Alshaabi et al. *EPJ Data Science* (2021) 10:15
https://doi.org/10.1140/epjds/s13688-021-00271-0

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020

Thayer Alshaabi[1,2,3*], David Rushing Dewhurst[1,2,4], Joshua R. Minot[1,2], Michael V. Arnold[1,2], Jane L. Adams[1,2], Christopher M. Danforth[1,2,5] and Peter Sheridan Dodds[1,2,3]

*Correspondence: thayer.alshaabi@uvm.edu
[1]Vermont Complex Systems Center, University of Vermont, Burlington, VT 05405, USA
[2]Computational Story Lab, University of Vermont, Burlington, VT 05405, USA
Full list of author information is available at the end of the article

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.

Keywords: NLP; Sociolinguistics; Social contagion; Twitter; Signal processing

1 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 [1–4].

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 [1, 2, 5, 6]. 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 [7–16], 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.

1.1 Background and motivation

Social contagion has been extensively studied across many disciplines including marketing [17–20], finance [21–24], sociology [25–27], and medicine [28–30]. 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 [31–34]. 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 [6, 35], with attention paid to the spreading of certain kinds of messages, such as rumours [36–40], misinformation and "fake news" [41–44]. Several models have also been proposed to predict the spread of information on Twitter [45], while other models have shown the differences in which various topics can propagate throughout social networks [46, 47]. Studies have also investigated the extent to which information spread on Twitter can have an echo chamber effect [48–50].

The body of research shows overwhelming evidence that retweeting is a key instrument of social contagion on Twitter [3, 51]. One of the earliest analysis of Twitter by Kwak *et al.* [52] 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 [2]. 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 [5]. Recent work shows how retweets of officials can be either alarming or reassuring amid the COVID–19 pandemic [53, 54]. Statistical features of retweets reveal a strong association between links and hashtags in most retweeted messages [1]. 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 [2]. The use of retweets empirically alters the visibility of information and how fast messages can spread on the platform [4].

Other studies have quantified language usage on social media [55, 56], particularly on Twitter [57, 58]. While investigators have studied the use of retweets in the context of social contagion using network-based approaches [35, 46, 54, 59], little research has been done regarding the statistical variability of retweets across the vast majority of languages. In this paper, 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.

1.2 Overview

We structure our paper 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 [60] 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 [61].

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.* [62]).

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 A. We conclude with an analytical validation of our contagion ratios (Appendix B), and the impact of tweet-length on language detection (Appendix C). We also provide an online appendix at: http://compstorylab.org/storywrangler/papers/tlid/.

2 Tweet language identification

Twitter is a well-structured streaming source of sociotechnical data, allowing for the study of dynamical linguistics and cultural phenomena [63, 64]. 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 [65–68]. Researchers have nevertheless shown that Twitter's collective conversation mirrors the dynamics of local and global events [69] including earthquakes [70], flu and influenza [71, 72], crowdsourcing and disaster relief [73, 74], major political affairs [75], and fame dynamics for political figures and celebrities [76]. Moreover, analyses of social media data and digital text corpora over the last decade have advanced natural language processing (NLP) research [77–79] such as language detection [80–83], sentiment analysis [84–88], word embeddings [89–92], and machine translation [93–95].

LID is often referred to as a solved problem in NLP research [96–100], 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 [101].

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 [81–83]. 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 [102]. 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 on the platform over time.

2.1 Open-source tools for LID

Several studies have looked closely at language identification and detection for short-text [103–110], particularly on Twitter where users are limited to a few characters per tweet (140 prior to the last few months of 2017, 280 thereafter [111]). Researchers have outlined common challenges specific to this platform [112, 113].

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 [81, 82] (see Appendix 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,

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

and multilingual posts on Twitter makes language detection much more difficult compared to LID of formal documents [114]. 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 [81, 115, 116].

Researchers have evaluated off-the-shelf LID tools on substantial subsets of Twitter data for a limited number of languages [81, 82, 116]. 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 [80] 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 *et al.* [82] 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.* [90] 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 model to carry out various NLP tasks such as LID.

FastText [60] is a recently proposed approach for text classification that uses n-gram features similar to the model described by Mikolov *et al.* [117]. FastText employs various tricks [91, 92, 118] 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 [119] that uses the word embeddings produced by the model. The proposed tool uses a hierarchical softmax function [60, 117] 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 [119] as FastText-LID throughout the paper. The authors show that FastText-LID is on par with deep learning models [120, 121] in terms of accuracy and consistency, yet orders of magnitude faster in terms of inference and training time [60]. They also show that FastText-LID outperforms previously introduced LID tools such as langid.[4]

[2] http://code.google.com/p/chromium-compact-language-detector/

[3] https://github.com/CLD2Owners/cld2

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

2.2 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 [119] due to its consistent and reliable performance in terms of inference time and accuracy.

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 *et al.* [103] to improve language classification [82, 122]. 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 [123, 124]. We provide a list of all language labels assigned by FastText-LID compared to the ones served by Twitter-LID in Table 1.

We subsequently extract day-scale time series and Zipf distributions for uni-, bi-, 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.* [125] for technical details about our project.

3 Results and discussion

3.1 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 Sect. 2.2, we have implemented a simple pipeline to preprocess messages and obtain language labels using FastText-LID [119]. 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 in Dodds *et al.* [126]).

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. 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

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

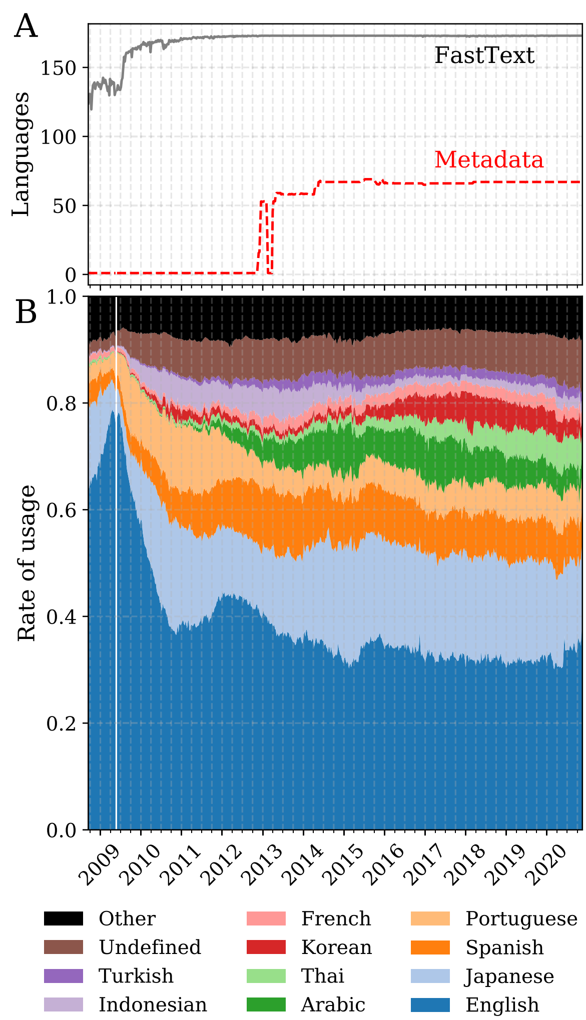

Figure 1 Language time series for the Twitter historical feed and FastText-LID classified tweets. (**A**) Number of languages reported by Twitter-LID (red) and classified by FastText-LID (black) since September 2008. Fluctuations in late 2012 and early 2013 for the Twitter language time series are indicative of inconsistent classifications. (**B**) Rate of usage by language using FastText-LID maintains consistent behavior throughout 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

the fluctuations in the time series during the 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. 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. 7). 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 [127–130]. 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. 8–9. We point the reader's attention to Appendix 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). We have over 118 billion messages between 2009-01-01 and 2019-12-31 spanning 173 languages. English is the 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. 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 [131–134], K-pop, and political events [76]. "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.

3.2 Quantifying Twitter's social contagion: 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 [135]. 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) [136] a message, distinct from a message reply [137].

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

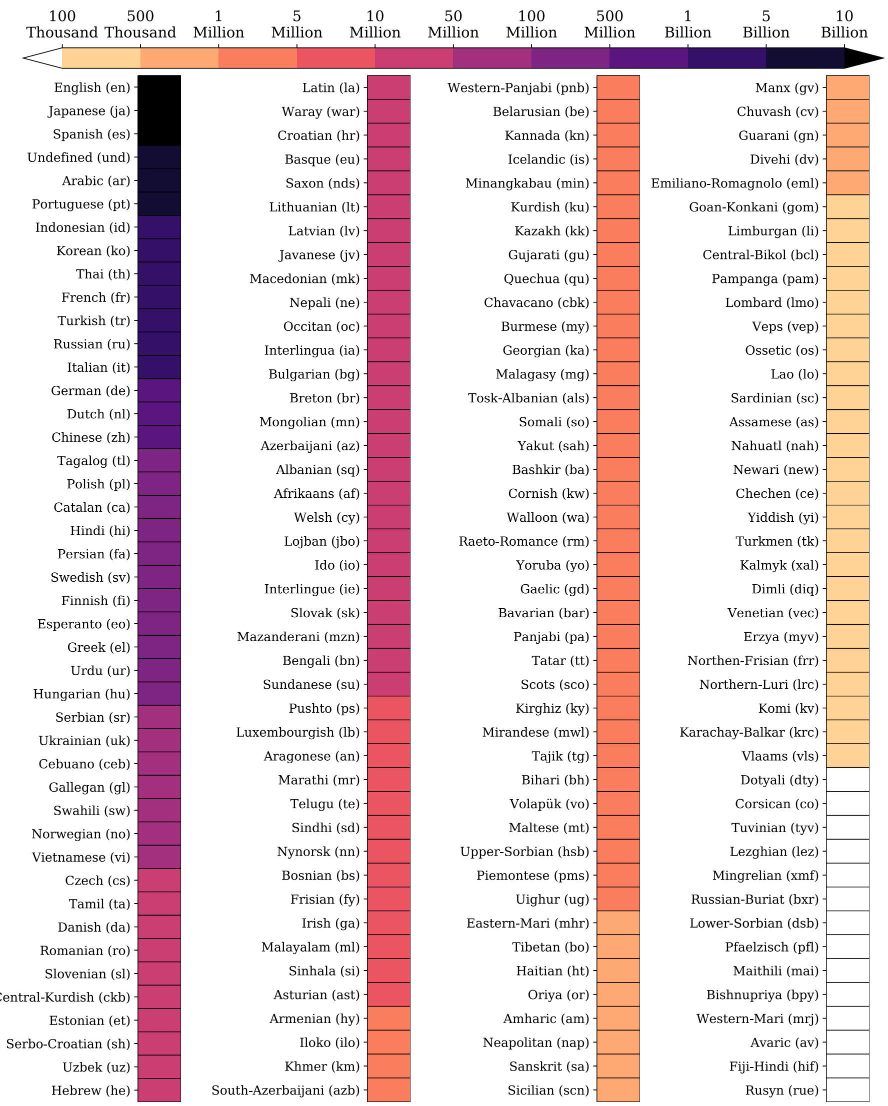

Figure 2 Overall dataset statistics. Number of messages captured in our dataset as classified by the FastText-LID algorithm between 2009-01-01 and 2019-12-31, which sums up to approximately 118 billion messages throughout that period (languages are sorted by popularity). This collection represents roughly 10% of all messages ever posted

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:

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

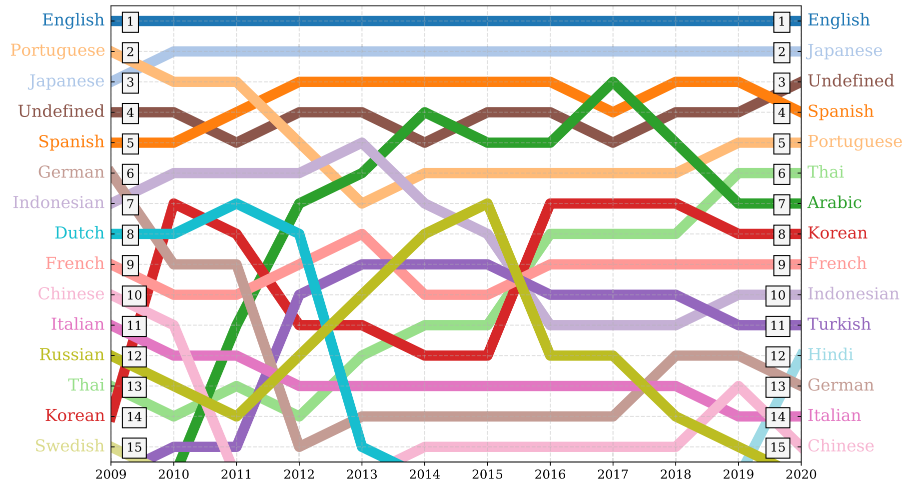

Figure 3 Annual average rank of the most used languages on Twitter between 2009 and 2019. English and Japanese show the most consistent rank time series. Spanish, and Portuguese are also relatively stable over time. Undefined—which covers a wide variety of content such as emojis, links, pictures, and 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

3.3 Measuring social and linguistic 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:

$$R_{\ell,t} = f^{(\mathrm{RT})}_{\ell,t} / f^{(\mathrm{OT})}_{\ell,t}.$$

In 2018, the contagion ratio exceeded 1, indicating a higher number of retweeted messages than organic messages (Fig. 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 [138], along with several new policies to counter synthetic and manipulated media [139–141]. 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

[6] https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html

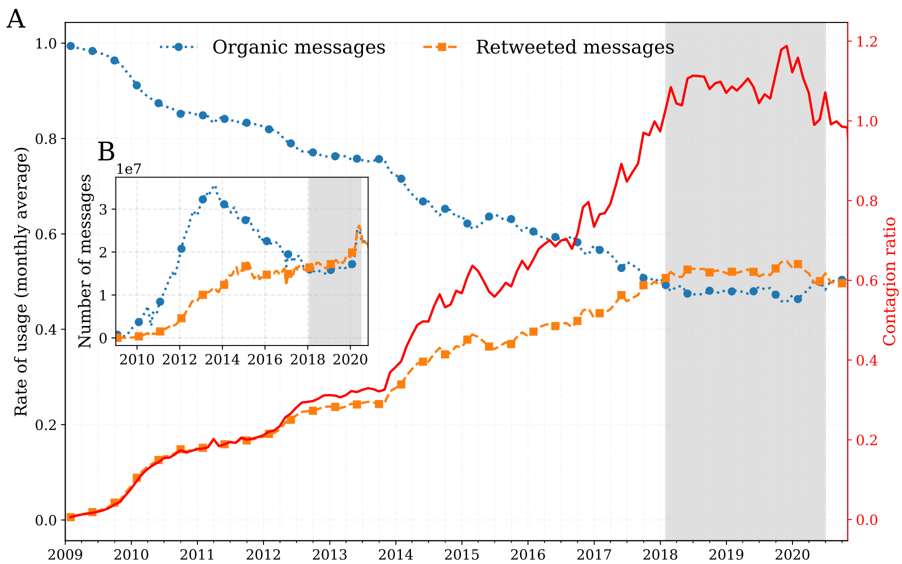

Figure 4 Timeseries for organic messages, retweeted messages, and average contagion ratio for all languages. (**A**) Monthly average rate of usage of organic messages ($p^{(\mathrm{OT})}_{t,\ell}$, blue), and retweeted messages ($p^{(\mathrm{RT})}_{t,\ell}$, orange). The solid red line highlights the 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

networks.[7] We investigate the robustness of our findings regarding contagion ratios in light of some of these changes in Appendix B and Appendix 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 $p_{t,\ell}$ for the top 30 ranked languages of 2019 in Fig. 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 $p^{(\mathrm{RT})}_{t,\ell}$ than organic messages $p^{(\mathrm{OT})}_{t,\ell}$. Thai appears to be the first language to have made this transition in late 2013. In Fig. 6, we show a heatmap of the average contagion ratio $R_{\ell,t}$ 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 and Table 3, respectively.

[7] https://help.twitter.com/en/using-twitter/twitter-conversations

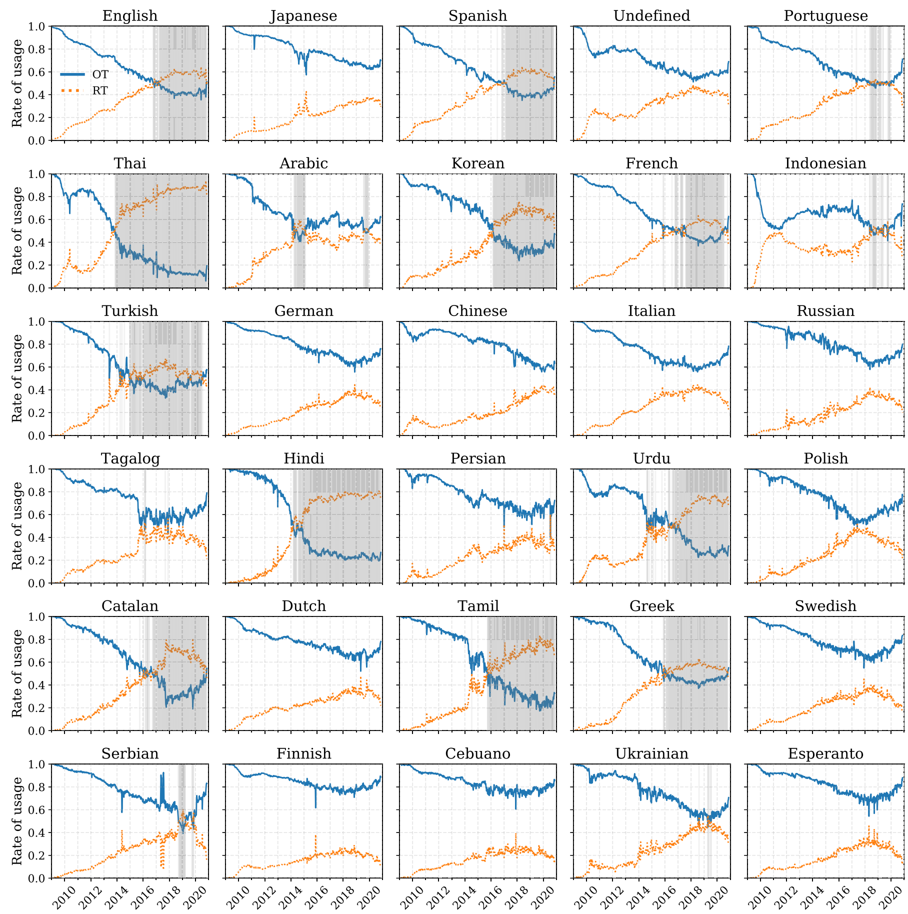

Figure 5 Weekly rate of usage of the top 30 languages (sorted by popularity). For each language, we show a weekly average rate of usage for organic messages ($p_{t,\ell}^{(\mathrm{OT})}$, blue) compared to retweeted messages ($p_{t,\ell}^{(\mathrm{RT})}$, orange). The areas highlighted in light shades 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

4 Concluding remarks

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 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 paper, 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

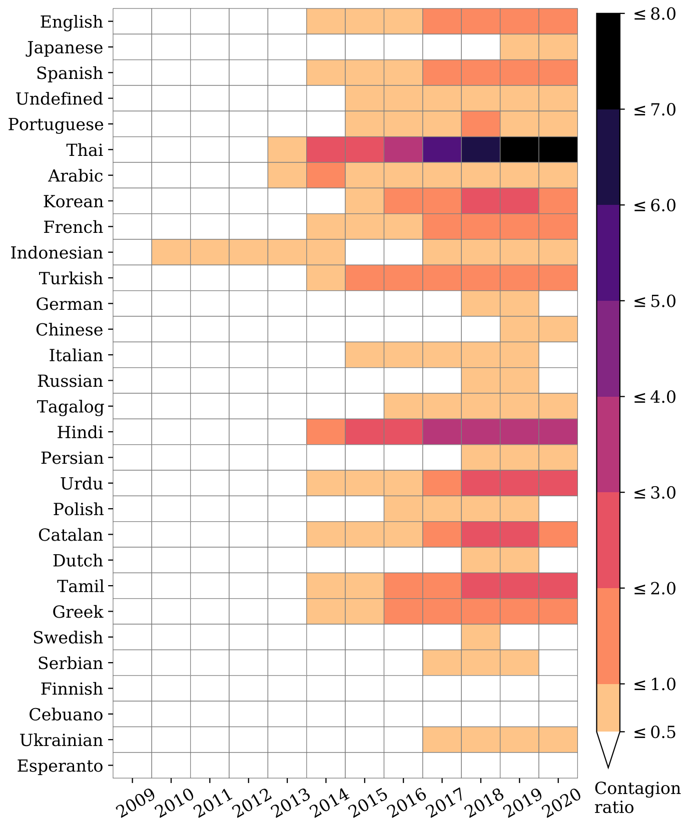

Figure 6 Timelapse of contagion ratios. The average ratio is plotted against year for the top 30 ranked languages of 2019. Colored cells indicate a ratio higher than 0.5 whereas ratios below 0.5 are colored in white. Table 2 shows the top 10 languages with the highest average contagion ratio per year, while Table 3 shows the bottom 10 languages with the lowest average contagion ratio per year

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 [60]. 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 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 [142]. Future research could focus on implementing accurate geotagging algorithms that assign tweets a probabilistic geographical location based on their text and user metadata, while fully respecting privacy through judicious use of masking algorithms.

Appendix 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 1. To evaluate the agreement between the two classifiers, we computed annual confusion matrices starting from 2013 to the end of 2019. In Fig. 7, we show confusion matrices for the 15 most dominate languages on Twitter for all tweets authored in 2013 (Fig. 7A) and 2019 (Fig. 7B).

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. 8, 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:

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

where $\mathcal{C}_\ell^{\mathrm{F}}$ is the number of messages captured by FastText-LID for language ℓ, and $\mathcal{C}_\ell^{\mathrm{T}}$ 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. 8A and Fig. 8B, 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. 8C, we display the joint distribution of all languages captured by

Table 1 Language codes for both FastText-LID and Twitter-LID tools

Language	Fast-Text	Twit-ter	Language	Fast-Text	Twit-ter	Language	Fast-Text	Twit-ter
Afrikaans	af	–	Haitian	ht	ht	Pfaelzisch	pfl	–
Albanian	sq	–	Hebrew	he	he	Piemontese	pms	–
Amharic	am	am	Hindi	hi	hi	Polish	pl	pl
Arabic	ar	ar	Hungarian	hu	hu	Portuguese	pt	pt
Aragonese	an	–	Icelandic	is	is	Pushto	ps	ps
Armenian	hy	hy	Ido	io	–	Quechua	qu	–
Assamese	as	–	Iloko	ilo	–	Raeto-Romance	rm	–
Asturian	ast	–	Indonesian	id	in	Romanian	ro	ro
Avaric	av	–	Inuktitut	–	iu	Russian-Buriat	bxr	–
Azerbaijani	az	–	Interlingua	ia	–	Russian	ru	ru
Bashkir	ba	–	Interlingue	ie	–	Rusyn	rue	–
Basque	eu	eu	Irish	ga	–	Sanskrit	sa	–
Bavarian	bar	–	Italian	it	it	Sardinian	sc	–
Belarusian	be	–	Japanese	ja	ja	Saxon	nds	–
Bengali	bn	bn	Javanese	jv	–	Scots	sco	–
Bihari	bh	–	Kalmyk	xal	–	Serbian	sr	sr
Bishnupriya	bpy	–	Kannada	kn	kn	Serbo-Croatian	sh	–
Bosnian	bs	bs	Karachay-Balkar	krc	–	Sicilian	scn	–
Breton	br	–	Kazakh	kk	–	Sindhi	sd	sd
Bulgarian	bg	bg	Khmer	km	km	Sinhala	si	si
Burmese	my	my	Kirghiz	ky	–	Slovak	sk	–
Catalan	ca	ca	Komi	kv	–	Slovenian	sl	sl
Cebuano	ceb	–	Korean	ko	ko	Somali	so	–
Cherokee	–	chr	Kurdish	ku	–	Shona	–	sn
Central-Bikol	bcl	–	Lao	lo	lo	South-Azerbaijani	azb	–
Central-Kurdish	ckb	ckb	Latin	la	–	Spanish	es	es
Chavacano	cbk	–	Latvian	lv	lv	Sundanese	su	–
Chechen	ce	–	Lezghian	lez	–	Swahili	sw	–
Chinese-Simplified	–	zh-cn	Limburgan	li	–	Swedish	sv	sv
Chinese-Traditional	–	zh-tw	Lithuanian	lt	lt	Tagalog	tl	tl
Chinese	zh	zh	Lojban	jbo	–	Tajik	tg	–
Chuvash	cv	–	Lombard	lmo	–	Tamil	ta	ta
Cornish	kw	–	Lower-Sorbian	dsb	–	Tatar	tt	–
Corsican	co	–	Luxembourgish	lb	–	Telugu	te	te
Croatian	hr	–	Macedonian	mk	–	Thai	th	th
Czech	cs	cs	Maithili	mai	–	Tibetan	bo	bo
Danish	da	da	Malagasy	mg	–	Tosk-Albanian	als	–
Dimli	diq	–	Malayalam	ml	ml	Turkish	tr	tr
Divehi	dv	dv	Malay	ms	msa	Turkmen	tk	–
Dotyali	dty	–	Maltese	mt	–	Tuvinian	tyv	–
Dutch	nl	nl	Manx	gv	–	Uighur	ug	ug
Eastern-Mari	mhr	–	Marathi	mr	mr	Ukrainian	uk	uk
Egyptian-Arabic	arz	–	Mazanderani	mzn	–	Upper-Sorbian	hsb	–
Emiliano-Romagnolo	eml	–	Minangkabau	min	–	Urdu	ur	ur
English	en	en	Mingrelian	xmf	–	Uzbek	uz	–
Erzya	myv	–	Mirandese	mwl	–	Venetian	vec	–
Esperanto	eo	–	Mongolian	mn	–	Veps	vep	–
Estonian	et	et	Nahuatl	nah	–	Vietnamese	vi	vi
Fiji-Hindi	hif	–	Neapolitan	nap	–	Vlaams	vls	–
Filipino	–	fil	Nepali	ne	ne	Volapük	vo	–
Finnish	fi	fi	Newari	new	–	Walloon	wa	–
French	fr	fr	Northen-Frisian	frr	–	Waray	war	–
Frisian	fy	–	Northern-Luri	lrc	–	Welsh	cy	cy
Gaelic	gd	–	Norwegian	no	no	Western-Mari	mrj	–
Gallegan	gl	–	Nynorsk	nn	–	Western-Panjabi	pnb	–
Georgian	ka	ka	Occitan	oc	–	Wu-Chinese	wuu	–
German	de	de	Oriya	or	or	Yakut	sah	–
Goan-Konkani	gom	–	Ossetic	os	–	Yiddish	yi	–
Greek	el	el	Pampanga	pam	–	Yoruba	yo	–
Guarani	gn	–	Panjabi	pa	pa	Yue-Chinese	yue	–
Gujarati	gu	gu	Persian	fa	fa	Undefined	und	und

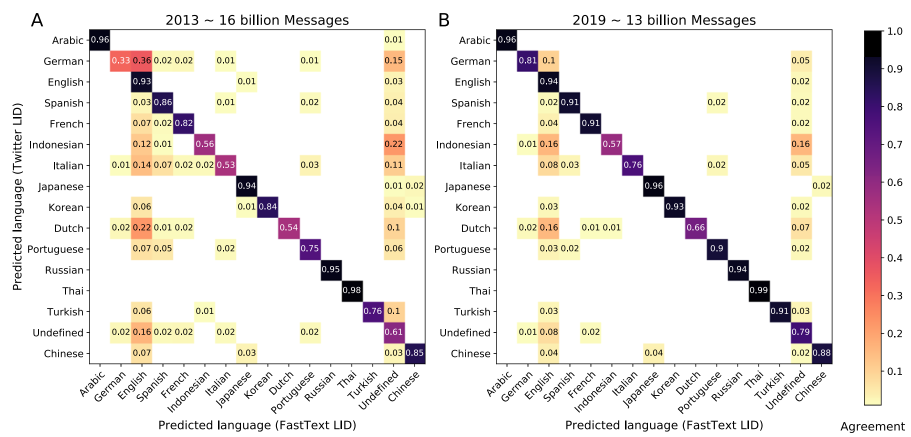

Figure 7 Language identification confusion matrices. We show a subset of the full confusion matrix for top-15 languages on Twitter. (**A**) Confusion matrix for tweets authored in 2013. The matrix indicates substantial disagreement between the two classifiers during 2013, the first year of Twitter's 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 0.01 are colored in white to indicate very minor disagreement between the two classifiers

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. 9.

Appendix 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 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 and Table 3, respectively. We then compare the new set of annual contagion ratios with the original ones discussed in Sect. 3.3. 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 4, 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

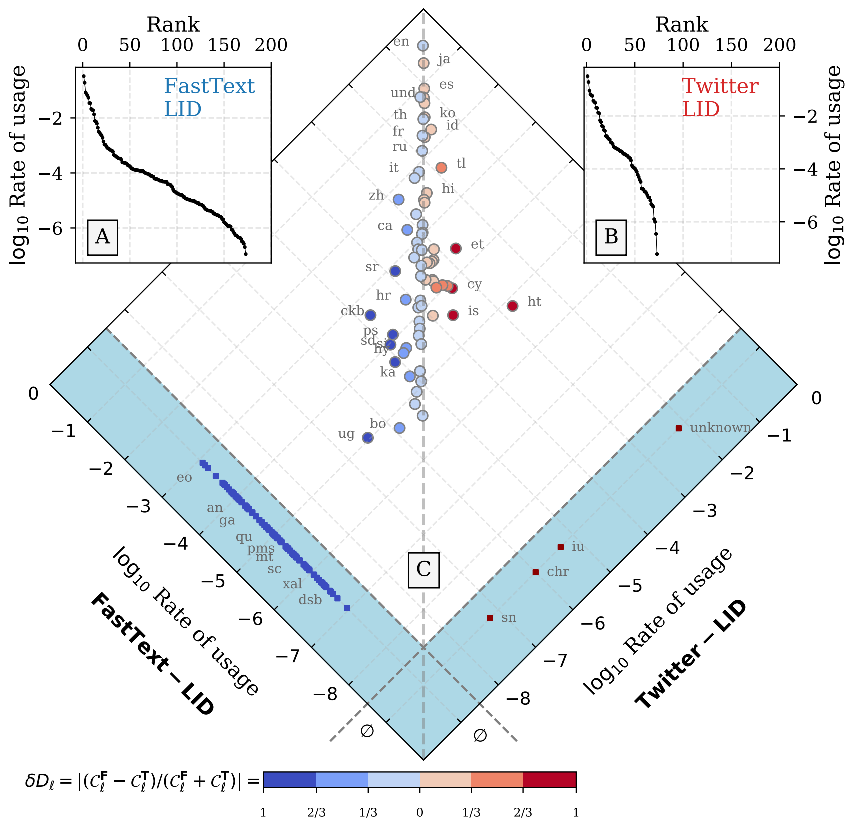

Figure 8 Language Zipf distributions. (**A**) Zipf distribution [143] of all languages captured by FastText-LID model. (**B**) Zipf distribution for languages captured by Twitter-LID algorithm(s). The vertical axis in both panels reports rate of usage of all messages $p_{t,\ell}$ between 2014 and 2019, 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 $\mathcal{C}_\ell^{\mathrm{F}}$ is the number of messages captured by FastText-LID for language ℓ, and $\mathcal{C}_\ell^{\mathrm{T}}$ 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 1, and an online appendix of all languages is also available here: http://compstorylab.org/storywrangler/papers/tlid/files/fasttext_twitter_timeseries.html

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 5.

In Fig. 10, 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. By contrast, we note high disagreement on Catalan messages. The two classifiers start off with unusual disagreement in 2014 ($\delta = 0.52$). 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

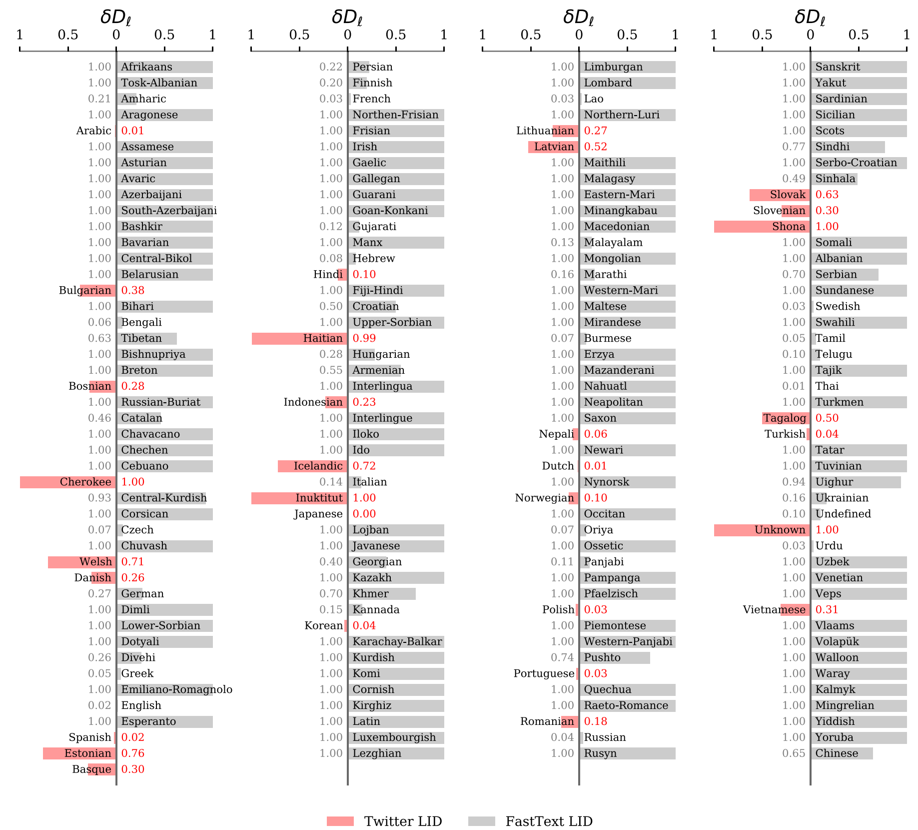

Figure 9 Language identification divergence. A normalized ratio difference δD_ℓ (i.e., divergence) closer to zero implies strong agreement, whereby both classifiers captured approximately the same number of messages over the last decade. Grey bars indicate higher rate of messages captured by FastText-LID, whereas red bars highlight higher rate of messages captured by Twitter-LID

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

Language	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

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.

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

Table 3 Bottom 10 languages with the lowest annual average contagion ratio (sorted by 2019)

Language	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 4 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

Table 5 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

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.

Appendix 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 [111]. 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

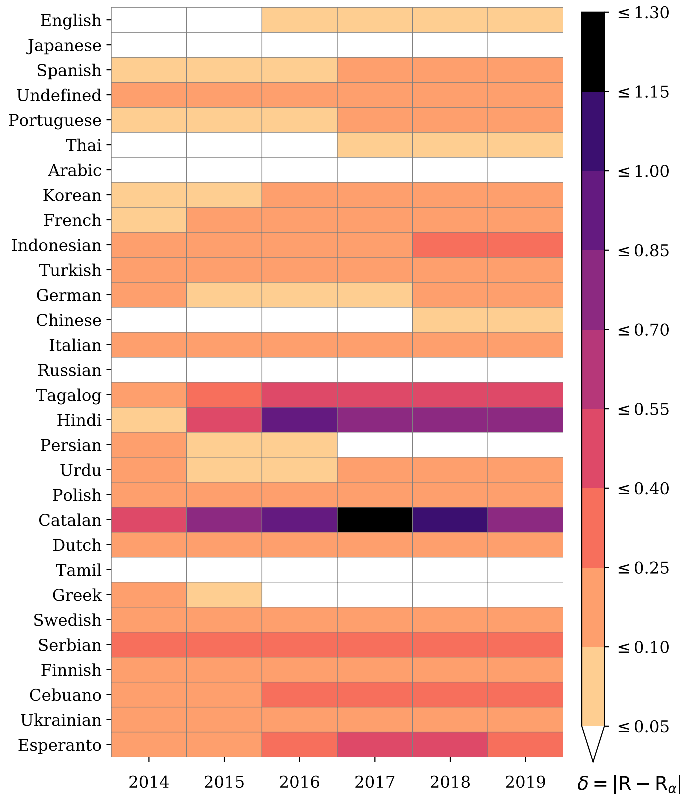

Figure 10 Margin of error for contagion ratios. We compute the annual average of contagion ratios R for all messages in the top 30 ranked languages as classified by FastText-LID and described in Sect. 3.3. Similarly, we compute the annual average of contagion ratios R_α for the subset of messages that 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 0.05, whereas colored cells highlight values that are equal to, or above 0.05. We show the top 10 languages with the highest average values of δ's per year in Table 4. We also show the bottom 10 languages with the lowest average values of δ's per year in Table 5

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. 11, 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 built-in Python function to get the exact number of characters in a given message.[9] As anticipated, our results indicate a higher proportion of short messages

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

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

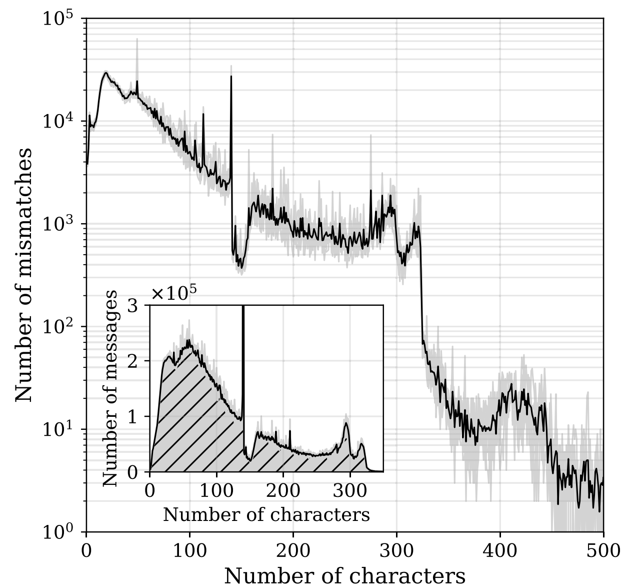

Figure 11 Language identification uncertainty as a function of tweet-length for top 10 most used languages on Twitter. We display the number of messages that were classified differently by Twitter-LID model and FastText-LID for the top-10 prominent languages as a function of the number of characters 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

Table 6 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	0.0853
Japanese	6.8×10^6	0.0268
Spanish	2.3×10^6	0.0558
Thai	2.2×10^6	0.0161
Portuguese	2.1×10^6	0.0565
Korean	1.7×10^6	0.0085
Arabic	1.5×10^6	0.0080
Indonesian	8.1×10^5	0.1203
French	7.9×10^5	0.1305
Turkish	5.6×10^5	0.0325

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 6 (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 0.25 are labeled as Undefined (und). On the other hand, messages with scores greater or equal to 0.25 but less than 0.5 are flagged as predictions with low confidence (low). Predictions that have scores in the range $[0.5, 0.75)$

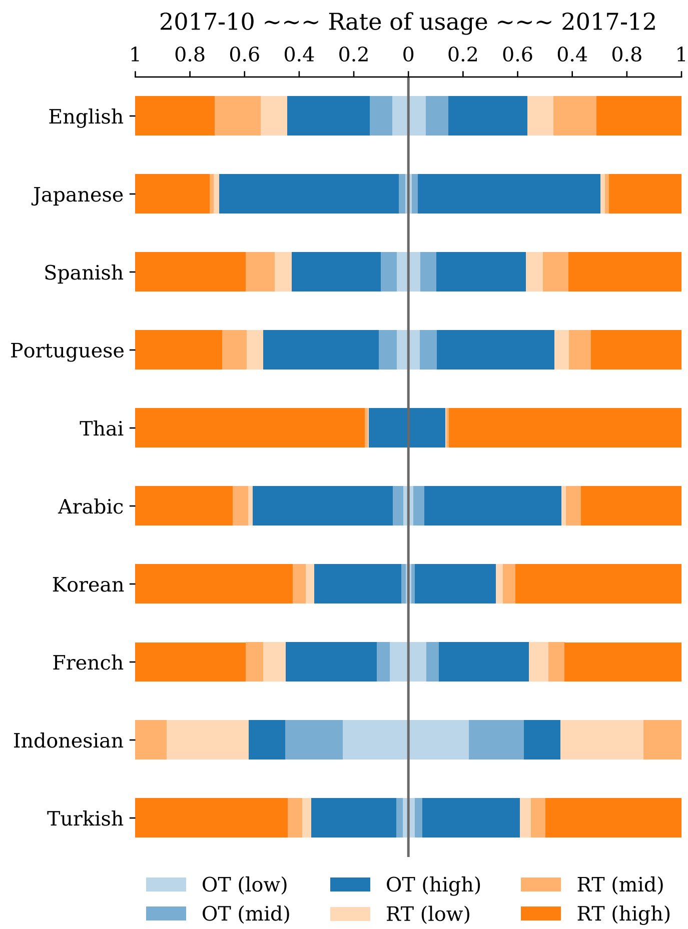

Figure 12 Confidence scores of the FastText-LID neural network predictions for the month before and after the shift to 280 characters. We categorize messages into four classes based on the confidence scores we get from FastText-LID's neural network. Predictions with confidence scores below 0.25 are labeled as Undefined (und). Messages with scores greater or equal to 0.25 but less than 0.5 are flagged as predictions with low confidence (low). Predictions that have scores in the range $[0.5, 0.75)$ 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

are considered moderate (mid), and messages with higher scores are labeled as predictions with high confidence (high).

In Fig. 12, 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. 13A, 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. Figure 13B and Fig. 13C show the growth of long messages on the platform, while the rate of usage for the most used languages re-

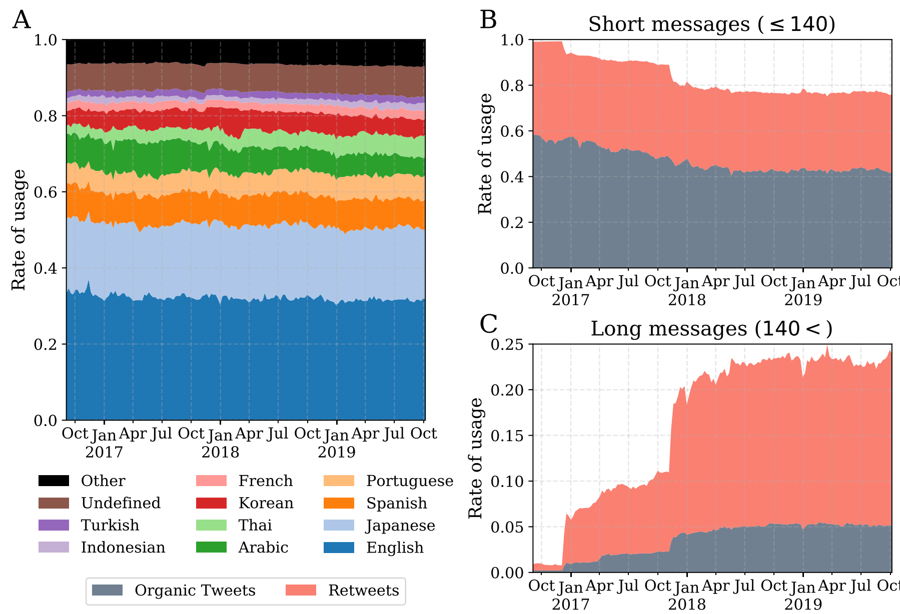

Figure 13 Weekly rate of usage for short and long messages. (**A**) Rate of usage for the top-10 used languages averaged at the week scale for the past three years. The introduction of long messages (i.e., above 140 but below 280 characters) does not change the overall language usage on the platform. (**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

mains consistent. In Fig. 13C, 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.

Acknowledgements
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.

Funding
The authors acknowledge financial support from NSF Big Data Grant #1447634 and MassMutual Life Insurance and Google.

Abbreviations
NLP, Natural Language Processing; LID, Language Identification and Detection; FastText-LID, The language detection tool based on FastText; CLD, Google's Compact Language Detector; GLM, General Linear Model; LKJ, Lewandowski–Kurowicka–Joe distribution; QT, Quote Tweets; a comment on a retweeted message; RT, Retweeted Tweets; repeated content (retweets) and the non-organic content found in Quote Tweets; OT, Organic Tweets; all publicly available new content including tweets, replies, and comments; AT, All Tweets; all messages on the platform.

Availability of data and materials
The datasets analysed during the current study and our source code are available in the https://gitlab.com/compstorylab/storywrangler repository.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
TA, CMD, and PSD conceived the idea; TA, DRD, JRM, and MVA collected and analyzed data; TA, DRD, CMD, and PSD wrote the manuscript; TA, DRD, JRM, MVA, JLA, CMD, and PSD edited the manuscript. All authors read and approved the final manuscript.

Author details
[1]Vermont Complex Systems Center, University of Vermont, Burlington, VT 05405, USA. [2]Computational Story Lab, University of Vermont, Burlington, VT 05405, USA. [3]Department of Computer Science, University of Vermont, Burlington, VT 05405, USA. [4]Charles River Analytics, Cambridge, MA 02138, USA. [5]Department of Mathematics & Statistics, University of Vermont, Burlington, VT 05405, USA.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 10 April 2020 Accepted: 17 March 2021 Published online: 31 March 2021

References

1. Suh B, Hong L, Pirolli P, Chi EH (2010) Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In: 2010 IEEE second international conference on social computing. IEEE, pp 177–184
2. Boyd D, Golder S, Lotan G (2010) Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: 2010 43rd Hawaii international conference on system sciences. IEEE, pp 1–10. https://doi.org/10.1109/HICSS.2010.412
3. Nagarajan M, Purohit H, Sheth A (2010) A qualitative examination of topical tweet and retweet practices. In: Proceedings of the international AAAI conference on web and social media, vol 4
4. Hodas NO, Lerman K (2012) 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. IEEE, pp 249–257
5. Harrigan N, Achananuparp P, Lim E-P (2012) Influentials, novelty, and social contagion: the viral power of average friends, close communities, and old news. Soc Netw 34(4):470–480
6. Hodas NO, Lerman K (2014) The simple rules of social contagion. Sci Rep 4:4343
7. Goffman W, Newill VA (1964) Generalization of epidemic theory: an application to the transmission of ideas. Nature 204:225–228
8. Daley DJ, Kendall DG (1965) Stochastic rumours. J Inst Math Appl 1:42–55
9. Schelling TC (1971) Dynamic models of segregation. J Math Sociol 1:143–186
10. Granovetter M (1978) Threshold models of collective behavior. Am J Sociol 83(6):1420–1443
11. Dodds PS, Watts DJ (2004) Universal behavior in a generalized model of contagion. Phys Rev Lett 92:218701
12. Dodds PS, Watts DJ (2005) A generalized model of social and biological contagion. J Theor Biol 232:587–604. https://doi.org/10.1016/j.jtbi.2004.09.006
13. Centola D, Macy MW (2007) Complex contagions and the weakness of long ties. Am J Sociol 113:702–734
14. Ugander J, Backstrom L, Marlow C, Kleinberg J (2012) Structural diversity in social contagion. Proc Natl Acad Sci 109(16):5962–5966
15. Cozzo E, Banos RA, Meloni S, Moreno Y (2013) Contact-based social contagion in multiplex networks. Phys Rev E 88(5):050801
16. Bessi A, Coletto M, Davidescu GA, Scala A, Caldarelli G, Quattrociocchi W (2015) Science vs conspiracy: collective narratives in the age of misinformation. PLoS ONE 10(2):0118093
17. Bass F (1969) A new product growth model for consumer durables. Manag Sci 15:215–227
18. Van den Bulte C, Joshi YV (2007) New product diffusion with influentials and imitators. Mark Sci 26(3):400–421
19. Trusov M, Bucklin RE, Pauwels K (2009) Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. J Mark 73(5):90–102
20. Iyengar R, Van den Bulte C, Valente TW (2011) Opinion leadership and social contagion in new product diffusion. Mark Sci 30(2):195–212
21. Kelly M, O Grada C (2000) Market contagion: evidence from the panics of 1854 and 1857. Am Econ Rev 90(5):1110–1124
22. Cipriani M, Guarino A (2008) Herd behavior and contagion in financial markets. B E J Theor Econ 8(1):1–56
23. Hirshleifer D, Teoh SH (2009) Thought and behavior contagion in capital markets. In: Hens T, Schenk-Hoppé KR (eds) Handbook of financial markets: dynamics and evolution. Handbooks in finance. North-Holland, San Diego, pp 1–56. http://www.sciencedirect.com/science/article/pii/B9780123742582500051
24. Fenzl T, Pelzmann L (2012) Psychological and social forces behind aggregate financial market behavior. J Behav Finance 13(1):56–65
25. Hamilton JD, Hamilton LC (1981) Models of social contagion. J Math Sociol 8(1):133–160
26. Bovasso G (1996) A network analysis of social contagion processes in an organizational intervention. Hum Relat 49(11):1419–1435
27. Fagan J, Wilkinson DL, Davies G (2007) In: Flannery DJ, Vazsonyi AT, Waldman IDE (eds) Social contagion of violence. Cambridge handbooks in psychology. Cambridge University Press, Cambridge, pp 688–724. https://doi.org/10.1017/CBO9780511816840.037

28. Christakis NA, Fowler JH (2013) Social contagion theory: examining dynamic social networks and human behavior. Stat Med 32(4):556–577
29. Papachristos AV, Wildeman C, Roberto E (2015) Tragic, but not random: the social contagion of nonfatal gunshot injuries. Soc Sci Med 125:139–150
30. Pollack CE, Soulos PR, Herrin J, Xu X, Christakis NA, Forman HP, Yu JB, Killelea BK, Wang S-Y, Gross CP (2017) The impact of social contagion on physician adoption of advanced imaging tests in breast cancer. J Natl Cancer Inst 109(8):330
31. Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489(7415):295–298
32. Kramer AD, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci 111(24):8788–8790
33. Ellison NB, Vitak J, Gray R, Lampe C (2014) Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. J Comput-Mediat Commun 19(4):855–870
34. Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2016) The rise of social bots. Commun ACM 59(7):96–104
35. Lerman K, Ghosh R (2010) 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
36. Borge-Holthoefer J, Moreno Y (2012) Absence of influential spreaders in rumor dynamics. Phys Rev E 85(2):026116
37. Kwon S, Cha M, Jung K, Chen W, Wang Y (2013) Prominent features of rumor propagation in online social media. In: 2013 IEEE 13th international conference on data mining. IEEE, pp 1103–1108
38. Ozturk P, Li H, Sakamoto Y (2015) Combating rumor spread on social media: the effectiveness of refutation and warning. In: 2015 48th Hawaii international conference on system sciences. IEEE, pp 2406–2414
39. Kaligotla C, Yücesan E, Chick SE (2015) An agent based model of spread of competing rumors through online interactions on social media. In: 2015 winter simulation conference (WSC). IEEE, pp 3985–3996
40. Zubiaga A, Liakata M, Procter R, Wong Sak Hoi G, Tolmie P (2016) Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE 11(3):0150989
41. Del Vicario M, Bessi A, Zollo F, Petroni F, Scala A, Caldarelli G, Stanley HE, Quattrociocchi W (2016) The spreading of misinformation online. Proc Natl Acad Sci 113(3):554–559
42. Spohr D (2017) Fake news and ideological polarization: filter bubbles and selective exposure on social media. Bus Inf Rev 34(3):150–160
43. Shao C, Ciampaglia GL, Varol O, Yang K-C, Flammini A, Menczer F (2018) The spread of low-credibility content by social bots. Nat Commun 9(1):1–9
44. Törnberg P (2018) Echo chambers and viral misinformation: modeling fake news as complex contagion. PLoS ONE 13(9):0203958
45. Zaman TR, Herbrich R, Van Gael J, Stern D (2010) Predicting information spreading in Twitter. In: Workshop on computational social science and the wisdom of crowds. NIPS, vol 104. Citeseer, pp 17599–17601
46. Romero DM, Meeder B, Kleinberg J (2011) 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, pp 695–704
47. Weng L, Flammini A, Vespignani A, Menczer F (2012) Competition among memes in a world with limited attention. Nat Sci Rep 2:335
48. Colleoni E, Rozza A, Arvidsson A (2014) Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. J Commun 64(2):317–332
49. Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R (2015) Tweeting from left to right: is online political communication more than an echo chamber? Psychol Sci 26(10):1531–1542. https://doi.org/10.1177/0956797615594620. PMID: 26297377
50. Barberá P (2015) Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Polit Anal 23(1):76–91. https://doi.org/10.1093/pan/mpu011
51. Stieglitz S, Dang-Xuan L (2012) 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. IEEE, pp 3500–3509
52. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, pp 591–600
53. Rao HR, Vemprala N, Akello P, Valecha R (2020) Retweets of officials' alarming vs reassuring messages during the COVID-19 pandemic: implications for crisis management. Int J Inf Manag 55:102187
54. Mønsted B, Sapieżyński P, Ferrara E, Lehmann S (2017) Evidence of complex contagion of information in social media: an experiment using Twitter bots. PLoS ONE 12(9):0184148
55. Cha M, Haddadi H, Benevenuto F, Gummadi K (2010) Measuring user influence in Twitter: the million follower fallacy. In: Proceedings of the international AAAI conference on web and social media, vol 4
56. Fitch WT (2017) Empirical approaches to the study of language evolution. Psychon Bull Rev 24(1):3–33
57. Bolhuis JJ, Okanoya K, Scharff C (2010) Twitter evolution: converging mechanisms in birdsong and human speech. Nat Rev Neurosci 11(11):747–759
58. Kim S, Weber I, Wei L, Oh A (2014) Sociolinguistic analysis of Twitter in multilingual societies. In: Proceedings of the 25th ACM conference on hypertext and social media, pp 243–248
59. Fábrega J, Paredes P (2013) Social contagion and cascade behaviors on Twitter. Information 4(2):171–181
60. Joulin A, Grave E, Bojanowski P, Mikolov T (2017) Bag of tricks for efficient text classification. In: Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers. Association for Computational Linguistics, Valencia, pp 427–431. https://www.aclweb.org/anthology/E17-2068
61. Twitter (2019) Developer application program interface (API). https://developer.twitter.com/en/docs/ads/campaign-management/api-reference
62. Hong L, Convertino G, Chi E (2011) Language matters in Twitter: a large scale study. In: Proceedings of the international AAAI conference on web and social media, vol 5
63. Zubiaga A, Spina D, Martínez R, Fresno V (2015) Real-time classification of Twitter trends. J Assoc Inf Sci Technol 66(3):462–473

64. Dewhurst DR, Alshaabi T, Kiley D, Arnold MV, Minot JR, Danforth CM, Dodds PS (2020) The shocklet transform: a decomposition method for the identification of local, mechanism-driven dynamics in sociotechnical time series. EPJ Data Sci 9(1):3
65. Mellon J, Prosser C (2017) Twitter and Facebook are not representative of the general population: political attitudes and demographics of British social media users. Res Polit 4(3):2053168017720008
66. Ke Q, Ahn Y-Y, Sugimoto CR (2017) A systematic identification and analysis of scientists on Twitter. PLoS ONE 12(4):1–17. https://doi.org/10.1371/journal.pone.0175368
67. Mitchell A, Hitlin P (2019) Twitter reaction to events often at odds with overall public opinion. Pew Research Center: Internet, Science & Tech
68. Wojcik S, Hughes A (2019) How Twitter users compare to the general public. Pew Research Center: Internet, Science & Tech
69. Palen L, Anderson KM (2016) Crisis informatics—new data for extraordinary times. Science 353(6296):224–225
70. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web. Association for Computing Machinery, New York, pp 851–860. https://doi.org/10.1145/1772690.1772777
71. Lampos V, Cristianini N (2010) Tracking the flu pandemic by monitoring the social web. In: 2010 2nd international workshop on cognitive information processing, pp 411–416. https://doi.org/10.1109/CIP.2010.5604088
72. Culotta A (2010) Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the first workshop on social media analytics. SOMA 10. Assoc. Comput. Mach., New York, pp 115–122. https://doi.org/10.1145/1964858.1964874
73. Pickard G, Pan W, Rahwan I, Cebrian M, Crane R, Madan A, Pentland A (2011) Time-critical social mobilization. Science 334(6055):509–512
74. Gao H, Barbier G, Goolsby R (2011) Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intell Syst 26(3):10–14
75. Steinert-Threlkeld ZC, Mocanu D, Vespignani A, Fowler J (2015) Online social networks and offline protest. EPJ Data Sci 4(1):19
76. Dodds PS, Minot JR, Arnold MV, Alshaabi T, Adams JL, Dewhurst DR, Reagan AJ, Danforth CM (2019) Fame and ultrafame: measuring and comparing daily levels of 'being talked about' for United States' presidents, their rivals, God, countries, and K-pop. http://arxiv.org/abs/1910.00149
77. Ritter A, Clark S, Mausam EO (2011) Named entity recognition in tweets: an experimental study. In: Proceedings of the 2011 conference on empirical methods in natural language processing. Association for Computational Linguistics, Edinburgh, pp 1524–1534. https://www.aclweb.org/anthology/D11-1141
78. Ritter A, Mausam EO, Clark S (2012) Open domain event extraction from Twitter. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. KDD'12. Assoc. Comput. Mach., New York, pp 1104–1112. https://doi.org/10.1145/2339530.2339704
79. Hirschberg J, Manning CD (2015) Advances in natural language processing. Science 349(6245):261–266
80. Lui M, Baldwin T (2012) langid.py: an off-the-shelf language identification tool. In: Proceedings of the ACL 2012 system demonstrations. Association for Computational Linguistics, pp 25–30
81. Bergsma S, McNamee P, Bagdouri M, Fink C, Wilson T (2012) Language identification for creating language-specific Twitter collections. In: Proceedings of the second workshop on language in social media. LSM'12. Association for Computational Linguistics, pp 65–74
82. Lui M, Baldwin T (2014) Accurate language identification of Twitter messages. In: Proceedings of the 5th workshop on language analysis for social media (LASM). Association for Computational Linguistics, Gothenburg, pp 17–25. https://doi.org/10.3115/v1/W14-1303. https://www.aclweb.org/anthology/W14-1303
83. Williams J, Dagli C (2017) 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). Association for Computational Linguistics, Valencia, pp 73–83. https://doi.org/10.18653/v1/W17-1209. https://www.aclweb.org/anthology/W17-1209
84. Dodds PS, Harris KD, Kloumann IM, Bliss CA, Danforth CM (2011) Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE 6(12):e26752. https://doi.org/10.1371/journal.pone.0026752
85. Chu Z, Gianvecchio S, Wang H, Jajodia S (2012) Detecting automation of Twitter accounts: are you a human, bot, or cyborg? IEEE Trans Dependable Secure Comput 9(6):811–824
86. Kharde V, Sonawane S (2016) Sentiment analysis of Twitter data: a survey of techniques. Int J Comput Appl 139(11):5–15. https://doi.org/10.5120/ijca2016908625
87. Kryvasheyeu Y, Chen H, Obradovich N, Moro E, Van Hentenryck P, Fowler J, Cebrian M (2016) Rapid assessment of disaster damage using social media activity. Sci Adv 2(3):1500779
88. Kursuncu U, Gaur M, Lokala U, Thirunarayan K, Sheth A, Arpinar IB (2019) In: Agarwal N, Dokoohaki N, Tokdemir S (eds) Predictive analysis on Twitter: techniques and applications. Springer, Cham, pp 67–104. https://doi.org/10.1007/978-3-319-94105-9_4
89. Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, pp 1532–1543. https://doi.org/10.3115/v1/D14-1162. https://www.aclweb.org/anthology/D14-1162
90. Devlin J, Chang M-W, Lee K, Toutanova K (2019) 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). Association for Computational Linguistics, Minneapolis, pp 4171–4186. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
91. Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A (2018) Advances in pre-training distributed word representations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki

92. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) 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), Miyazaki. https://www.aclweb.org/anthology/L18-1550
93. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318
94. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015. Conference track proceedings
95. Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics, Lisbon, pp 1412–1421. https://doi.org/10.18653/v1/D15-1166
96. McNamee P (2005) Language identification: a solved problem suitable for undergraduate instruction. J Comput Sci Coll 20(3):94–101
97. Hughes B, Baldwin T, Bird S, Nicholson J, MacKinlay A (2006) Reconsidering language identification for written language resources. In: Proceedings of the fifth international conference on language resources and evaluation (LREC'06). European Language Resources Association (ELRA), Genoa. http://www.lrec-conf.org/proceedings/lrec2006/pdf/459_pdf.pdf
98. Grothe L, De Luca EW, Nürnberger A (2008) A comparative study on language identification methods. In: Proceedings of the sixth international conference on language resources and evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech
99. Lui M, Baldwin T (2011) Cross-domain feature selection for language identification. In: Proceedings of 5th international joint conference on natural language processing. Asian Federation of Natural Language Processing, Chiang Mai, pp 553–561. https://www.aclweb.org/anthology/I11-1062
100. Lui M, Lau JH, Baldwin T (2014) Automatic detection and language identification of multilingual documents. Trans Assoc Comput Linguist 2:27–40. https://doi.org/10.1162/tacl_a_00163
101. Michel J-B, Shen YK, Aiden AP, Veres A, Gray MK, Pickett JP, Hoiberg D, Clancy D, Norvig P, Orwant J et al (2011) Quantitative analysis of culture using millions of digitized books. Science 331(6014):176–182
102. Roomann-Kurrik A (2013) Introducing new metadata for tweets. Twitter
103. Tromp E, Pechenizkiy M (2011) Graph-based N-gram language identification on short texts. In: Proceedings of Benelearn 2011, pp 27–34
104. Elfardy H, Diab M (2012) Token level identification of linguistic code switching. In: Proceedings of COLING 2012: posters. The COLING 2012 Organizing Committee, Mumbai, pp 287–296
105. Carter S, Weerkamp W, Tsagkias M (2013) Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. Lang Resour Eval 47(1):195–215. https://doi.org/10.1007/s10579-012-9195-y
106. Steinmetz K (2013) What Twitter says to linguists. Time Inc. http://content.time.com/time/subscriber/article/0,33009,2150609,00.html
107. Goldszmidt M, Najork M, Paparizos S (2013) Boot-strapping language identifiers for short colloquial postings. In: Blockeel H, Kersting K, Nijssen S, Železný F (eds) Machine learning and knowledge discovery in databases. Springer, Berlin, pp 95–111
108. Nguyen D, Trieschnigg D, Cornips L (2015) Audience and the use of minority languages on Twitter. In: Proceedings of the international AAAI conference on web and social media, vol 9
109. Vilares D, Alonso MA, Gómez-Rodríguez C (2015) 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. Association for Computational Linguistics, Lisboa, pp 2–8. https://doi.org/10.18653/v1/W15-2902
110. Rijhwani S, Sequiera R, Choudhury M, Bali K, Maddila C (2017) Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In: Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers), pp 1971–1982. https://doi.org/10.18653/v1/P17-1180
111. Rosen A (2017) Tweeting made easier. https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html
112. Batrinca B, Treleaven PC (2015) Social media analytics: a survey of techniques, tools and platforms. AI & Society 30(1):89–116
113. Giachanou A, Crestani F (2016) Like it or not: a survey of Twitter sentiment analysis methods. ACM Comput Surv 49(2):28. https://doi.org/10.1145/2938640
114. Pla F, Hurtado L-F (2017) Language identification of multilingual posts from Twitter: a case study. Knowl Inf Syst 51(3):965–989
115. Zubiaga A, San Vicente I, Gamallo P, Pichel JR, Alegria I, Aranberri N, Ezeiza A, Fresno V (2016) Tweetlid: a benchmark for tweet language identification. Lang Resour Eval 50(4):729–766
116. Blodgett SL, Wei J, O'Connor B (2017) A dataset and classifier for recognizing social media English. In: Proceedings of the 3rd workshop on noisy user-generated text. Association for Computational Linguistics, Copenhagen, pp 56–61. https://doi.org/10.18653/v1/W17-4408
117. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. In: Bengio Y, LeCun Y (eds) 1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013. Workshop track proceedings. http://arxiv.org/abs/1301.3781
118. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135–146
119. Facebook AI Research (2017) FastText language identification. https://fasttext.cc/docs/en/language-identification.html
120. Conneau A, Schwenk H, Barrault L, Lecun Y (2017) 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. Association for Computational Linguistics, Valencia, pp 1107–1116

121. Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R (eds) Advances in neural information processing systems, vol 28. Curran Associates, Red Hook
122. Bergsma S, Dredze M, Van Durme B, Wilson T, Yarowsky D (2013) 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, pp 1010–1019
123. Twitter (2019) Rules and filtering. https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/premium-operators
124. Phillips A, Davis M (2009) Best current practice (BCP): tags for identifying languages. Technical report, Network Working Group IETF, California, USA
125. Alshaabi T, Adams JL, Arnold MV, Minot JR, Dewhurst DR, Reagan AJ, Danforth CM, Dodds PS (2020) Storywrangler: a massive exploratorium for sociolinguistic, cultural, socioeconomic, and political timelines using Twitter. http://arxiv.org/abs/2003.03667
126. Dodds PS et al (2020) Long-term word frequency dynamics derived from Twitter are corrupted: a bespoke approach to detecting and removing pathologies in ensembles of time series. https://arxiv.org/abs/2008.11305
127. Ringbom H (2006) Cross-linguistic similarity in foreign language learning. Multilingual Matters, Bristol. https://doi.org/10.21832/9781853599361
128. Borer H (1984) Parametric syntax: case studies in semitic and romance languages. de Gruyter, Berlin. https://doi.org/10.1515/9783110808506
129. Samoilenko A, Karimi F, Edler D, Kunegis J, Strohmaier M (2016) Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. EPJ Data Sci 5(1):9
130. Jin H, Toyoda M, Yoshinaga N (2017) Can cross-lingual information cascades be predicted on Twitter? In: International conference on social informatics. Springer, Berlin, pp 457–472
131. Hussain M, Howard P (2012) Democracy's fourth wave? Information technologies and the fuzzy causes of the Arab Spring. SSRN Electron J 57. https://doi.org/10.2139/ssrn.2029711
132. Wolfsfeld G, Segev E, Sheafer T (2013) Social media and the Arab Spring: politics comes first. Int J Press Polit 18(2):115–137
133. Dewey T, Kaden J, Marks M, Matsushima S, Zhu B (2012) The impact of social media on social unrest in the Arab Spring. Int Policy Program 5:8
134. Cottle S (2011) Media and the Arab uprisings of 2011. Journalism 12(5):647–659
135. Stone B (2009) Retweet limited rollout. Twitter
136. Shu C (2015) Twitter officially launches its "retweet with comment" feature. TechCrunch
137. Stone B (2007) Are you Twittering @ me? Twitter. https://blog.twitter.com/official/en_us/a/2007/are-you-twittering-me.html
138. Gadde V, Beykpour K (2020) Additional steps we're taking ahead of the 2020 US election. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html
139. Roth Y, Achuthan A (2020) 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
140. Roth Y, Pickles N (2020) Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html
141. Gadde V, Beykpour K (2020) Expanding our policies to further protect the civic conversation. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html
142. Twitter (2019) Tweet geospatial metadata. https://developer.twitter.com/en/docs/tutorials/tweet-geo-metadata
143. Zipf GK (1949) Human behaviour and the principle of least-effort. Addison-Wesley, Cambridge