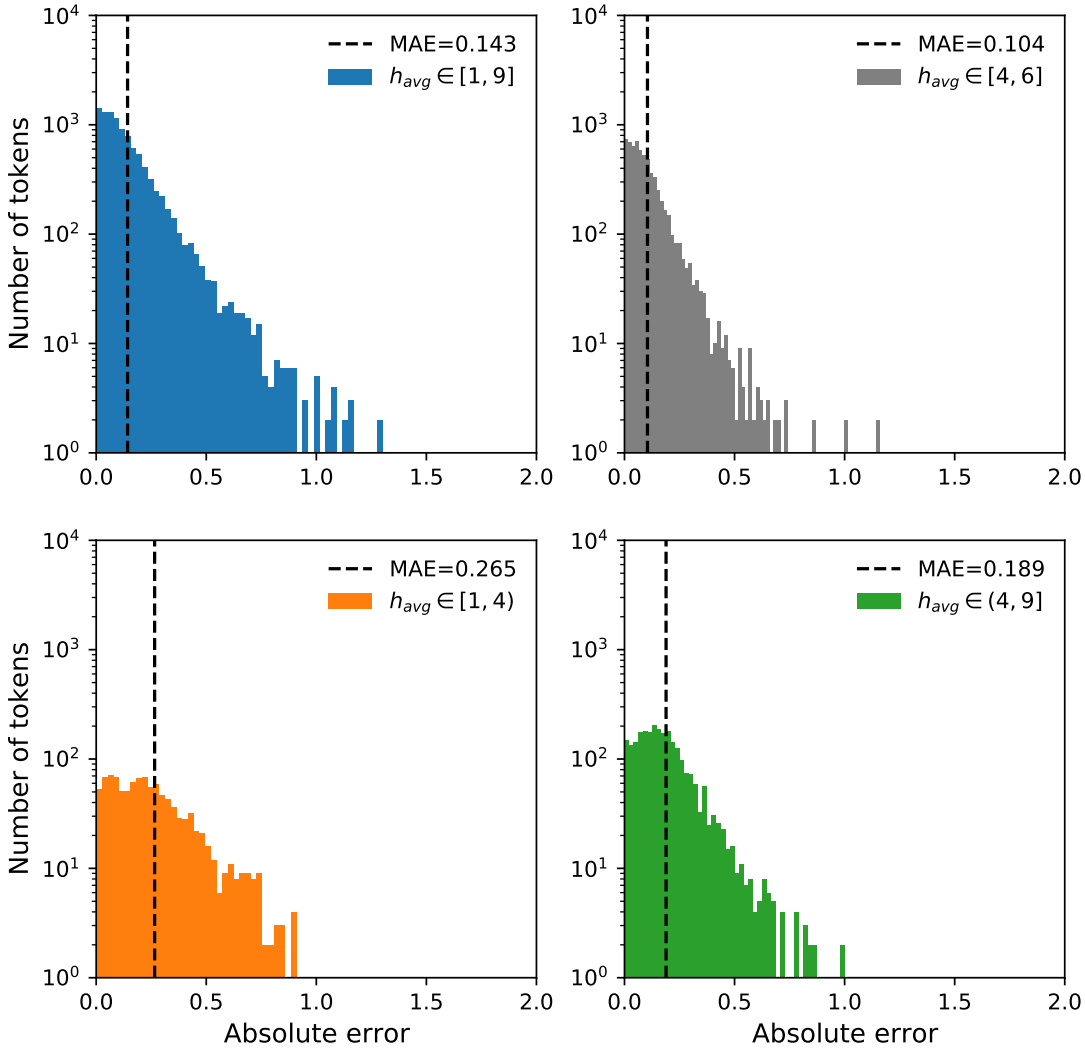

Figure 5.5.4: ***Error distributions for the Dictionary model.*** *We display mean absolute errors for predictions using the Dictionary model on all words in labMT. Again, we categorize the happiness scores into three groups: negative ($h_{avg} \in [1, 4)$, orange), neutral ($h_{avg} \in [4, 6]$, grey), and positive ($h_{avg} \in (6, 9]$, green). Similar to the Token model, most words have an MAE less than 1 with the exception of a few outliers. While the Dictionary model outperforms the Token model, we still observe a higher MAE for negative and positive terms compared to neutral expressions.*

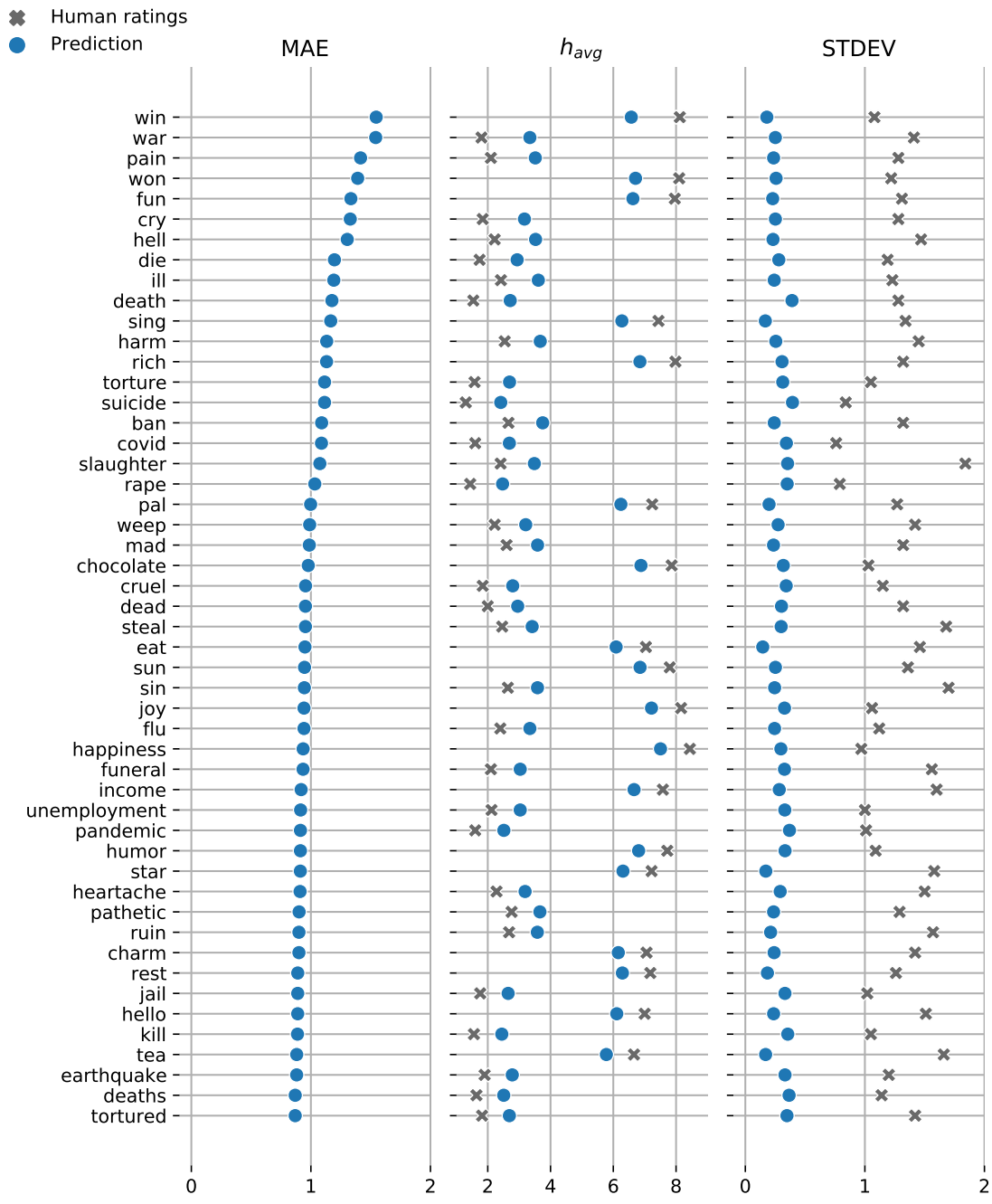

Figure 5.5.5: ***Token model: Top-50 words with the highest mean absolute error.*** *Model predictions are shown in blue and the crowdsourced annotations are displayed in grey. While still maintaining relatively low MAE, most of our predictions are conservative—marginally underestimating words with extremely high happiness scores, and overestimating words with low happiness scores.*

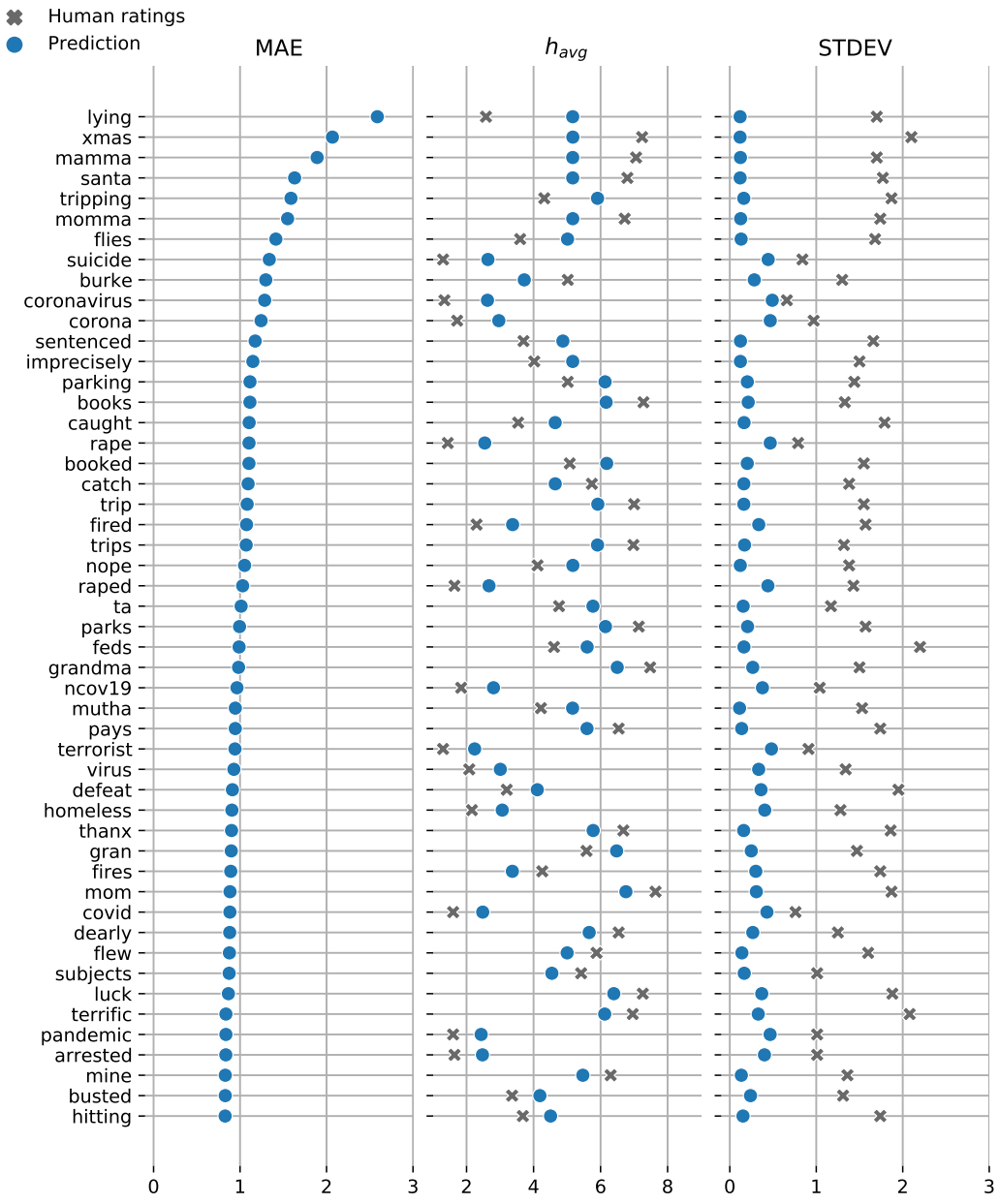

Figure 5.5.6: ***Dictionary model: Top-50 words with the highest mean absolute error.*** *Model predictions are shown in blue and the crowdsourced annotations are displayed in grey. Note, the vast majority of words with relatively high MAE also have high standard deviations of AMT ratings. Words that have multiple definitions will have a neutral score (e.g., lying). A neutral happiness score is also often predicted for words because we are unable to obtain good definitions for them to use as input. Although we have definitions for most words in our dataset, we still have a little over 1500 words with missing definitions. Most of these words are names (e.g., 'Burke'), and slang (e.g., 'xmas', and 'ta').*

	All words		*Out-of-sample*		*All words*	
	STDEV Human Ratings	**Variance** Human Ratings	**MAE** Token Model	**MAE** Dictionary Model	**MAE** Token Model	**MAE** Dictionary Model
Average	1.38	1.17	0.64	0.50	0.16	0.14
25^{th} *Percentile*	1.18	1.09	0.63	0.49	0.05	0.05
50^{th} *Percentile*	1.36	1.17	0.64	0.50	0.12	0.10
75^{th} *Percentile*	1.56	1.25	0.65	0.51	0.22	0.19
85^{th} *Percentile*	1.69	1.30	0.65	0.51	0.30	0.25
95^{th} *Percentile*	1.90	1.38	0.66	0.52	0.48	0.40

Table 5.5.1: We report summary statistics comparing our models to the annotated ratings reported in labMT. Each word in the labMT lexicon is scored by 50 distinct individuals and the final happiness score is derived by taking the average score of these evaluations [81]. We report the standard deviation and variance of the ratings as a baseline to assess the human's confidence in the reported scores. Comparing our predictions with the annotations crowdsourced via AMT, our MAEs are on par with the margin of error we observe in the reported scores in labMT.

can use the observed average variance of the ratings (1.17) as a baseline to assess the human's confidence in the reported scores. Comparing our models to that baseline, we note that all models offer consistent predictions with similar expectations to a random and reliable reviewer from AMT. See Table 5.5.1 for further statistical details.

In Figs. 5.5.5 and 5.5.6, we display the top-50 words with the highest mean absolute error for the Token and Dictionary models, respectively. While the models always predict the right emotional attitude outlining each word based on its lexical polarity, they undershoot scores for happy words, and overshoot scores for sad expressions.

One possible explanation of this systematic behavior is the lack of words with

extreme happiness scores in the labMT lexicon. It is possible to train models with a smaller but balanced subset of the dataset to overcome that challenge. Doing so, however, would reduce the size of training/validation samples substantially. Still, our margin of error is relatively low compared to human ratings. Future investigations may test and improve the models by examining larger sentiment lexicons.

Another key factor that plays a big role in our prediction error is obtaining good word definitions, or the lack thereof, to use as input for our Dictionary model. Surprisingly, outsourcing definitions from online dictionaries for a large set of words is rather challenging, especially if you opt-out of reliable but paid services. In our work, we choose not to use an urban dictionary or any services with paid APIs. We use a completely free online dictionary API that is available online at `https://dictionaryapi.dev`.

While we do have definitions for most words in our dataset, we have a total of 1518 words with missing definitions. Most of these words are names, abbreviations, and slangs (e.g., 'xams', 'foto', 'nvm', and 'lmao'). Words with multiple definitions can also cancel each other's score (e.g., lying).

Notably, the vast majority of words with relatively high MAE also have high standard deviations of AMT ratings. To further investigate our prediction accuracy, we examine the overlap between the predictions and human ratings. In particular, we compute the intersection over union (IOU) between the predicted happiness score $h'_{avg} \pm \sigma'$, and the corresponding value from the annotated ratings $h_{avg} \pm \sigma$.

Using the Token model, we find that our model underestimates the happiness score for 'win'—the only word with a prediction that falls outside the range of human annotated happiness scores. The rest of the predicted happiness scores lie well

within the range of scores crowdsourced via AMT for any given word. Similarly, our Dictionary model slightly underestimates the happiness scores for 'mamma' while overestimating the scores for 'lying', and 'coronavirus'.

5.6 Discussion

As the growing demand for sentiment-aware intelligent systems increases, we will continue to see developments for both lexicon-based models and contextual language models. While contextualized models are suitable for a wide set of applications, lexicon-based models are used by computational linguistics, journalists, and data scientists who are interested in studying how individual words contribute to sentiment trends.

Lexicon dictionaries, however, have to be updated periodically to support new words and expressions that were not considered when the dictionaries were assembled. In this paper, we proposed two models for predicting sentiment scores to augment semantic dictionaries using word embeddings and transfer learning. Our first model establishes a baseline using a shallow neural network initialized with pre-trained word embeddings, while our second model features a deep Transformer-based network that brings into play word definitions to estimate their lexical polarity. Our results and evaluation of both models demonstrate human-level performance on LabMT.

Although both models can predict scores for novel words, we acknowledge a few shortcomings of our strategies. Our Token model relies on subword information to estimate a happiness score for any given word. For example, using subwords for 'coronavirus' yields a good estimate given that it has 'virus' as part of it. By contrast,

parsing character-level n-grams for other words (e.g., 'covid') may not reveal any further information. We can overcome that hurdle by using the word definition as input to our Dictionary model to gauge its happiness score. Words, however, often have different meanings based on context. Finding good definitions may be challenging, especially for slang, informal expressions, and abbreviations. We recommend using the Dictionary model whenever it is possible to outsource a good definition of the word.

A natural next step would be to develop similar models for other languages. There are two possible approaches: building a model for each language or a multilingual model. Fortunately, FastText [32] provides pre-trained word embeddings for over 100 languages. Therefore, it is easy to upgrade the Token model to support other languages. Updating the Dictionary model is also a straightforward task by simply adopting a multilingual Transformer-based model pre-trained with several languages (e.g., Multilingual BERT [73]).

Another vast space of improvements would be to adopt our proposed strategies to develop prediction models for other semantic dictionaries. Researchers can further fine-tune our models to predict other sentiment scores. For example, the happiness scores in the labMT [81] dataset are closely aligned with the valence scores in the NRC-VAD lexicon [206]. We envision future work developing similar models to predict other semantic differentials such as arousal and dominance [206], EPA [213], and SocialSent [115].

More importantly, researchers would need to fine-tune the models using annotated scores for words and expressions in other languages. We caution against translating words and using the same English scores because most words do not have a one-to-one

mapping into other languages, and are often used to express different meanings by the native speakers of any given language. Our primary goal is to provide an easy and robust method to augment semantic dictionaries to empower researchers to maintain and expand them at a relatively low cost using today's state-of-the-art NLP methods.

5.7 Acknowledgments

We are grateful for the computing resources provided by the Vermont Advanced Computing Core and financial support from Google and the Massachusetts Mutual Life Insurance Company. We thank Anne Marie Stupinski and Julia Zimmerman for their insightful discussion and suggestions. Computations were performed on the Vermont Advanced Computing Core supported in part by NSF award No. OAC-1827314.

Chapter 6

Concluding remarks

In the first case study, we presented an alternative approach for language detection and identification of tweets to overcome the challenge of missing language labels in the historical Decahose stream. We have also observed a recent, but growing passive behavior on Twitter leading to an increasing preference for retweets over original tweets for most languages. Further investigations will be needed to shed light on this sociotechnical phenomenon, but possible partial causes might lie in changes of native social amplification mechanisms that may be intrinsically magnifying algorithmic bias on Twitter, and other social media platforms with similar features.

In the second case study, we proposed Storywrangler as an interactive sociotechnical instrument that automatically discovers narratively trending storylines. In building Storywrangler, our goal was to create an evolving platform for the research community to enable future developments in computational linguistics, natural language processing, and data-driven studies of social and behavioral science. The Storywrangler project has already empowered a wide set of interdisciplinary applications—characterizing online collective attention for natural disasters [8]; quantifying language changes surrounding mental health on Twitter [276]; computational timeline reconstruction of the stories surrounding famous individuals [82, 85]; exploring public engagement with the US Presidents [203]; blending search queries with social media data to forecast economic indicators [178]; and examining the resurgence in collective attention toward black victims of fatal police violence in the wake of George Floyd's murder [309].

In the third case study, we also presented an example analysis for using Storywrangler to curate a principled set of day-scale time series of unigrams and bigrams across different languages to track discussions of the COVID-19 pandemic on Twitter.

We showcased example time series plots, including visual comparison with COVID-19 confirmed cases and deaths. In addition to our preliminary analysis, we shared a data repository for current and retrospective investigations to better understand the discourse of the COVID-19 pandemic among vastly diverse societies on social media.

Finally, in the last case study, we proposed a tool for augmenting semantic dictionaries using word embeddings and transfer learning. The framework reduced the need for crowdsourcing scores from human annotators while still providing similar, and often better, results compared with random reviewers from Amazon Mechanical Turk at a fraction of the cost. Although our models can be fine-tuned to predict scores for any semantic lexicon, we focused on predicting happiness scores for the Hedonometer to further improve the instrument's ability to capture the emotional valance of various events on social media.

The previous chapters exemplify the critical value of NLP instruments to study modern sociotechnical systems, particularly social media platforms. While the technical challenges of designing and maintaining such tools are staggering in terms of optimization and scalability, ease of use and open access have been focal to the development of these applications. Developing more sophisticated NLP instruments remains an open challenge to help us understand and disentangle the underlying properties of language in evolving sociotechnical systems.

Bibliography

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, USA, 2016. USENIX Association. ISBN 9781931971331.

[2] J. Abello, P. Broadwell, and T. R. Tangherlini. Computational folkloristics. *Communications of the ACM*, 55(7):60–70, 2012.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, Portland, Oregon, 2011. Association for Computational Linguistics.

[4] J. Allen, B. Howland, M. Mobius, D. Rothschild, and D. J. Watts. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14), 2020.

[5] T. Alshaabi, J. L. Adams, M. V. Arnold, J. R. Minot, D. R. Dewhurst, A. J. Reagan, C. M. Danforth, and P. S. Dodds. Storywrangler: A massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. *Science Advances*, 2021. In press.

[6] T. Alshaabi, M. V. Arnold, J. R. Minot, J. L. Adams, D. R. Dewhurst, A. J. Reagan, R. Muhamad, C. M. Danforth, and P. S. Dodds. How the world's collective attention is being paid to a pandemic: COVID-19 related n-gram time series for 24 languages on Twitter. *Plos One*, 16(1):1–13, 2021.

[7] T. Alshaabi, D. R. Dewhurst, J. R. Minot, M. V. Arnold, J. L. Adams, C. M. Danforth, and P. S. Dodds. The growing amplification of social media: Measur-

ing temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020. *EPJ Data Science*, 10(15), 2021.

[8] M. V. Arnold, D. R. Dewhurst, T. Alshaabi, J. R. Minot, J. L. Adams, C. M. Danforth, and P. S. Dodds. Hurricanes and hashtags: Characterizing online collective attention for natural disasters. *Plos One*, 16(5):e0251762, 2021.

[9] S. Asur and B. A. Huberman. Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 492–499. IEEE, 2010.

[10] Ł. Augustyniak, P. Szymański, T. Kajdanowicz, and W. Tuligłowicz. Comprehensive study on lexicon-based ensemble classification sentiment analysis. *Entropy*, 18(1):4, 2016.

[11] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[12] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur. Opinion mining and sentiment analysis. In *2016 3rd international Conference on Computing for Sustainable Global Development (INDIACom)*, pages 452–455. IEEE, 2016.

[13] D. Bamman, J. Eisenstein, and T. Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

[14] P. Barbera, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542, 2015.

[15] F. Bass. A new product growth model for consumer durables. *Management Science*, 15:215–227, 1969.

[16] K. C. Bathina, M. Ten Thij, L. Lorenzo-Luaces, L. A. Rutter, and J. Bollen. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*, 5(4):458–466, 2021.

[17] B. Batrinca and P. C. Treleaven. Social media analytics: A survey of techniques, tools and platforms. *AI & Society*, 30(1):89–116, 2015.

[18] G. Baxter and I. Sommerville. Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, 23(1):4–17, 2011.

[19] D. Beeferman, W. Brannon, and D. Roy. RadioTalk: A large-scale corpus of talk radio transcripts. In *Proceedings of Interspeech 2019*, pages 564–568. International Speech Communication Association, 2019.

[20] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

[21] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[22] A. I. Bento, T. Nguyen, C. Wing, F. Lozano-Rojas, Y.-Y. Ahn, and K. Simon. Evidence from internet search data shows information-seeking responses to news of local COVID-19 cases. *Proceedings of the National Academy of Sciences*, 117 (21):11220–11222, 2020.

[23] S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson. Language identification for creating language-specific Twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, pages 65–74. Association for Computational Linguistics, 2012.

[24] S. Bergsma, M. Dredze, B. Van Durme, T. Wilson, and D. Yarowsky. Broadly improving user classification via communication-based name and location clustering on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1019, 2013.

[25] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS One*, 10(2):e0118093, 2015.

[26] E. Bevensee, M. Aliapoulios, Q. Dougherty, J. Baumgartner, D. McCoy, and J. Blackburn. SMAT: The social media analysis toolkit. In *Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media*, volume 14, 2020.

[27] P. Block, M. Hoffman, I. J. Raabe, J. B. Dowd, C. Rahal, R. Kashyap, and M. C. Mills. Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world. *Nature Human Behaviour*, pages 1–9, 2020.

[28] S. L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130. Association for Computational Linguistics, 2016.

[29] S. L. Blodgett, J. Wei, and B. O'Connor. A dataset and classifier for recognizing social media English. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61. Association for Computational Linguistics, 2017.

[30] J. Bohannon. Google opens books to new cultural studies. *Science*, 330(6011): 1600–1600, 2010. ISSN 0036-8075.

[31] A. F. Boing, A. C. Boing, J. Cordes, R. Kim, and S. Subramanian. Quantifying and explaining variation in life expectancy at census tract, county, and state levels in the United States. *Proceedings of the National Academy of Sciences*, 117(30):17688–17694, 2020.

[32] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[33] E. Bokányi, D. Kondor, and G. Vattay. Scaling in words on twitter. *Royal Society Open Science*, 6(10):190027, 2019.

[34] J. J. Bolhuis, K. Okanoya, and C. Scharff. Twitter evolution: Converging mechanisms in birdsong and human speech. *Nature Reviews Neuroscience*, 11 (11):747–759, 2010.

[35] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[36] R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[37] H. Borer. *Parametric Syntax: Case Studies in Semitic and Romance Languages*. De Gruyter Mouton, Berlin, Boston, 01 Jan. 1984. ISBN 978-3-11-080850-6.

[38] J. Borge-Holthoefer and Y. Moreno. Absence of influential spreaders in rumor dynamics. *Physical Review E*, 85(2):026116, 2012.

[39] G. Bovasso. A network analysis of social contagion processes in an organizational intervention. *Human Relations*, 49(11):1419–1435, 1996.

[40] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In *2010 43rd Hawaii International Conference on System Sciences*, pages 1–10. IEEE, 2010.

[41] C. O. Buckee, S. Balsari, J. Chan, M. Crosas, F. Dominici, U. Gasser, Y. H. Grad, B. Grenfell, M. E. Halloran, M. U. Kraemer, et al. Aggregated mobility data could help fight COVID-19. *Science (New York, NY)*, 368(6487):145, 2020.

[42] L. Bursztyn, A. Rao, C. Roth, and D. Yanagizawa-Drott. Misinformation during a pandemic. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2020-44), 2020.

[43] L. Cabral and A. Hortacsu. The dynamics of seller reputation: Evidence from eBay. *The Journal of Industrial Economics*, 58(1):54–78, 2010.

[44] D. Caldara and M. Iacoviello. Measuring geopolitical risk. *FRB International Finance Discussion Paper*, (1222), 2018.

[45] H. A. Carneiro and E. Mylonakis. Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases*, 49(10):1557–1564, 2009.

[46] S. Carter, W. Weerkamp, and M. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215, Mar 2013. ISSN 1574-0218.

[47] D. Centola and M. W. Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113:702–734, 2007.

[48] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, 2010.

[49] E. Chen, K. Lerman, and E. Ferrara. Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020.

[50] H. Chen, W. Xu, C. Paris, A. Reeson, and X. Li. Social distance and SARS memory: Impact on the public awareness of 2019 novel coronavirus (COVID-19) outbreak, 2020. Available online at https://www.medrxiv.org/content/10.1101/2020.03.11.20033688v1.

[51] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668. Association for Computational Linguistics, 2017.

[52] Q. Chen, C. Min, W. Zhang, G. Wang, X. Ma, and R. Evans. Unpacking the black box: How to promote citizen engagement through government social media during the COVID-19 crisis. *Computers in Human Behavior*, page 106380, 2020.

[53] E. Chenoweth and M. J. Stephan. Drop your weapons: When and why civil resistance works. *Foreign Affairs*, 93:94, 2014.

[54] H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.

[55] N. A. Christakis and J. H. Fowler. Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine*, 32(4):556–577, 2013.

[56] H. Christenson. Hathitrust: A research library at web scale. *Library Resources & Technical Services*, 55:93–102, 2011.

[57] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012.

[58] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala. The COVID-19 social media infodemic, 2020. Available online at http://arxiv.org/abs/2003.05004.

[59] M. Cipriani and A. Guarino. Herd behavior and contagion in financial markets. *The BE Journal of Theoretical Economics*, 8(1), 2008.

[60] R. Cohen and D. Ruths. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 2013.

[61] E. Colleoni, A. Rozza, and A. Arvidsson. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, 64(2):317–332, 2014.

[62] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116. Association for Computational Linguistics, 2017.

[63] M. Conway and D. O'Connor. Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*, 9: 77–82, 2016.

[64] G. Coppersmith, M. Dredze, and C. Harman. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, 2014.

[65] S. Cottle. Media and the Arab uprisings of 2011. *Journalism*, 12(5):647–659, 2011.

[66] E. Cozzo, R. A. Banos, S. Meloni, and Y. Moreno. Contact-based social contagion in multiplex networks. *Physical Review E*, 88(5):050801, 2013.

[67] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the first workshop on social media analytics*, pages 115–122, 2010.

[68] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, and R. Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988. Association for Computational Linguistics, 2019.

[69] D. J. Daley and D. G. Kendall. Stochastic rumours. *Journal of the Institute of Mathematics and its Applications*, 1:42–55, 1965.

[70] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554–559, 2016.

[71] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, and H. Larson. The pandemic of social media panic travels faster than the COVID-19 outbreak, 2020.

[72] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[73] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[74] T. Dewey, J. Kaden, M. Marks, S. Matsushima, and B. Zhu. The impact of social media on social unrest in the Arab Spring. *International Policy Program*, 5:8, 2012.

[75] D. R. Dewhurst, T. Alshaabi, D. Kiley, M. V. Arnold, J. R. Minot, C. M. Danforth, and P. S. Dodds. The shocklet transform: A decomposition method for the identification of local, mechanism-driven dynamics in sociotechnical time series. *EPJ Data Science*, 9(1):3, 2020.

[76] P. S. Dodds and C. M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4): 441–456, 2010.

[77] P. S. Dodds and D. J. Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92:218701, 2004.

[78] P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232:587–604, 2005.

[79] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS One*, 6(12):1–1, 12 2011.

[80] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *Plos One*, 6(12):e26752, 2011.

[81] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth. Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, 112(8):2389–2394, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1411678112.

[82] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, A. J. Reagan, and C. M. Danforth. Fame and Ultrafame: Measuring and comparing daily levels of 'being talked about' for United States' presidents,

their rivals, God, countries, and K-pop, 2019. Available online at http://arxiv.org/abs/1910.00149.

[83] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, T. J. Gray, M. R. Frank, A. J. Reagan, and C. M. Danforth. Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems, 2020. Available online at http://arxiv.org/abs/2002.09770.

[84] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, D. R. Dewhurst, A. J. Reagan, and C. M. Danforth. Long-term word frequency dynamics derived from Twitter are corrupted: A bespoke approach to detecting and removing pathologies in ensembles of time series, 2020. Available online at https://arxiv.org/abs/2008.11305.

[85] P. S. Dodds, J. R. Minot, M. V. Arnold, T. Alshaabi, J. L. Adams, A. J. Reagan, and C. M. Danforth. Computational timeline reconstruction of the stories surrounding Trump: Story turbulence, narrative control, and collective chronopathy, 2020. Available online at https://arxiv.org/abs/2008.07301.

[86] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 2020.

[87] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, ICRL'21, 2021.

[88] H. Elfardy and M. Diab. Token level identification of linguistic code switching. In *Proceedings of COLING 2012: Posters*, pages 287–296, Mumbai, India, dec 2012. The COLING 2012 Organizing Committee.

[89] N. B. Ellison, J. Vitak, R. Gray, and C. Lampe. Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, 19(4):855–870, 2014.

[90] F. E. Emery and E. L. Trist. Socio-technical systems. management science, models and techniques. *C.W. Churchman & M. Verhurst (Eds), Management Science, Models and Techniques*, pages 83–97, 1960.

[91] J. Fábrega and P. Paredes. Social contagion and cascade behaviors on Twitter. *Information*, 4(2):171–181, 2013.

[92] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[93] T. Fenzl and L. Pelzmann. Psychological and social forces behind aggregate financial market behavior. *Journal of Behavioral Finance*, 13(1):56–65, 2012.

[94] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[95] W. T. Fitch. Empirical approaches to the study of language evolution. *Psychonomic Bulletin & Review*, 24(1):3–33, 2017.

[96] I. Flaounas, O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini. Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style and gender. *Digital journalism*, 1(1):102–116, 2013.

[97] J. W. Forrester. Counterintuitive behavior of social systems. *Theory and Decision*, 2(2):109–140, 1971.

[98] V. Gadde and K. Beykpour. Expanding our policies to further protect the civic conversation. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html, 2020.

[99] V. Gadde and K. Beykpour. Additional steps we're taking ahead of the 2020 us election. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html, 2020.

[100] R. J. Gallagher, M. R. Frank, L. Mitchell, A. J. Schwartz, A. J. Reagan, C. M. Danforth, and P. S. Dodds. Generalized word shift graphs: A method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1):4, 2021.

[101] H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, 2011.

[102] H. Gao, G. Barbier, and R. Goolsby. Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3):10–14, 2011.

[103] F. W. Geels. From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research Policy*, 33(6-7):897–920, 2004.

[104] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457 – 472, 1992.

[105] M. Gerlach and F. Font-Clos. A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22 (1):126, 2020.

[106] A. Giachanou and F. Crestani. Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys*, 49(2), June 2016. ISSN 0360-0300.

[107] A. Giddens. *The Constitution of Society: Outline of the Theory of Structuration*. Outline of the Theory of Structuration. University of California Press, 1984. ISBN 9780520052925.

[108] W. Goffman and V. A. Newill. Generalization of epidemic theory: An application to the transmission of ideas. *Nature*, 204:225–228, 1964.

[109] S. Gohil, S. Vuik, and A. Darzi. Sentiment analysis of health care tweets: Review of the methods used. *JMIR Public Health and Surveillance*, 4(2):e43, 2018.

[110] M. Goldszmidt, M. Najork, and S. Paparizos. Boot-strapping language identifiers for short colloquial postings. In H. Blockeel, K. Kersting, S. Nijssen, and F. Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 95–111, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40991-2.

[111] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[112] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018. URL https://www.aclweb.org/anthology/L18-1550.

[113] L. Grothe, E. W. De Luca, and A. Nürnberger. A comparative study on language identification methods. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[114] J. D. Hamilton and L. C. Hamilton. Models of social contagion. *Journal of Mathematical Sociology*, 8(1):133–160, 1981.

[115] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605. Association for Computational Linguistics, 2016.

[116] L. K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[117] F. M. Harper and J. A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4), 2015. ISSN 2160-6455.

[118] N. Harrigan, P. Achananuparp, and E.-P. Lim. Influentials, novelty, and social contagion: The viral power of average friends, close communities, and old news. *Social Networks*, 34(4):470–480, 2012.

[119] Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[120] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975. ISSN 00905364.

[121] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[122] J. Hirschberg and C. D. Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[123] D. Hirshleifer and S. H. Teoh. Thought and behavior contagion in capital markets. In T. Hens and K. R. Schenk-Hoppé, editors, *Handbook of Financial Markets: Dynamics and Evolution*, Handbooks in Finance, pages 1–56. North-Holland, San Diego, 2009. URL http://www.sciencedirect.com/science/article/pii/B9780123742582500051.

[124] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735.

[125] N. O. Hodas and K. Lerman. How visibility and divided attention constrain social contagion. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing*, pages 249–257. IEEE, 2012.

[126] N. O. Hodas and K. Lerman. The simple rules of social contagion. *Scientific Reports*, 4:4343, 2014.

[127] M. D. Hoffman and A. Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014. ISSN 1532-4435.

[128] L. Hollink, A. Bedjeti, M. van Harmelen, and D. Elliott. A corpus of images and text in online news. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1377–1382, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

[129] G. Hollis and C. Westbury. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, 23(6):1744–1756, 2016.

[130] G. Hollis, C. Westbury, and L. Lefsrud. Extrapolating human judgments from skip-gram vector representations of word meaning. *Quarterly Journal of Experimental Psychology*, 70(8):1603–1619, 2017.

[131] D. Holtz, M. Zhao, S. G. Benzell, C. Y. Cao, M. A. Rahimian, J. Yang, J. Allen, A. Collis, A. Moehring, T. Sowrirajan, D. Ghosh, Y. Zhang, P. S. Dhillon, C. Nicolaides, D. Eckles, and S. Aral. Interdependence and the cost of uncoordinated responses to COVID-19. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.2009522117.

[132] J. Hong, W. Crichton, H. Zhang, D. Y. Fu, J. Ritchie, J. Barenholtz, B. Hannel, X. Yao, M. Murray, G. Moriba, M. Agrawala, and K. Fatahalian. Analyzing who and what appears in a decade of US cable TV news, 2020. Available online at https://arxiv.org/abs/2008.06007.

[133] L. Hong, G. Convertino, and E. Chi. Language matters in Twitter: A large scale study. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[134] S. Hong and D. Nadler. Does the early bird move the polls? The use of the social media tool 'Twitter' by US politicians and its impact on public opinion. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*, pages 182–186, 2011. ISBN 9781450307628.

[135] B. Hughes, T. Baldwin, S. Bird, J. Nicholson, and A. MacKinlay. Reconsidering language identification for written language resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*

(LREC'06), Genoa, Italy, may 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/459_pdf.pdf.

[136] M. Hussain and P. Howard. Democracy's fourth wave? information technologies and the fuzzy causes of the arab spring. *SSRN Electronic Journal*, 57, 03 2012.

[137] D. Hymes. Models of the interaction of language and social setting. *Journal of Social Issues*, 23(2):8–28, 1967.

[138] M. Ienca and E. Vayena. On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature medicine*, 26(4):463–464, 2020.

[139] R. Iyengar, C. Van den Bulte, and T. W. Valente. Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212, 2011.

[140] K. H. Jamieson and D. Albarracin. The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US. *The Harvard Kennedy School Misinformation Review*, 2020.

[141] J. Jiang, E. Chen, S. Yan, K. Lerman, and E. Ferrara. Political polarization drives online conversations about COVID-19 in the United States. *Human Behavior and Emerging Technologies*, 2(3):200–211, 2020.

[142] H. Jin, M. Toyoda, and N. Yoshinaga. Can cross-lingual information cascades be predicted on Twitter? In *International Conference on Social Informatics*, pages 457–472. Springer, 2017.

[143] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. 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*, pages 427–431. Association for Computational Linguistics, 2017.

[144] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. 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*, pages 427–431. Association for Computational Linguistics, 2017.

[145] C. Kaligotla, E. Yücesan, and S. E. Chick. An agent based model of spread of competing rumors through online interactions on social media. In *2015 Winter Simulation Conference (WSC)*, pages 3985–3996. IEEE, 2015.

[146] A. M. Kaplan and M. Haenlein. Users of the world, unite! The challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[147] R. E. Kasperson, O. Renn, P. Slovic, H. S. Brown, J. Emel, R. Goble, J. X. Kasperson, and S. Ratick. The social amplification of risk: A conceptual framework. *Risk analysis*, 8(2):177–187, 1988.

[148] Q. Ke, Y. Ahn, and C. R. Sugimoto. A systematic identification and analysis of scientists on Twitter. *PLOS ONE*, 12(4):1–17, 2017.

[149] Q. Ke, Y.-Y. Ahn, and C. R. Sugimoto. A systematic identification and analysis of scientists on Twitter. *PLoS One*, 12(4):1–17, 04 2017.

[150] M. Kelly and C. O Grada. Market contagion: Evidence from the panics of 1854 and 1857. *American Economic Review*, 90(5):1110–1124, 2000.

[151] V. Kharde and S. Sonawane. Sentiment analysis of Twitter data: A survey of techniques. *International Journal of Computer Applications*, 139(11):5–15, Apr 2016. ISSN 0975-8887.

[152] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre. Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons*, 54(3):241–251, 2011.

[153] S. Kim, I. Weber, L. Wei, and A. Oh. Sociolinguistic analysis of Twitter in multilingual societies. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pages 243–248, 2014.

[154] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations*, 2015.

[155] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. ISSN 0027-8424.

[156] A. Koplenig. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—reconstructing the composition of the German corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1):169–188, 2015. ISSN 2055–7671.

[157] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, 62: 148–158, 2016.

[158] M. U. G. Kraemer, C.-H. Yang, B. Gutierrez, C.-H. Wu, B. Klein, D. M. Pigott, L. du Plessis, N. R. Faria, R. Li, W. P. Hanage, J. S. Brownstein, M. Layan, A. Vespignani, H. Tian, C. Dye, O. G. Pybus, and S. V. Scarpino. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, 2020.

[159] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[160] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation and active learning. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94, pages 231–238, Cambridge, MA, USA, 1994. MIT Press.

[161] Y. Kryvasheyeu, H. Chen, N. Obradovich, E. Moro, P. Van Hentenryck, J. Fowler, and M. Cebrian. Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3):e1500779, 2016.

[162] T. Kudo and J. Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71. Association for Computational Linguistics, 2018.

[163] U. Kursuncu, M. Gaur, U. Lokala, K. Thirunarayan, A. Sheth, and I. B. Arpinar. *Predictive Analysis on Twitter: Techniques and Applications*, pages 67–104. Springer International Publishing, Cham, 2019. ISBN 978-3-319-94105-9.

[164] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, 2010.

[165] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108. IEEE, 2013.

[166] R. J. Ladle, R. A. Correia, Y. Do, G.-J. Joo, A. C. Malhado, R. Proulx, J.-M. Roberge, and P. Jepson. Conservation culturomics. *Frontiers in Ecology and the Environment*, 14(5):269–275, 2016.

[167] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. In *2010 2nd International Workshop on Cognitive Information Processing*, pages 411–416. Institute of Electrical and Electronics Engineers, 2010.

[168] V. Lampos and N. Cristianini. Tracking the flu pandemic by monitoring the social web. pages 411–416, 07 2010.

[169] V. Lampos, S. Moura, E. Yom-Tov, I. J. Cox, R. McKendry, and M. Edelstein. Tracking COVID-19 using online search, 2020. Available online at http://arxiv.org/abs/2003.08086.

[170] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations*, 2020.

[171] M. Laver, K. Benoit, and J. Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, pages 311–331, 2003.

[172] K. Lee, L. He, M. Lewis, and L. Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197. Association for Computational Linguistics, 2017.

[173] K. Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 2011.

[174] K. Lerman and R. Ghosh. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[175] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T.-L. Gao, W. Duan, K. K.-f. Tsoi, and F.-Y. Wang. Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo. *IEEE Transactions on Computational Social Systems*, 7(2):556–562, 2020.

[176] M. Li, Q. Lu, Y. Long, and L. Gui. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456, 2017.

[177] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2). *Science*, 2020.

[178] Y. Li, A. Ahani, H. Zhan, K. Foley, T. Alshaabi, K. Linnell, P. S. Dodds, C. M. Danforth, and A. Fox. Blending search queries with social media data to improve forecasts of economic indicators, 2021. Available online at https://arxiv.org/abs/2107.06096.

[179] W. Liu and D. Ruths. What's in a name? Using first names as features for gender inference in Twitter. In *AAAI Spring Symposium: Analyzing Microtext*, volume SS-13-01 of *AAAI Technical Report*. AAAI, 2013.

[180] E. Loper and S. Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.

[181] L. López and X. Rodó. The end of social confinement and COVID-19 re-emergence risk. *Nature Human Behaviour*, 4(7):746–755, 2020.

[182] M. Lui and T. Baldwin. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand, nov 2011. Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I11-1062.

[183] M. Lui and T. Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30. Association for Computational Linguistics, 2012.

[184] M. Lui and T. Baldwin. Accurate language identification of Twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25. Association for Computational Linguistics, 2014.

[185] M. Lui, J. H. Lau, and T. Baldwin. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40, 2014.

[186] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics, 2015.

[187] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics, 2011.

[188] M. Malik, H. Lamba, C. Nakos, and J. Pfeffer. Population bias in geotagged tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.

[189] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. ISBN 9780262303798.

[190] M. S. Mayzner and M. E. Tresselt. Tables of single-letter and digram frequency counts for various word-length and letter-position combinations. *Psychonomic Monograph Supplements*, 1965.

[191] P. McNamee. Language identification: A solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101, 2005. ISSN 1937-4771.

[192] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[193] J. Mellon and C. Prosser. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3):2053168017720008, 2017.

[194] J. Mellon and C. Prosser. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3):2053168017720008, 2017.

[195] J. Merritt and S. Niequist. *Learning to Speak God from Scratch: Why Sacred Words Are Vanishing–and How We Can Revive Them*. Crown Publishing Group, 2018. ISBN 9781601429315.

[196] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[197] J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. A. Lieberman. Quantitative analysis of culture using millions of digitized books. *Science Magazine*, 331:176–182, 2011.

[198] W. Mieder. *Proverbs: A Handbook*. Greenwood folklore handbooks. Greenwood Press, 2004. ISBN 9780313326981.

[199] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Y. Bengio and Y. LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[200] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

[201] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[202] G. A. Miller, E. B. Newman, and E. A. Friedman. Length-frequency statistics for written English. *Information and Control*, 1(4):370–389, 1958.

[203] J. R. Minot, M. V. Arnold, T. Alshaabi, C. M. Danforth, and P. S. Dodds. Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter. *Plos One*, 16(4):e0248880, 2021.

[204] G. Mishne, N. S. Glance, et al. Predicting movie sales from blogger sentiment. In *AAAI spring symposium: computational approaches to analyzing weblogs*, pages 155–158, 2006.

[205] A. Mitchell and P. Hitlin. Twitter reaction to events often at odds with overall public opinion. *Pew Research Center: Internet, Science & Tech*, 2019. URL https://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/.

[206] S. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184. Association for Computational Linguistics, 2018.

[207] B. Mønsted, P. Sapieżyński, E. Ferrara, and S. Lehmann. Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PloS One*, 12(9):e0184148, 2017.

[208] M. Nagarajan, H. Purohit, and A. Sheth. A qualitative examination of topical tweet and retweet practices. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, 2010.

[209] T. Nasukawa and J. Yi. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, pages 70–77, 2003.

[210] D. Nguyen, D. Trieschnigg, and L. Cornips. Audience and the use of minority languages on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.

[211] S. V. Nuti, B. Wayda, I. Ranasinghe, S. Wang, R. P. Dreyer, S. I. Chen, and K. Murugiah. The use of google trends in health care research: A systematic review. *PloS one*, 9(10):e109583, 2014.

[212] G. S. O'Keeffe, K. Clarke-Pearson, et al. The impact of social media on children, adolescents, and families. *Pediatrics*, 127(4):800–804, 2011.

[213] C. E. Osgood. Studies on the generality of affective meaning systems. *American Psychologist*, 17(1):10, 1962.

[214] P. Ozturk, H. Li, and Y. Sakamoto. Combating rumor spread on social media: The effectiveness of refutation and warning. In *2015 48th Hawaii International Conference on System Sciences*, pages 2406–2414. IEEE, 2015.

[215] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[216] L. Palen and K. M. Anderson. Crisis informatics new data for extraordinary times. *Science*, 353(6296):224–225, 2016.

[217] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008. ISSN 1554-0669.

[218] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[219] A. V. Papachristos, C. Wildeman, and E. Roberto. Tragic, but not random: The social contagion of nonfatal gunshot injuries. *Social Science & Medicine*, 125:139–150, 2015.

[220] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[221] E. A. Pechenick, C. M. Danforth, and P. S. Dodds. Characterizing the google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*, 10(10):1–24, 2015.

[222] E. A. Pechenick, C. M. Danforth, and P. S. Dodds. Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not. *Journal of Computational Science*, 21:24–37, 2017. ISSN 1877-7503.

[223] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.

[224] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020.

[225] M. Peters, W. Ammar, C. Bhagavatula, and R. Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765. Association for Computational Linguistics, 2017.

[226] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics, 2018.

[227] J. Pfeffer, K. Mayer, and F. Morstatter. Tampering with Twitter's sample API. *EPJ Data Science*, 7(1):50, 2018.

[228] A. Phillips and M. Davis. Best current practice (BCP): Tags for identifying languages. Technical report, Network Working Group IETF, California, USA, Technical report, 2009.

[229] S. T. Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, 2014.

[230] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland. Time-critical social mobilization. *Science*, 334(6055):509–512, 2011.

[231] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland. Time-critical social mobilization. *Science*, 334(6055):509–512, 2011.

[232] F. Pla and L.-F. Hurtado. Language identification of multilingual posts from Twitter: A case study. *Knowledge and Information Systems*, 51(3):965–989, 2017.

[233] C. E. Pollack, P. R. Soulos, J. Herrin, X. Xu, N. A. Christakis, H. P. Forman, J. B. Yu, B. K. Killelea, S.-Y. Wang, and C. P. Gross. The impact of social contagion on physician adoption of advanced imaging tests in breast cancer. *Journal of the National Cancer Institute*, 109(8):djw330, 2017.

[234] D. M. W. Powers. Applications and explanations of Zipf's law. In *New Methods in Language Processing and Computational Natural Language Learning*, 1998. URL https://www.aclweb.org/anthology/W98-1218.

[235] D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. Studying user income through language, behaviour and affect in social media. *PLOS ONE*, 10(9):1–17, 2015.

[236] D. D. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[237] R. B. Primack, H. Higuchi, and A. J. Miller-Rushing. The impact of climate change on cherry trees and other species in Japan. *Biological Conservation*, 142 (9):1943–1949, 2009. ISSN 0006-3207. The Conservation and Management of Biodiversity in Japan.

[238] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

[239] H. R. Rao, N. Vemprala, P. Akello, and R. Valecha. Retweets of officials' alarming vs reassuring messages during the COVID–19 pandemic: Implications for crisis management. *International Journal of Information Management*, page 102187, 2020.

[240] A. J. Reagan, C. M. Danforth, B. Tivnan, J. R. Williams, and P. S. Dodds. Sentiment analysis methods for understanding large-scale texts: A case for using continuum-scored words and word shift graphs. *EPJ Data Science*, 6: 1–21, 2017.

[241] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):1–29, 2016.

[242] S. Rijhwani, R. Sequiera, M. Choudhury, K. Bali, and C. Maddila. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. pages 1971–1982, 01 2017.

[243] E. Riloff. An empirical study of automated dictionary construction for information extraction in three domains. *Artificial Intelligence*, 85(1-2):101–134, 1996.

[244] H. Ringbom. *Cross-linguistic Similarity in Foreign Language Learning*. Multilingual Matters, Bristol, 2006. ISBN 978-1-85359-936-1.

[245] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics, 2011.

[246] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1104–1112, 2012. ISBN 9781450314626.

[247] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, pages 695–704, 2011.

[248] A. Roomann-Kurrik. Introducing new metadata for tweets. https://blog.twitter.com/developer/en_us/a/2013/introducing-new-metadata-for-tweets.html, 2013.

[249] A. Rosen. Tweeting made easier. https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html, 2017.

[250] Y. Roth and A. Achuthan. Building rules in public: Our approach to synthetic & manipulated media. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html, 2020.

[251] Y. Roth and N. Pickles. Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html, 2020.

[252] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986. ISBN 026268053X.

[253] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, 2010. ISBN 9781605587998.

[254] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860, 2010. ISBN 9781605587998.

[255] A. Samoilenko, F. Karimi, D. Edler, J. Kunegis, and M. Strohmaier. Linguistic neighbourhoods: Explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ Data Science*, 5(1):9, 2016.

[256] C. S. Sanders Peirce. Prolegomena to an Apology for Pragmaticism. *The Monist*, 16(4):492–546, 2015. ISSN 0026-9662.

[257] E. Sandhaus. The New York Times Annotated Corpus, 2008.

[258] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing. MIT Press, 2019.

[259] T. C. Schelling. Dynamic models of segregation. *Journal of Mathematical Sociology*, 1:143–186, 1971.

[260] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[261] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics, 2016.

[262] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1):1–9, 2018.

[263] Y. Shimshoni, N. Efron, and Y. Matias. On the predictability of search trends. 2009.

[264] C. Shu. Twitter officially launches its "retweet with comment" feature. https://techcrunch.com/2015/04/06/retweetception/, 2015.

[265] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3-4): 425–440, 1955. ISSN 0006-3444.

[266] B. Snyder and R. Barzilay. Multiple aspect ranking using the good grief algorithm. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 300–307. Association for Computational Linguistics, 2007.

[267] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics, 2013.

[268] D. Spohr. Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, 34(3):150–160, 2017.

[269] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[270] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler. Online social networks and offline protest. *EPJ Data Science*, 4(1):19, 2015.

[271] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler. Online social networks and offline protest. *EPJ Data Science*, 4(1):19, 2015.

[272] K. Steinmetz. What Twitter says to linguists, Sep 2013. URL http://content.time.com/time/subscriber/article/0,33009,2150609,00.html.

[273] S. Stieglitz and L. Dang-Xuan. Political communication and influence through microblogging—An empirical analysis of sentiment in Twitter messages and retweet behavior. In *2012 45th Hawaii International Conference on System Sciences*, pages 3500–3509. IEEE, 2012.

[274] B. Stone. Are you twittering @ me? https://blog.twitter.com/official/en_us/a/2007/are-you-twittering-me.html, 2007. URL https://blog.twitter.com/official/en_us/a/2007/are-you-twittering-me.html.

[275] B. Stone. Retweet limited rollout. https://blog.twitter.com/official/en_us/a/2009/retweet-limited-rollout.html, 2009.

[276] A. M. Stupinski, T. Alshaabi, M. V. Arnold, J. L. Adams, J. R. Minot, M. Price, P. S. Dodds, and C. M. Danforth. Quantifying language changes surrounding mental health on Twitter, 2021. Available online at https://arxiv.org/abs/2106.01481.

[277] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *2010 IEEE Second International Conference on Social Computing*, pages 177–184. IEEE, 2010.

[278] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2):267–307, 2011. ISSN 0891-2017. doi: 10.1162/COLI_a_00049.

[279] H. Tang, S. Tan, and X. Cheng. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7):10760–10773, 2009.

[280] T. R. Tangherlini and P. Leonard. Trawling in the sea of the Great Unread: Sub-corpus topic modeling and Humanities research. *Poetics*, 41(6):725–749, 2013. ISSN 0304-422X.

[281] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter. PHOAKS: A system for sharing recommendations. *Communications of the ACM*, 40(3):59–62, 1997.

[282] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[283] M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335. Association for Computational Linguistics, 2006.

[284] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso. The computational limits of deep learning, 2020. Available online at https://arxiv.org/abs/2007.05558.

[285] P. Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS One*, 13(9):e0203958, 2018.

[286] E. L. Trist. *The evolution of socio-technical systems*, volume 2. Ontario Quality of Working Life Centre Toronto, 1981.

[287] E. L. Trist and K. W. Bamforth. Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human Relations*, 4(1):3–38, 1951.

[288] E. Tromp and M. Pechenizkiy. Graph-based N-gram language identification on short texts. *Proceedings of Benelearn 2011*, pages 27–34, 01 2011.

[289] M. Trusov, R. E. Bucklin, and K. Pauwels. Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *Journal of Marketing*, 73(5):90–102, 2009.

[290] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, 2010.

[291] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 417–424. Association for Computational Linguistics, 2002.

[292] Twitter. Developer application program interface (API). https://developer.twitter.com/en/docs/ads/campaign-management/api-reference, 2019.

[293] Twitter. Rules and filtering. https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/premium-operators, 2019.

[294] Twitter. Tweet geospatial metadata. https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata, 2019.

[295] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16): 5962–5966, 2012.

[296] J. J. Van Bavel, K. Baicker, P. S. Boggio, V. Capraro, A. Cichocka, M. Cikara, M. J. Crockett, A. J. Crum, K. M. Douglas, J. N. Druckman, et al. Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, pages 1–12, 2020.

[297] C. Van den Bulte and Y. V. Joshi. New product diffusion with influentials and imitators. *Marketing Science*, 26(3):400–421, 2007.

[298] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[299] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez. Sentiment analysis on monolingual, multilingual and code-switching Twitter corpora. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–8. Association for Computational Linguistics, 2015.

[300] Q.-H. Vuong, Q.-K. Bui, V.-P. La, T.-T. Vuong, V.-H. T. Nguyen, M.-T. Ho, H.-K. T. Nguyen, and M.-T. Ho. Cultural additivity: Behavioural insights from the interaction of Confucianism, Buddhism and Taoism in folktales. *Palgrave Communications*, 4(1):1–15, 2018.

[301] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Nature Scientific Reports*, 2:335, 2012.

[302] J. Williams and C. Dagli. Twitter language identification of similar languages and dialects without ground truth. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 73–83. Association for Computational Linguistics, 2017.

[303] J. R. Williams, P. R. Lessard, S. Desu, E. M. Clark, J. P. Bagrow, C. M. Danforth, and P. S. Dodds. Zipf's law holds for phrases, not words. *Nature Scientific Reports*, 5:12209, 2015.

[304] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics, 2005.

[305] S. Wojcik and A. Hughes. How Twitter users compare to the general public. *Pew Research Center: Internet, Science & Tech*, 2019. URL https://www.pewinternet.org/2019/04/24/sizing-up-twitter-users/.

[306] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020.

[307] G. Wolfsfeld, E. Segev, and T. Sheafer. Social media and the Arab Spring: Politics comes first. *The International Journal of Press/Politics*, 18(2):115–137, 2013.

[308] J. T. Woolley and G. Peters. The American presidency project, 1999. Available online at http://www.presidency.ucsb.edu/.

[309] H. Wu, R. J. Gallagher, T. Alshaabi, J. L. Adams, J. R. Minot, M. V. Arnold, B. F. Welles, R. Harp, P. S. Dodds, and C. M. Danforth. Say their names: Resurgence in collective attention toward Black victims of fatal police violence following the death of George Floyd, 2021. Available online at https://arxiv.org/abs/2106.10281.

[310] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. Available online at https://arxiv.org/abs/1609.08144.

[311] B. Xu, B. Gutierrez, S. Mekaru, K. Sewalk, L. Goodwin, A. Loskill, E. L. Cohn, Y. Hswen, S. C. Hill, M. M. Cobo, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. *Scientific data*, 7(1):1–6, 2020.

[312] B. Xu, M. U. Kraemer, B. Gutierrez, S. Mekaru, K. Sewalk, A. Loskill, L. Wang, E. Cohn, S. Hill, A. Zarebski, et al. Open access epidemiological data from the COVID-19 outbreak. *The Lancet Infectious Diseases*, 2020.

[313] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane. Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR)*, 50(2):1–33, 2017.

[314] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[315] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

[316] A. Younus, M. A. Qureshi, F. F. Asar, M. Azam, M. Saeed, and N. Touheed. What do the average twitterers say: A Twitter model for public opinion analysis in the face of major political events. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 618–623. Institute of Electrical and Electronics Engineers, 2011.

[317] A. Z. Yu, S. Ronen, K. Hu, T. Lu, and C. A. Hidalgo. Pantheon 1.0, a manually verified dataset of globally famous biographies. *Scientific Data*, 3(1):1–16, 2016.

[318] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern. Predicting information spreading in Twitter. In *Workshop on Computational Social Science and the Wisdom of Crowds, NIPS*, volume 104, pages 17599–601. Citeseer, 2010.

[319] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[320] X. Zheng, J. Han, and A. Sun. A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671, 2018.

[321] G. K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA, 1949.

[322] G. K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA, 1949.

[323] A. Zubiaga, D. Spina, R. Martínez, and V. Fresno. Real-time classification of Twitter trends. *Journal of the Association for Information Science and Technology*, 66(3):462–473, 2015.

[324] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS One*, 11(3):e0150989, 2016.

[325] A. Zubiaga, I. San Vicente, P. Gamallo, J. R. Pichel, I. Alegria, N. Aranberri, A. Ezeiza, and V. Fresno. Tweetlid: A benchmark for tweet language identification. *Language Resources and Evaluation*, 50(4):729–766, 2016.