The growing echo chamber of social media: Measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020

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

[1] *Vermont Complex Systems Center, The University of Vermont, Burlington, VT 05405.*
[2] *Computational Story Lab, The University of Vermont, Burlington, VT 05405.*
[3] *Department of Computer Science, The University of Vermont, Burlington, VT 05405.*
[4] *Department of Mathematics & Statistics, The University of Vermont, Burlington, VT 05405.*

(Dated: March 10, 2020)

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, and Portuguese being the most dominant. To quantify each language's level of being a Twitter 'echo chamber' 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.

I. INTRODUCTION

Twitter is a well-structured streaming source of sociotechnical data allowing for the study of dynamical linguistics and cultural phenomena [1–3]. Of course, like many other social platforms, Twitter represents only a subsample of the publicly declared views of utterances, and interactions of hundreds of millions of individuals, organizations, and automated accounts (Twitter social bots) around the world [4–7]. Researchers have, nevertheless, shown that Twitter's collective conversation mirrors the dynamics of local and global events [8] including earthquakes [9], flu and influenza [10, 11], crowdsourcing and disaster relief [12, 13], major political affairs [14, 15], and fame dynamics for political figures and celebrities [16]. Moreover, analyses of social media data and digital text corpora over the last decade have advanced Natural Language Processing (NLP) research [17–19], sentiment detection [20, 21], word representation [22–26], text summarization [27–29], and network science [30–34].

Language Identification (LID) is often referred to as a solved problem in NLP research [35–40], especially for properly formatted documents, such as, books, newspapers, and other long-form digital texts. Language detection for tweets, however, is a much more challenging task due to the nature of the platform. Every day, millions of text snippets are posted to Twitter and written in many different 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 [41].

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. Many researchers have proposed manually-labeled datasets of Twitter messages [42–44], promisingly showing that most off-the-shelf LID methods perform relatively well when tested on annotated tweets.

Here, we use the LID model `FastText` [45, 46] to identify and explore the evolution of over 150 languages in over 118 billion messages collected via Twitter's 10% random sample (decahose) from 2008 to 2020 [47]. For messages posted after 2013, we also analyze language labels provided by Twitter's proprietary LID algorithm. 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 rather than share new content. Finally, we present some analytical results related to the contagion dynamics of Twitter.

II. TWEET LANGUAGE IDENTIFICATION

Several studies have looked closely at short-text LID [43, 48–56], particularly on Twitter where users are limited to a small number of characters per tweet (140 prior to the last few months of 2017, 280 thereafter [57]). These studies all share a strong consensus that short text language identification on Twitter is an exceptionally dif-

* thayer.alshaabi@uvm.edu
† peter.dodds@uvm.edu

ficult task.

Many methods have been proposed to classify the language of an individual tweet. Researchers have evaluated off-the-shelf LID tools on substantial subsets of Twitter data for a limited number of languages [42, 43]. For example, Google's Compact Language Detector (CLD-1 [58], and CLD-2 [59]) are 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 [60] 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 was also proposed by [43] that combines both Google's `CLD` and `langid` to improve classification accuracy for Twitter data.

As of early 2013, Twitter introduced language predictions classified by their internal algorithm in the historical data feed [61]. Since the LID algorithm used by Twitter is proprietary, we can only refer to a simple evaluation of their own model [62]. Our analysis of Twitter's language labels indicates that 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 location 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 predictor variables are found in the message itself.

`FastText` [45, 46, 63] is a recently proposed approach for text classification that uses pre-engineered n-gram features similar to the model described by [64]. `FastText` employs various tricks [24, 25, 63] in order to train a simple neural network using stochastic gradient descent and a linearly decaying learning rate for text classification. The model uses a hierarchical softmax function [45, 64] to efficiently compute the probability distribution over the predefined classes (i.e., languages). The authors show that `FastText` is on par with deep learning models [65–67] in terms of accuracy and consistency, yet orders of magnitude faster in terms of inference and training time [45, 46].

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 complected in case of a tie, or when models are completely unclear on a given tweet. However, 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 of the LID models outlined above 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.

Therefore, we use `FastText` to obtain language labels for tweets 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 in Ref. [48] and further tested in Ref. [68] and [43] to improve language classification. 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` 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 [69, 70]. We provide a list of all language labels assigned by `FastText` compared to the ones served by Twitter in Tab. S1.

III. COMPARISON WITH HISTORICAL FEED

We have collected a random 10% sample of all public tweets posted on the Twitter platform starting in September 2008. Using the steps described in Sec. II, we have implemented a pipeline to run `FastText` on this dataset. Our source code along with our documentation is publicly available online on a Gitlab repository [71]. Here, we evaluate our results by comparing the language labels obtained by `FastText` to those found in the metadata provided by Twitters 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 within ensembles of interconnected time series in Ref. [72]). Unfortunately, we are unable to objectively evaluate the performance of both classifiers due to the lack of verified ground-truth language labels, and some ambiguity in Twitter's sampling mechanism [73]. Nevertheless, we show that our results of language usage over time are on par with Twitter's estimation for most recent years.

We find that our classification of tweets using `FastText` notably improves the consistency of language labels when compared to the labels served with the his-

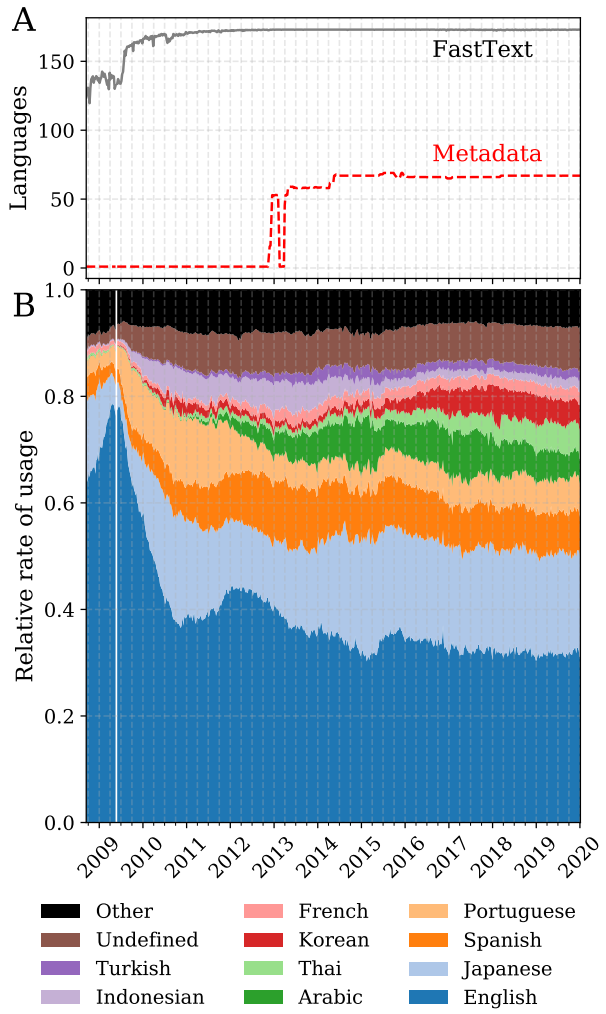

FIG. 1. **Language time series for the Twitter historical feed and `FastText` classified tweets. (A)** Number of languages reported by Twitter (red, dashed) and classified by `FastText` (black, solid) since September 2008. Fluctuations in late 2012 and early 2013 for the Twitter language time series are indicative of inconsistent classifications. **(B)** shows that the relative rate of usage by language using `FastText` was not severely affected by our missing data and maintained a consistent behavior throughout the last decade. The change in language distribution days when Twitter was relatively immature can be readily seen—for instance, English accounted for an exceedingly high proportion of activity on the platform, owing to Twitter's inception in an English speaking region.

torical feed. This observation is visible in the time series of the language classes shown in Fig. 1. Daily values are averaged weekly, so the y-axis reports results at a daily resolution. Fig. 1A shows that Twitter served a widely varying number of language tags for several months following the introduction of a language prediction. 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` 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 2000's—but stabilizes before Twitter introduces language labels. `FastText` classifies roughly 173 languages on average, including some rare languages, so the occasional dropout of a language seen in this time series is expected. We note that the fluctuations in the time series during the early years of Twitter (before 2012) and the first week of 2017 are primarily caused by some unexpected service outages which resulted in missing data. Nonetheless, Fig. 1B shows that the overall relative rate of usage by language was not impaired by the missing date, and maintained a consistent behavior throughout the last decade.

In order to take a closer look at the language labels classified by `FastText` compared to those found in the historical feed, we have collected both the language label predicted by Twitter and that obtained by `FastText` for every tweet in our dataset. We then computed confusion matrices to get an objective estimate of the agreement between the two classifiers on a large collection of tweets over time. Upon inspection of the computed confusion matrices, we find major disagreement during the first few years of Twitter's introduction of the LID feature to the platform—which is expected for several reasons outlined above. More importantly, both classifiers seem to agree on the predicted language of the majority of tweets, especially for recent years (see Figs. S1–S3). 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 [74–77]. On a similar note, the classifiers disagree on some of the tweets that were classified as Undefined. Interestingly, the two classifiers strongly disagree on tweets classified as Indonesian or Dutch. Again, there is no way for us to objectively evaluate the language labels of these tweets due to the lack of verified ground-truth language labels. Nevertheless, we show that our results of average language usage over time are on par with Twitters estimation for most recent years as illustrated in Fig. S8. Further analyses can be found Sec. V B.

IV. RESULTS AND DISCUSSION

Sociolinguistics is a field of study that explores how language evolves with respect to cultural norms, education, gender, and ethnicity among different societies [78–81]. Several studies have quantified language evolution on social media [82, 83], particularly on Twitter [84, 85]. In Fig. 2, we show yearly average rank of the 15 most used languages on Twitter between 2009 and 2019. This figure is an adaptation of the so-called sankey (alluviall) diagram [86–88], visualizing the flow of annual rank dynamics for a set of different languages. For ease of description,

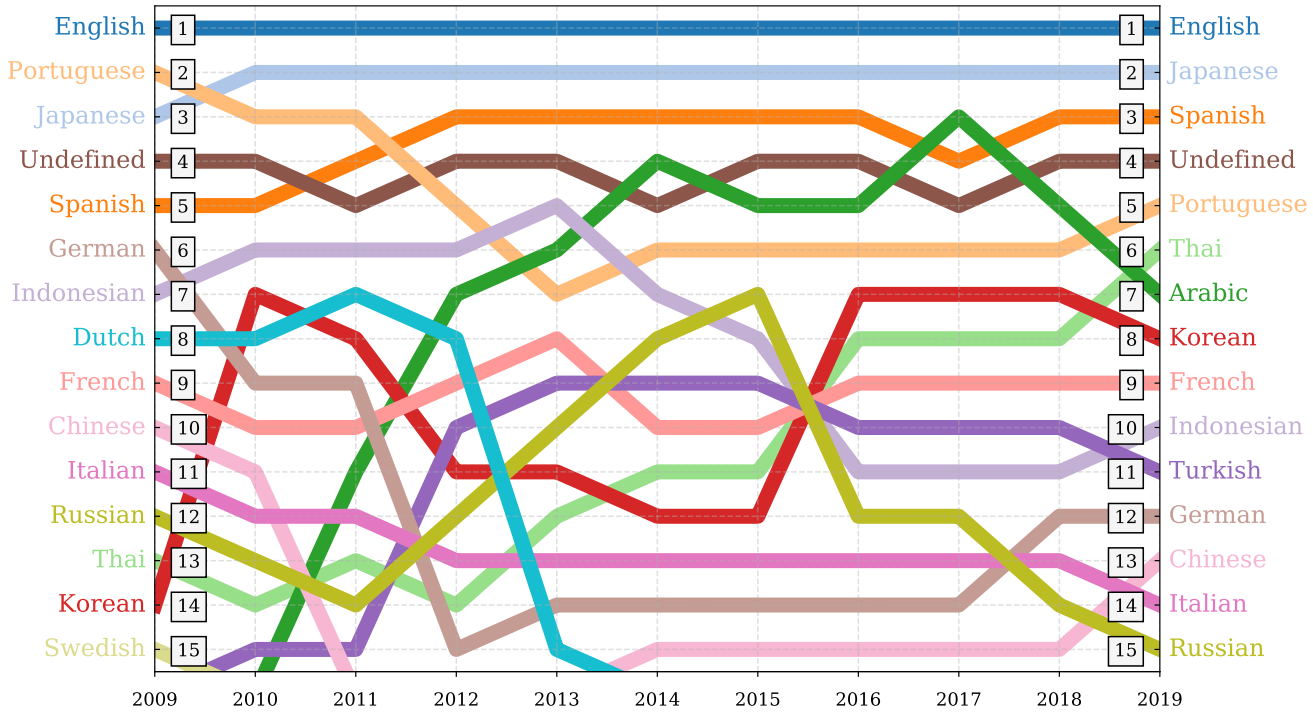

FIG. 2. **Alluvial plot visualizing the yearly average rank of the most used languages on Twitter between 2009 and 2019.** English and Japanese show the most consistent rank time series. Spanish, and Portuguese are also relatively stable over time. Undefined—which covers a wide variety of content such 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. It is worth noting that some languages like Russian, German, Indonesian, and Dutch moved down in rank. This shift is not necessarily due to a dramatic drop in the relative 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.

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 [89–92], K-pop, and political events [16]. Undefined is especially interesting as it covers a wide range of content such as emojis, memes, and other media shared on Twitter but would not be necessarily 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 relative rate of usage of these languages, but it is rather an artifact prompted by the growth of other languages on Twitter. We present some preliminary statistics of the number of messages captured in our dataset per language throughout the last decade in Figs. S9–S20.

A. Quantifying Twitter's Echo Chamber: 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" (RT) feature as a way of spreading information. Encoding a behavioral feature initially invented by users, Twitter formalized the retweet feature in November 2009 [93]. 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" [94] a message, distinct from a message reply [95].

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 different types:

'Organic Messages' (OT): All publicly available new content on the platform including *tweets*, *replies*, and *comments*, where $\mathcal{C}_{\ell,t}$ represent the number of messages in the decahose between times $t-1$ and t for a given language

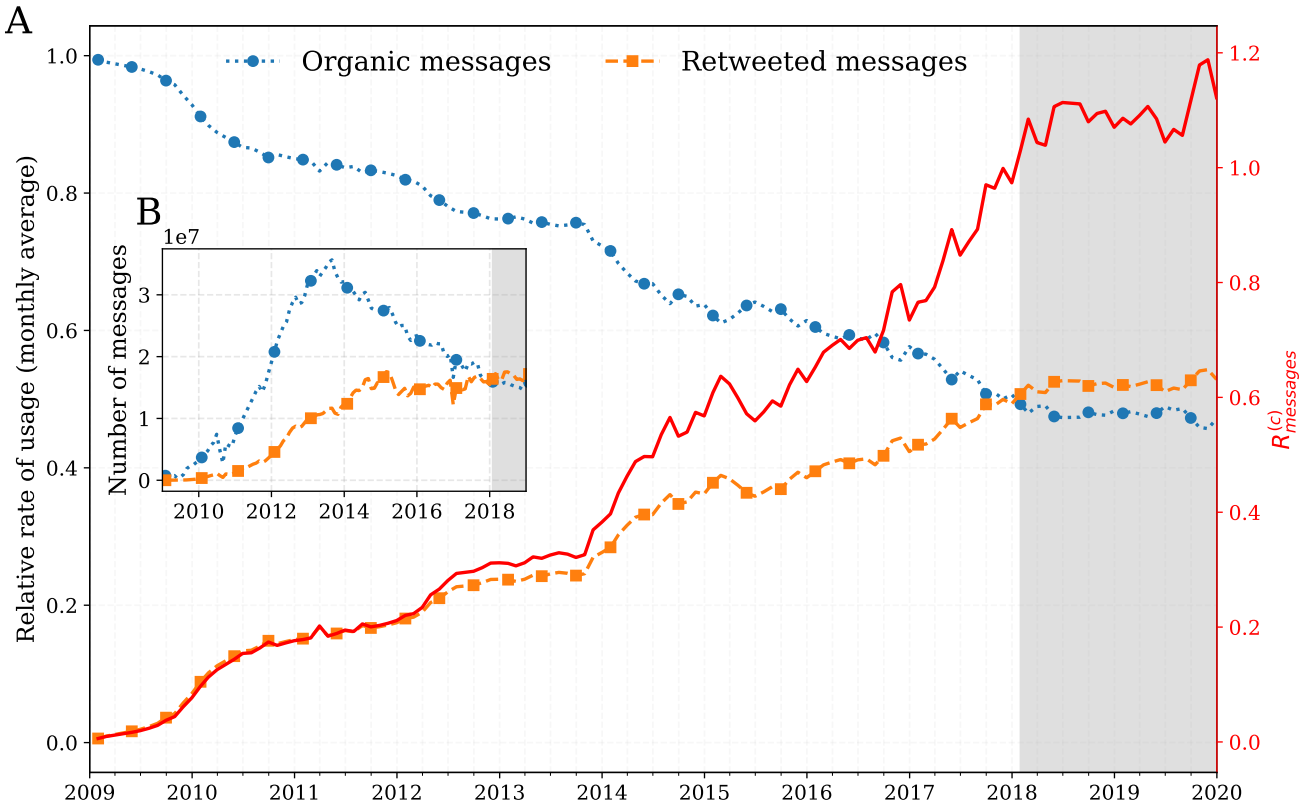

FIG. 3. **Timeseries for organic messages (blue), retweeted messages (orange), and average contagion ratio (red) for all languages.** Here we show monthly average ratio of retweet messages to organic messages for all languages. The dotted blue and dashed orange lines show a monthly average relative rate of usage for organic messages (*tweets*, *replies*, *comments*), versus retweeted messages respectively. The solid red line highlights the steady rise of the contagion ratio as defined in Sec. IV B. Inset **(B)** shows the number of organic messages compared to retweeted messages. The areas shaded in light grey starting in early 2018 highlights an interesting shift on the platform where the number of retweeted messages exceed the number of organic messages. An interactive version of the figure for all languages is available in an online appendix: http://compstorylab.org/share/papers/alshaabi2020a/ratio_timeseries.html

ℓ defined as follows:

$$\mathcal{O}_{\ell,t} = \frac{\mathcal{C}^{(\mathrm{OT})}_{\ell,t}}{\mathcal{C}^{(\mathrm{AT})}_{\ell,t}}, \tag{1}$$

where $\mathcal{C}^{(\mathrm{AT})}_{\ell,t}$ represent the overall number of messages in our dataset between times $t-1$ and t for a given language ℓ.

'Retweeted Messages' (RT): Repeated content (*retweets*) and the non-organic content found in Quote Tweets defined such that:

$$\mathcal{R}_{\ell,t} = \frac{\mathcal{C}^{(\mathrm{RT})}_{\ell,t}}{\mathcal{C}^{(\mathrm{AT})}_{\ell,t}}. \tag{2}$$

B. 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 contagion phenomenon. We compute the 'contagion ratio $R^{(c)}_{\mathrm{messages}}(\ell)$ as follows:

$$R^{(c)}_{\mathrm{messages}}(\ell) = \frac{\mathcal{C}^{(\mathrm{RT})}_{\ell,t}}{\mathcal{C}^{(\mathrm{OT})}_{\ell,t}}. \tag{3}$$

For all messages, in early 2018 the contagion ratio exceeded 1, indicating a higher number of retweeted messages than organic messages (Fig. 3). 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 message for every organic message at the end of 2019. In

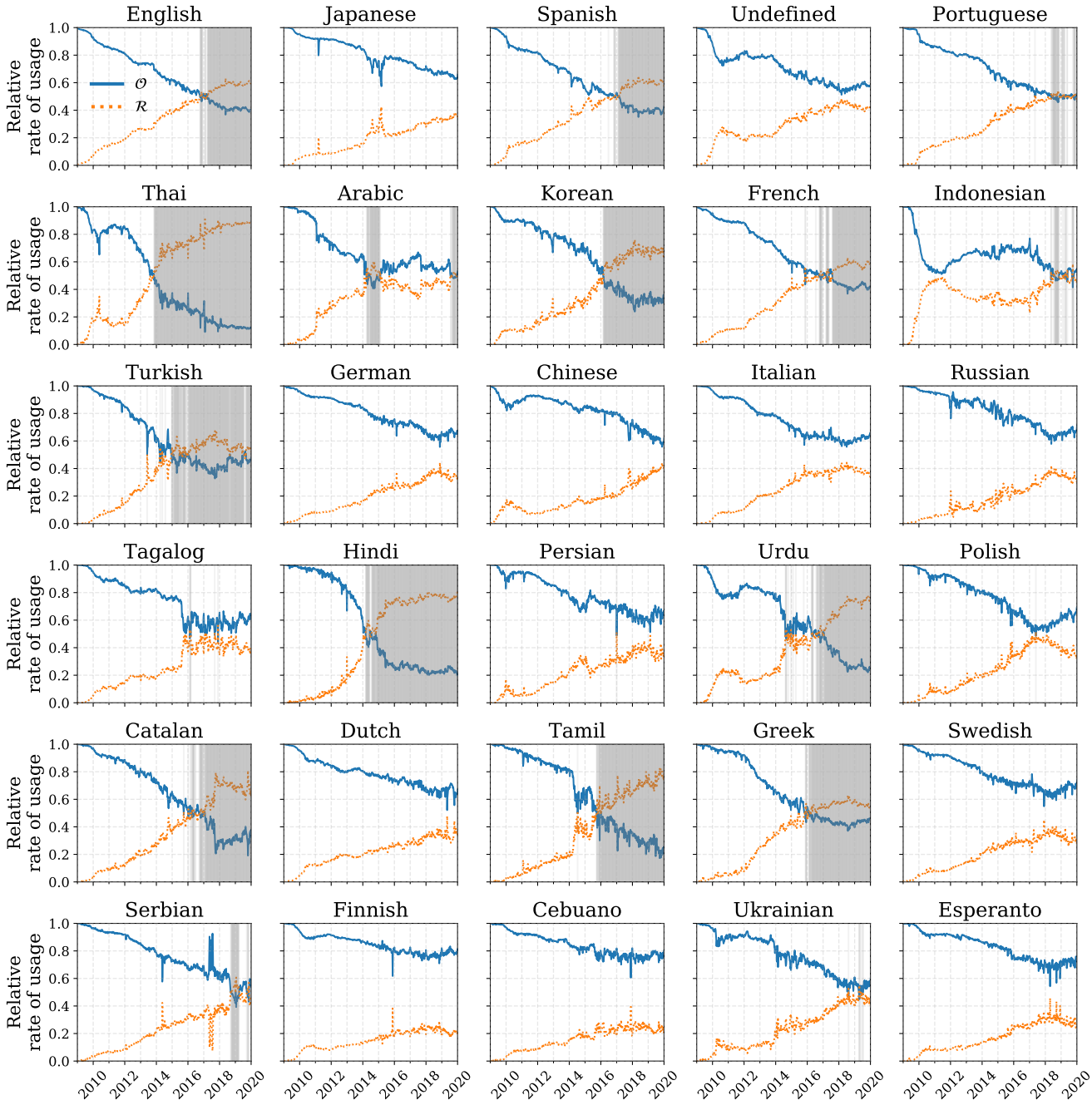

FIG. 4. **Weekly relative rate of usage of the top 30 ranked languages of 2019 (sorted by popularity).** We show the relative rate of `Organic Messages` $\mathcal{O}$ (in blue) compared to `Retweeted Messages` $\mathcal{R}$ (in orange). The areas highlighted in light shades of gray represent weeks where the relative 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/share/papers/alshaabi2020a/retweets_timeseries.html (53 MB).

Fig. 4, we show weekly aggregation of the relative rate of usage of the top 30 ranked languages of 2019. 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. Thai appears to be the first language to make this transition in late 2013.

The trend of increasing preference for retweeted messages is evident among most languages on Twitter, as illustrated in Fig. S4. In Fig. 5, we show a heatmap of the average contagion ratio for the top 30 most used

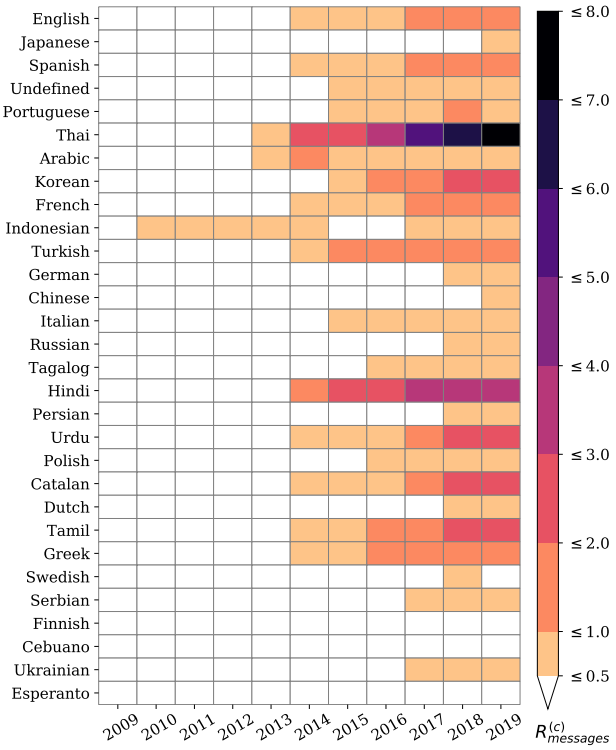

FIG. 5. **Timelapse of contagion ratios.** The average ratio is plotted against year for the top 30 ranked languages of 2019. Colored cells indicate a ratio higher than 0.5 whereas ratios below 0.5 are colored in white. Tab. S2 shows the top 10 languages with the highest average contagion ratio per year, while Tab. S3 shows the bottom 10 languages with the lowest average contagion ratio per year.

languages on Twitter per year. With the exception of Indonesian that showed a little bump between 2010 and 2013, most other languages began adopting 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. Interestingly, Japanese—the second most used language on the platform—does not exhibit this trend. Although most prevalent languages such as English, Spanish, Portuguese, Arabic, French, Indonesian, and Turkish, have a ratio higher than since 2017, there are a few languages that show a recent shift towards more organic content on the platform such as German, Russian, Polish, and Swedish. In Tab. S2 we show the top 10 languages with the highest average contagion ratio per year, while in Tab. S3 we show the bottom 10 languages with the lowest average contagion ratio per year.

Alternatively, we can borrow the concept of *gain* from signal processing to take a closer look at this interesting emerging behaviour of language dynamics on the platform. Gain is simply a contagion ratio amplifier. We present our findings using gain in Sec. V C. Similar to our observation of the contagion ratio, gain of retweets manifests the same behaviour indicating an increasing "market-share" for retweeted messages on the platform, particularly after 2018 (S5–S7).

There is a robust scaling relationship between number of messages and contagion ratio. We model this relationship using a Bayesian dynamic general linear model with measurement error, the details of which are presented on Appendix V D. We display this relationship and corresponding model fits in Fig. 6. In this figure we display only languages for which $R^{(c)}_{\text{messages}}(\ell) \in (0, 1)$ to focus on the dynamics of this region of ratio-space. However, we included all languages during model fitting. There is a significant linear relationship between $\log_{10}$ number of messages and $R^{(c)}_{\text{messages}}(\ell)$ for every year under study. The expected value of the slope of this linear relationship increases in each year. We conduct out-of-sample predictions and forecast that this increase in the slope of the linear relationship will continue during calendar year 2020.

V. CONCLUDING REMARKS

In this study, we present an alternative approach for obtaining language labels using `FastText` in order to overcome the challenge of missing labels in the decahose dataset. Our results comparing language usage over time largely agree with Twitter's estimation, particularly for recent years. However, `FastText` is not necessarily the best LID tool for language classification on Twitter. Future work ought to be done to improve language identification for short-text, particularly for social media outlets such as Twitter. We explored a new aspect of modern digital sociolinguistics on Twitter over the last decade. We found a recent tendency among most languages to *retweet* (spread information) rather than share new content. This recent rise of retweeted messages may suggest a systemic bias in the design of the platform, or perhaps human nature; it is much easier to repurpose or forward another individual's content than to post a new message. Different social and geographical communities have cultures of communication which will need to be explored in future work.

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. We thank Colin Van Oort, and Anne Marie Stupinski for their comments on the manuscript.

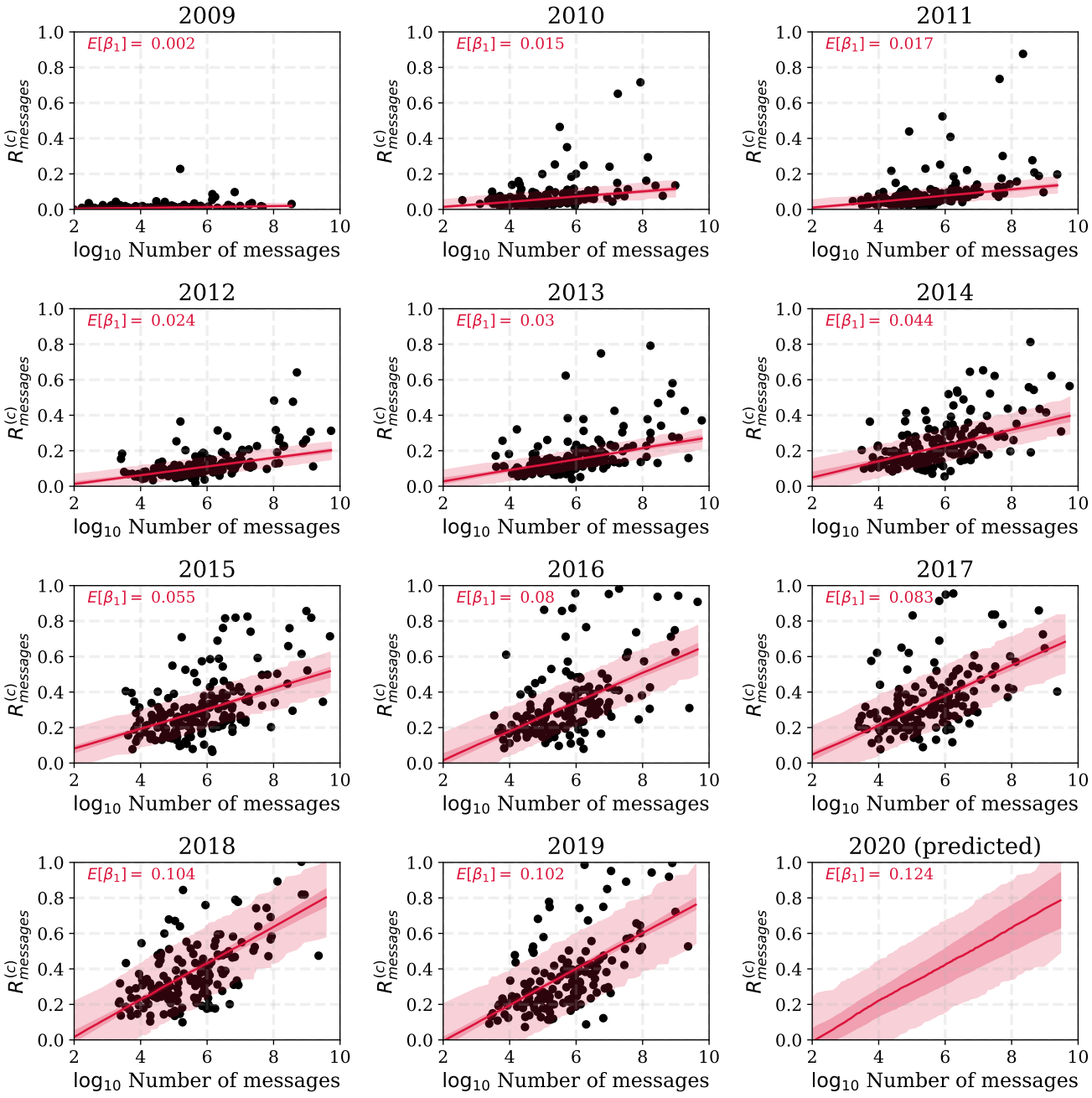

FIG. 6. **Contagion ratio of retweeted messages to organic messages as a function of number of messages by language.** We display points corresponding only to languages with a ratio value ranging between 0 and 1 to focus on the relationship in this region of ratio-space. However, we included all languages in our analysis. We fit a multi-stage Bayesian dynamic general linear model to this data. We describe the model in Appendix V D. There is a significant logarithmic scaling relationship between number of messages and the contagion ratio of retweeted messages to organic messages. The expected value of the slope of the linear model increases in each year under study. We perform out-of-sample predictions and forecast that this increase in slope will continue throughout the year 2020.

[1] Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLOS ONE*, 6(12):1–1, 12 2011.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] A. Mitchell and P. Hitlin. Twitter reaction to events often at odds with overall public opinion, 2019.

[7] S. Wojcik and A. Hughes. How Twitter users compare to the general public, 2019.

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

[9] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW 10, page 851860, New York, NY, USA, 2010. Association for Computing Machinery.

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

[11] A. Culotta. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA 10, page 115122, New York, NY, USA, 2010. Association for Computing Machinery.

[12] 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.

[13] 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.

[14] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler. Online social networks and offline protest. *EPJ Data Science*, 4(1):19, 2015.

[15] E. M. Cody, A. J. Reagan, P. S. Dodds, and C. M. Danforth. Public opinion polling with Twitter. *ArXiv*, abs/1608.02024, 2016.

[16] 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. *ArXiv*, abs/1910.00149, 2019.

[17] 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, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[18] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *KDD*, 2012.

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

[20] 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.

[21] 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.

[22] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[25] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[26] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzman, A. Joulin, and E. Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

[27] P. Fernandes, M. Allamanis, and M. Brockschmidt. Structured neural summarization. *arXiv preprint arXiv:1811.01824*, 2018.

[28] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[29] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*, 2019.

[30] M. D. Domenico, A. Lima, P.-N. Mougel, and M. Musolesi. The anatomy of a scientific rumor. In *Scientific reports*, 2013.

[31] U. Pavalanathan and J. Eisenstein. Confounds and consequences in geotagged Twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.

[32] K. Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[33] K. Kandhway. Modeling opinion dynamics in a social network using Markov random field. In *2018 10th International Conference on Communication Systems & Networks (COMSNETS)*, pages 631–636. IEEE, 2018.

[34] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425):374–378, 2019.

[35] G. Grefenstette. Comparing two language identification schemes. In *Proceedings of JADT*, volume 95, 1995.

[36] P. McNamee. Language identification: A solved problem suitable for undergraduate instruction. *J. Comput. Sci.*

Coll., 20(3):94101, Feb. 2005.

[37] 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).

[38] L. Grothe, E. W. De Luca, and A. Nürnberger. A comparative study on language identification methods. In *LREC*. Citeseer, 2008.

[39] 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.

[40] 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.

[41] 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, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[42] 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*, LSM 12, page 6574, USA, 2012. Association for Computational Linguistics.

[43] 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, Gothenburg, Sweden, Apr. 2014. Association for Computational Linguistics.

[44] J. Williams and C. Dagli. Twitter language identification of similar languages and dialects without ground truth. In *VarDial*, 2017.

[45] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[46] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[47] Twitter. Developer application program interface (API). https://developer.twitter.com/en/docs/ads/campaign-management/api-reference, 2019. [Online; accessed 01-October-2019].

[48] E. Tromp and M. Pechenizkiy. Graph-based n-gram language identification on short texts. *Proceedings of Benelearn 2011*, pages 27–34, 01 2011.

[49] 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.

[50] 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.

[51] K. Steinmetz. What Twitter says to linguists, Sep 2013.

[52] M. Goldszmidt, M. Najork, and S. Paparizos. Bootstrapping language identifiers for short colloquial postings. In *Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013)*. Springer Verlag, September 2013.

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

[54] 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, Lisboa, Portugal, Sept. 2015. Association for Computational Linguistics.

[55] 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.

[56] 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.

[57] A. Rosen. Tweeting made easier. https://blog.twitter.com/en_us/topics/product/2017/tweetingmadeeasier.html, 2017.

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

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

[60] 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.

[61] A. Roomann-Kurrik. Introducing new metadata for tweets. https://blog.twitter.com/developer/en_us/a/2013/introducing-new-metadata-for-tweets.html, 2013.

[62] https://blog.twitter.com/engineering/en\protect_us/a/2015/evaluating-language-identification-performance.html.

[63] 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.

[64] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[65] X. Zhang and Y. LeCun. Text understanding from scratch. *arXiv preprint arXiv:1502.01710*, 2015.

[66] Y. Xiao and K. Cho. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprint arXiv:1602.00367*, 2016.

[67] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun. Very deep convolutional networks for natural language processing. *arXiv preprint arXiv:1606.01781*, 2, 2016.

[68] 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.

[69] Twitter. Rules and filtering. https://developer.twitter.com/en/docs/tweets/rules-and-filtering/overview/premium-operators, 2019. [Online; accessed 01-October-2019].

[70] A. Phillips and M. Davis. Best current practice (BCP): Tags for identifying languages. Technical report, Network Working Group IETF, California, USA, Technical report, 2009.

[71] T. Alshaabi, J. Minot, and M. Arnold. Storywrangler: Twitter *n*-grams. https://gitlab.com/compstorylab/storywrangler, 2019.

[72] P. S. Dodds et al. Long-term word frequency dynamics derived from twitter are corrupted: A forensic data science approach to detecting and removing pathologies in ensembles of time series, 2020.

[73] J. Pfeffer, K. Mayer, and F. Morstatter. Tampering with Twitters sample api. *EPJ Data Science*, 7(1):50, 2018.

[74] H. Ringbom. *Cross-linguistic similarity in foreign language learning*, volume 21. Multilingual Matters, 2007.

[75] H. Borer. *Parametric syntax: Case studies in semitic and romance languages*, volume 13. Walter de Gruyter GmbH & Co KG, 2014.

[76] 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.

[77] 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.

[78] S. Pinker and P. Bloom. Natural language and natural selection. *Behavioral and brain sciences*, 13(4):707–727, 1990.

[79] M. D. Hauser, N. Chomsky, and W. T. Fitch. The faculty of language: What is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[80] R. Dunbar and R. I. M. Dunbar. *Grooming, gossip, and the evolution of language*. Harvard University Press, 1998.

[81] J. J. Gumperz and J. Cook-Gumperz. Studying language, culture, and society: Sociolinguistics or linguistic anthropology? 1. *Journal of Sociolinguistics*, 12(4):532–545, 2008.

[82] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *fourth international AAAI conference on weblogs and social media*, 2010.

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

[84] 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.

[85] 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.

[86] Sankey diagram.

[87] E. F. Sinar. Data visualization. In *Big Data at Work*, pages 129–171. Routledge, 2015.

[88] M. Rosvall and C. T. Bergstrom. Mapping change in large networks. *PloS one*, 5(1), 2010.

[89] P. N. Howard and M. M. Hussain. *Democracy's fourth wave?: Digital media and the Arab Spring*. Oxford University Press, 2013.

[90] 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.

[91] 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.

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

[93] B. Stone. Retweet limited rollout. https://blog.twitter.com/official/en_us/a/2009/retweet-limited-rollout.html, 2009.

[94] C. Shu. Twitter officially launches its retweet with comment feature. https://techcrunch.com/2015/04/06/retweetception/, 2015.

[95] B. Stone. Are you twittering @ me? https://blog.twitter.com/official/en_us/a/2007/are-you-twittering-me.html, 2007.

[96] G. K. Zipf. Human behaviour and the principle of least effort: An introduction to human ecology. 1949.

A. Tweet Language Identification

Algorithm 1 Tweet Language Identification

Input: Tweet text
Output: Language iso-code

```
text ← Filter(original_text)   ▷ Twitter-specific content
lang, conf ← FastText(text)    ▷ Language identification
if conf < .25 then
    return und
else
    return lang
end if
```

TABLE S1: Language codes for both FastText and Twitter language identification models

Language	FastText	Twitter
Afrikaans	af	-
Albanian	sq	-
Amharic	am	am
Arabic	ar	ar
Aragonese	an	-
Armenian	hy	hy
Assamese	as	-
Asturian	ast	-
Avaric	av	-
Azerbaijani	az	-
Bashkir	ba	-
Basque	eu	eu
Bavarian	bar	-
Belarusian	be	-
Bengali	bn	bn
Bihari	bh	-
Bishnupriya	bpy	-
Bosnian	bs	bs
Breton	br	-
Bulgarian	bg	bg
Burmese	my	my
Catalan	ca	ca
Cebuano	ceb	-
Cherokee	-	chr
Central-Bikol	bcl	-
Central-Kurdish	ckb	ckb
Chavacano	cbk	-
Chechen	ce	-
Chinese-Simplified	-	zh-cn
Chinese-Traditional	-	zh-tw
Chinese	zh	zh
Chuvash	cv	-
Cornish	kw	-
Corsican	co	-
Croatian	hr	-
Czech	cs	cs
Danish	da	da
Dimli	diq	-
Divehi	dv	dv
Dotyali	dty	-
Dutch	nl	nl
Eastern-Mari	mhr	-
Egyptian-Arabic	arz	-
Emiliano-Romagnolo	eml	-
English	en	en
Erzya	myv	-
Esperanto	eo	-
Estonian	et	et
Fiji-Hindi	hif	-
Filipino	-	fil
Finnish	fi	fi
French	fr	fr
Frisian	fy	-
Gaelic	gd	-
Gallegan	gl	-
Georgian	ka	ka
German	de	de
Goan-Konkani	gom	-
Greek	el	el
Guarani	gn	-
Gujarati	gu	gu
Haitian	ht	ht
Hebrew	he	he
Hindi	hi	hi
Hungarian	hu	hu
Icelandic	is	is
Ido	io	-
Iloko	ilo	-
Indonesian	id	in
Inuktitut	-	iu
Interlingua	ia	-
Interlingue	ie	-
Irish	ga	-
Italian	it	it
Japanese	ja	ja
Javanese	jv	-
Kalmyk	xal	-
Kannada	kn	kn
Karachay-Balkar	krc	-
Kazakh	kk	-
Khmer	km	km
Kirghiz	ky	-
Komi	kv	-
Korean	ko	ko
Kurdish	ku	-
Lao	lo	lo
Latin	la	-
Latvian	lv	lv
Lezghian	lez	-
Limburgan	li	-
Lithuanian	lt	lt
Lojban	jbo	-
Lombard	lmo	-
Lower-Sorbian	dsb	-
Luxembourgish	lb	-
Macedonian	mk	-
Maithili	mai	-
Malagasy	mg	-
Malayalam	ml	ml
Malay	ms	msa
Maltese	mt	-
Manx	gv	-
Marathi	mr	mr
Mazanderani	mzn	-
Minangkabau	min	-
Mingrelian	xmf	-
Mirandese	mwl	-
Mongolian	mn	-

Nahuatl	nah	-
Neapolitan	nap	-
Nepali	ne	ne
Newari	new	-
Northen-Frisian	frr	-
Northern-Luri	lrc	-
Norwegian	no	no
Nynorsk	nn	-
Occitan	oc	-
Oriya	or	or
Ossetic	os	-
Pampanga	pam	-
Panjabi	pa	pa
Persian	fa	fa
Pfaelzisch	pfl	-
Piemontese	pms	-
Polish	pl	pl
Portuguese	pt	pt
Pushto	ps	ps
Quechua	qu	-
Raeto-Romance	rm	-
Romanian	ro	ro
Russian-Buriat	bxr	-
Russian	ru	ru
Rusyn	rue	-
Sanskrit	sa	-
Sardinian	sc	-
Saxon	nds	-
Scots	sco	-
Serbian	sr	sr
Serbo-Croatian	sh	-
Sicilian	scn	-
Sindhi	sd	sd
Sinhala	si	si
Slovak	sk	-
Slovenian	sl	sl
Somali	so	-
Shona	-	sn
South-Azerbaijani	azb	-
Spanish	es	es
Sundanese	su	-
Swahili	sw	-
Swedish	sv	sv
Tagalog	tl	tl
Tajik	tg	-
Tamil	ta	ta
Tatar	tt	-
Telugu	te	te
Thai	th	th
Tibetan	bo	bo
Tosk-Albanian	als	-
Turkish	tr	tr
Turkmen	tk	-
Tuvinian	tyv	-
Uighur	ug	ug
Ukrainian	uk	uk
Upper-Sorbian	hsb	-
Urdu	ur	ur
Uzbek	uz	-
Venetian	vec	-
Veps	vep	-
Vietnamese	vi	vi
Vlaams	vls	-
Volapk	vo	-
Walloon	wa	-
Waray	war	-
Welsh	cy	cy
Western-Mari	mrj	-
Western-Panjabi	pnb	-
Wu-Chinese	wuu	-
Yakut	sah	-
Yiddish	yi	-
Yoruba	yo	-
Yue-Chinese	yue	-
Unknown	-	unknown
Undefined	und	und

B. Analytical comparison to the Decahose

In Fig. S2, we show the normalized ratio difference (Divergence) between the two classifiers for all activities between 2014 and 2019. Divergence is calculated as

$$\delta D_\ell = \left| \frac{\texttt{ft}_\ell - \texttt{tw}_\ell}{\texttt{ft}_\ell + \texttt{tw}_\ell} \right|, \tag{4}$$

where `ft` is the number of messages captured by `FastText` LID for language ℓ, and `tw` is the number of messages captured by `Twitter` LID for language ℓ.

In Fig S2A–B we show Zipf distributions [96] of all languages captured by `FastText`, Twitter's language identification algorithm(s) respectively. `FastText` recorded a total of 173 unique languages, whereas `Twitter` captured a total of 73 unique languages throughout that period. It is worth noting that, that some of the languages reported by Twitter were experimental and no longer available in recent years. As mentioned before, the two classifiers agree on most prevalent languages on the platform indicated by points near the vertical dashed gray line in Panel (C), specifically that they captured a similar number of activities between 2014 and end of 2019.

However, languages found left of this line are more prominent using the `FastText` LID model e.g., Chinese (zh), Central-Kurdish (ckb), Uighur (ug), Sindhi (sd). On the other hand, languages right of the line are identified more frequently by the `Twitter` LID model(s) e.g., Estonian (et), Haitian (ht). Languages found within the light-blue area exist in one classifier but not in the other such as Esperanto (eo), Interlingua (ia), Afrikaans (af), Inuktitut (iu), Cherokee (chr), Senegal (sn). It is worth noting that unknown is an artifical label that we added to flag messages with missing language labels in the metadata of our dataset.

C. Gain (Ratio Amplifier)

We define the 'contagion ratio $R^{(c)}_{\text{messages}}(\ell)$ as the ratio of retweeted messages to organic messages (Eq. 3). The contagion ratio is an intuitive and interpretable analytical measure. Alternatively, we can borrow the concept of *gain* from signal processing to take a closer look at this

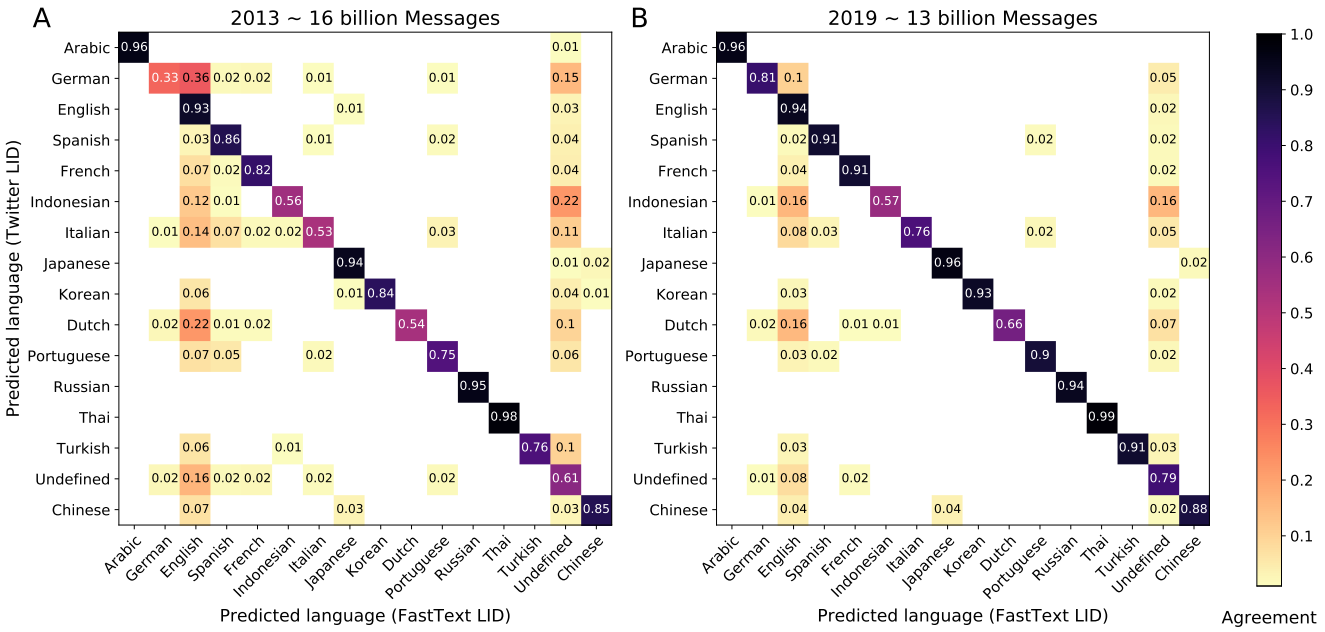

FIG. S1. **Language Identification Confusion Matrices.** Here we show a subset of the full confusion matrix for languages in the top-15 most frequently seen on Twitter. In Panel **(A)** we show the confusion matrix for tweets authored in 2013. The matrix indicates substantial disagreement between the two classifiers during 2013, the first year of Twitter's efforts to provide language labels. In Panel **(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.

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

TABLE S2. Top 10 languages with the highest average yearly contagion ratio (sorted by 2019).

interesting emerging behaviour of language dynamics on the platform.

Let $\mathcal{A}_t$ be the amount of signal observed at time step t for a given arbitrary system, where $\mathcal{X}_t$ is the contribution of the process we're interested in measuring, and $\mathcal{O}_t$ represents the amount of remaining signal such that:

$$\mathcal{A}_t = \mathcal{X}_t + \mathcal{O}_t. \tag{5}$$

Gain is then formally defined as

$$G_{\mathcal{X}}^{(t)} = 10 \log_{10} \left(\frac{\mathcal{A}_t}{\mathcal{A}_t - \mathcal{O}_t} \right). \tag{6}$$

Hence, gain measures the increase of $\mathcal{X}_t$ with respect to the amount of signal passed through the system. We note that the leading factor of 10 is simply a constant carried over from conversion to logarithmic Decibel (dB) units. In our application, however, we can redefine gain such that:

$$G_{\text{messages}}^{(c)}(\ell) = 10 \log_{10} \left[\frac{\mathcal{C}_{\ell,t}^{(AT)}}{\mathcal{C}_{\ell,t}^{(OT)}} \right], \tag{7}$$

where $\mathcal{C}_\ell^t$ represent the number of messages captured between time step $t-1$ and time step t for a given target

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 S3. Bottom 10 languages with the lowest average yearly contagion ratio (sorted by 2019).

language ℓ.

Similar to our observation of the ratio of retweeted messages to organic messages, gain manifests the same behaviour indicating an increasing marketshare for retweeted messages on the platform, particularly after 2018 as seen in Fig. S5. More importantly, gain represent an analytical approach of investigating the rise of retweeted content among different languages relative to the overall usage, or the number of messages captured for each language. Unlike ratio, gain allows us to factor in the number of messages for any given language, and then normalize out organic messages to explore how the signal of retweeted messages changes as the number of organic messages up or down.

Again, the trend of retweeted messages is readily seen among most languages on Twitter as shown in both Fig. S6 and Fig. S7. This time, however, gain offers a much finer resolution than ratio to explore this dramatic change over time. Compiling everything together, Fig. S6 shows average gain of retweeted messages as a function of the number of messages by language. One can arguably view this as a multi-objective optimization task to investigate whether maximizing spread/flow of information (ratio/gain) is correlated with maximizing the number of message for a given language. We note that a similar sort of scaling between number of messages and gain is evidently seen here as well. In fact, using gain we can see that the dominant languages on the platform with high values of gain now lie on the Pareto-Front (see Fig. S6, 2019). Points on the Pareto frontier are not dominated by any other points. In other words, `kn` is not on the Pareto frontier because it is dominated by several points, whereas `en`, `es` are not dominated by any other languages in 2019, and thus lie on the Pareto front. Moreover, the spread of languages in the gain-space seems to be growing in time. Although the trend for larger slope of gain as a function of number of messages continue to go up over time, the variance at the higher end of number of messages also seems to be increasing.

Furthermore, the finer resolution of gain compared to ratio becomes very apparent when we look at the heatmap of the average gain of retweeted messages in Fig. S7. We show average gain of every quarter of the year for the top 30 most used languages on Twitter in 2019. Some languages started to exhibit such behaviour as early as 2010. Among many other languages, Thai and Hindi seem to have the most dramatic shift in terms of the number of retweeted messages starting in early 2014. A few languages, on the other hand, exhibit major spikes only within specific years. For example, Indonesian has a spike that starts late 2010 and continue for about a year. Similarly, Arabic has a subsequent spike around 2014 followed by a spike in Turkish that starts in 2016 and peaks in late 2017.

D. Statistical analysis

We fit a multi-stage Bayesian dynamic general linear model (GLM) to model the relationship between $\log_{10} N$ and $R^{(c)}_{\text{messages}}(\ell)$. (We have denoted number of messages by N.) Though N is a discrete quantity, $\log_{10} N$ is a real-valued random variable, hence our usage of continuous latent variables throughout the model. We subscript variables with t to denote explicit dependence on time t measured in years. The model is composed of several pieces.

We model the generative process of $\log_{10} N$ as $\log_{10} N_t \sim \text{Skew-Normal}(\mu_t, \tau_t, \alpha_t)$. The parameter μ_t is the mean, while τ_t is the precision (inverse variance) and α_t is the skew. We chose a skewed-normal distribution because, though the distribution of $\log_{10} N$ does not have heavy tails, it does exhibit nonzero skew. Though we use the observed $\log_{10} N$ as the design variable in the GLM component of the model and not the latent μ_t, we model $\log_{10} N$ because we want to predict out-of-sample $\log_{10} N$ for $t = 2020$, which we accomplish using the predicted μ_t, τ_t, and α_t. We modeled the parameters of the skew-normal distribution using moderately-informative priors which we chose by inspecting the data: $\mu_t \sim \text{Normal}(5, 1)$, $\tau_t \sim \text{Gamma}(10, 1)$, and $\alpha_t \sim \text{Normal}(1, 1)$.

Given $\log_{10} N$, we fit a GLM of the form

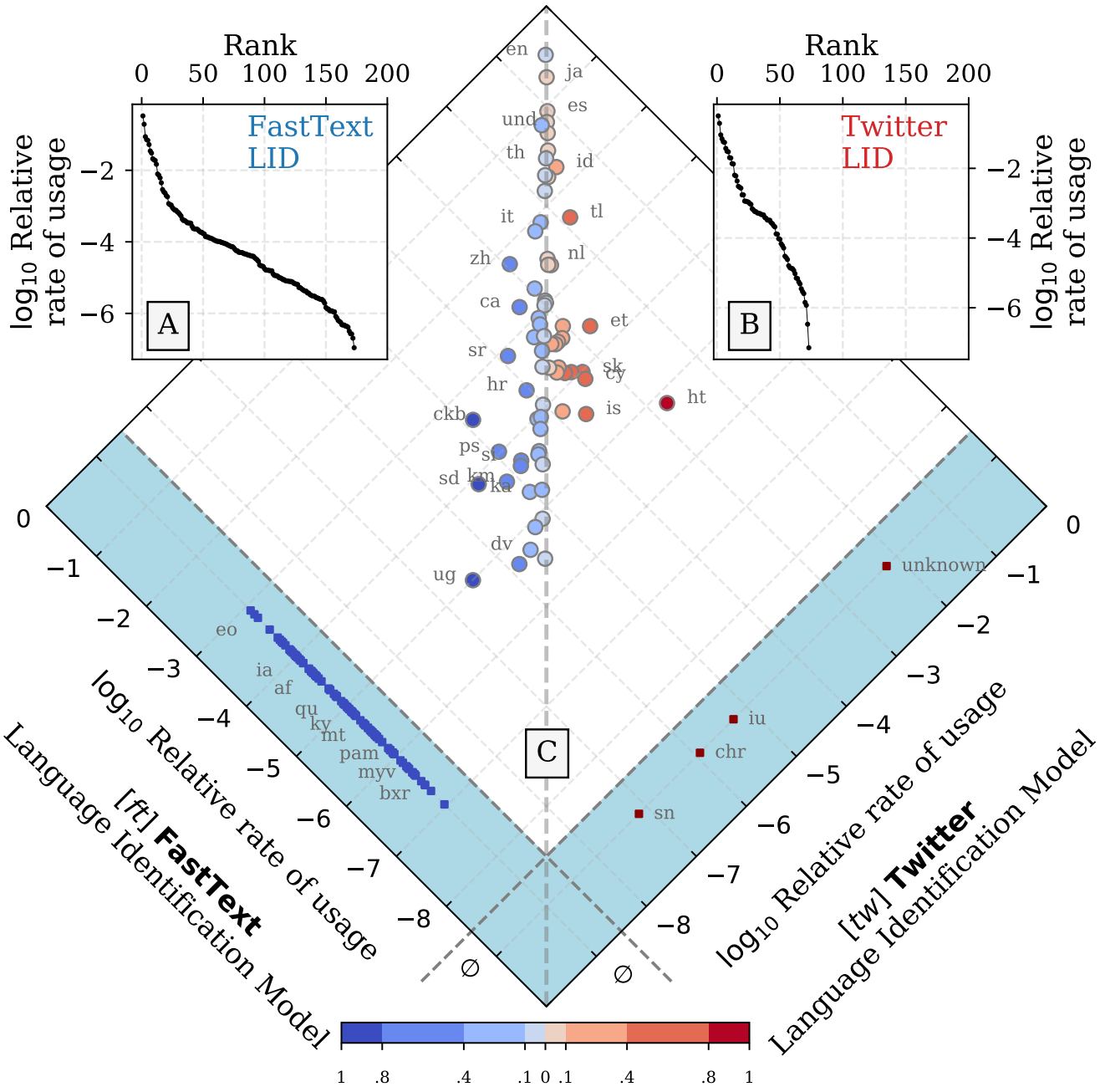

FIG. S2. **Language identification divergence.** Panel **(A)** shows a Zipf distribution [96] of all languages captured by `FastText` LID model, while Panel **(B)** shows the same distribution for languages captured by Twitter's language identification algorithm(s). The y-axis in both panels reports the relative rate of usage of all messages between 2014 and 2019, while the x-axis shows the rank of the corresponding language. `FastText` recorded a total of 173 unique languages throughout that period, some of which are considered rare languages. On the other hand, `Twitter` captured a total of 73 unique languages throughout that same period, some of which were experimental and no longer available in recent years. In panel **(C)**, languages located near the vertical dashed gray line signify agreement between `FastText` and Twitter's language-classifiers, specifically that they captured a similar number of activities 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 the `Twitter` LID model. Languages found within the light-blue area exist in one classifier but not in the other, where `FastText` is colored in blue and `Twitter` is colored in red. The color of the points highlights the normalized ratio difference (Divergence) between the two classifiers. Divergence is calculated as $|(\texttt{ft}_\ell - \texttt{tw}_\ell)/(\texttt{ft}_\ell + \texttt{tw}_\ell)|$, where `ft` is the number of messages captured by `FastText` LID for language ℓ, and `tw` is the number of messages captured by `Twitter` LID for language ℓ. Hence, points with darker colors indicate greater disagreement between the two classifiers as shown in the colorbar at the bottom of the plot. A lookup table for language labels can be found in the Appendix S1, and an online appendix of all languages is also available here: http://compstorylab.org/share/papers/alshaabi2020a/fasttext_twitter_timeseries.html.

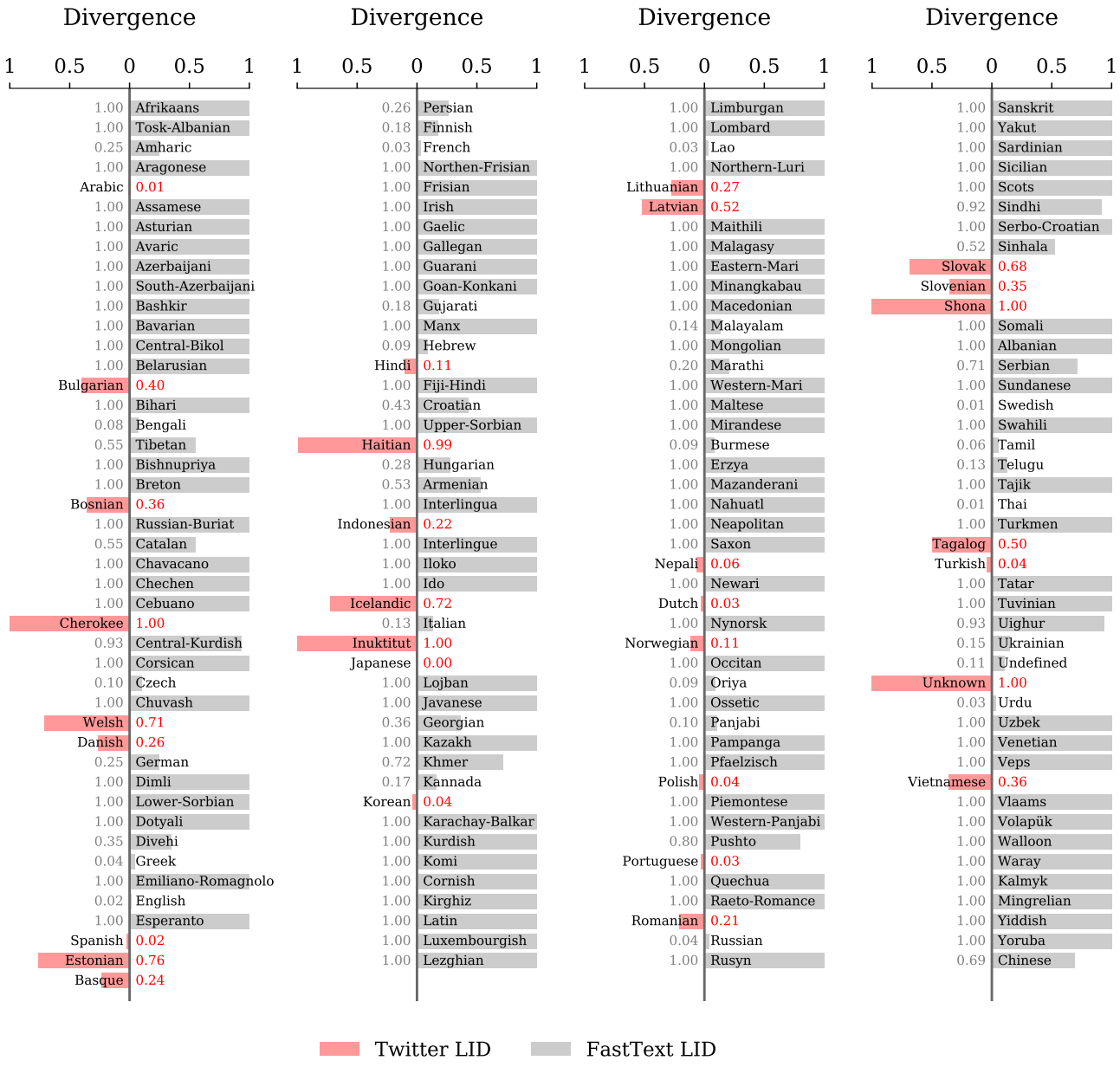

FIG. S3. **Relative language identification divergence.** A divergence value closer to zero indicate strong agreement between the two classifiers as they both captured approximately the same number of messages over the last decade. As the bars diverge away from the center we note higher relative rate of messages captured by one of the classifiers but not the other where `FastText LID` is highlighted in gray and `Twitter LID` highlighted in red.

$R^{(c)}_{\text{messages}}(\ell t) \sim \text{Laplace}(\beta_{0t} + \beta_{1t} \log_{10} N_{\ell,t}, b_t)$. We chose Laplace-distributed errors to account for the increased variance and heteroskedasticity in the distribution of $R^{(c)}_{\text{messages}}(\ell)$; we are concerned with the median behavior of this distribution for the purposes of this model, not the effects of outliers on the model. We placed centered normal priors on the regression coefficients, $\beta \sim \text{Normal}(0, 1)$, and a weakly-informative prior on the scale parameter, $b \sim \text{Inverse-Gamma}(6, 1)$. We fit a GLM of this form for each year $t \in \{2009, ..., 2019\}$.

We believe it is unlikely that the parameters of each GLM are independent of the parameters of the previous GLM; parameters of years $t + 1$ likely depend on parameters of year t. Collecting the vector of parameters as $z_t = (\mu_t, \tau_t, \alpha_t, \beta_{0t}, \beta_{1t}, b_t)$, we model this intertemporal dependence as $z_t \sim \text{Multivariate-Normal}(z_{t-1}, L)$. The lower triangular matrix L is the Cholesky decomposition of the covariance matrix of this process. The

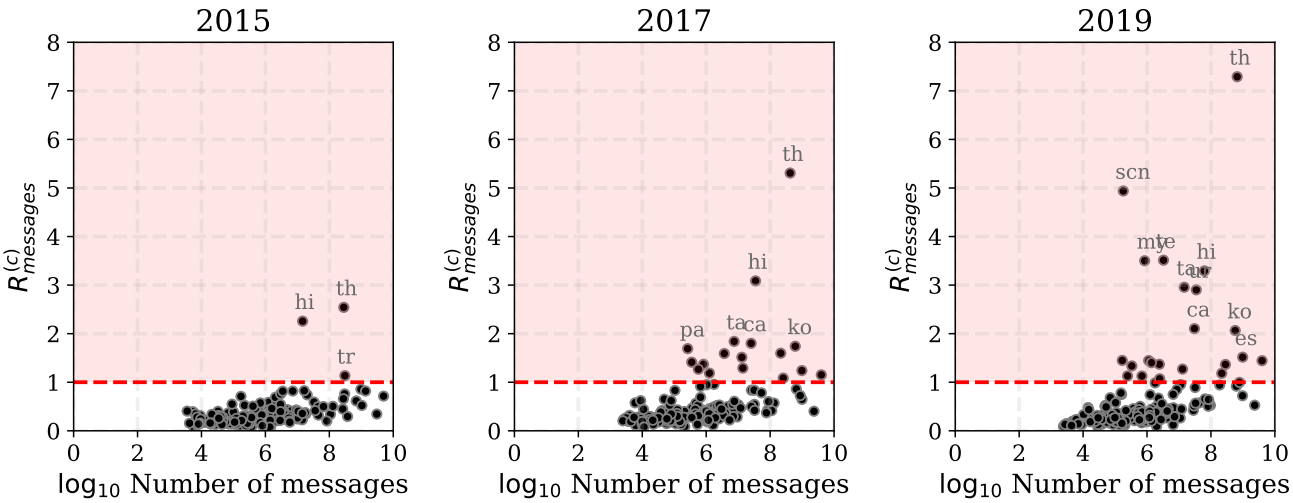

FIG. S4. **Yearly average contagion ratio for each language, plotted against the number of messages for 2015, 2017, and 2019.** The areas highlighted in red indicate a higher number of retweeted messages than organic messages. Several languages are highlighted including Thai (th), Hindi (hi), Korean (ko), Urdu (ur), Catalan (ca), and Tamil (ta). Japanese (ja)—the second most used language on the platform—does not exhibit this trend. A list of language ISO-codes is provided in Table S1.

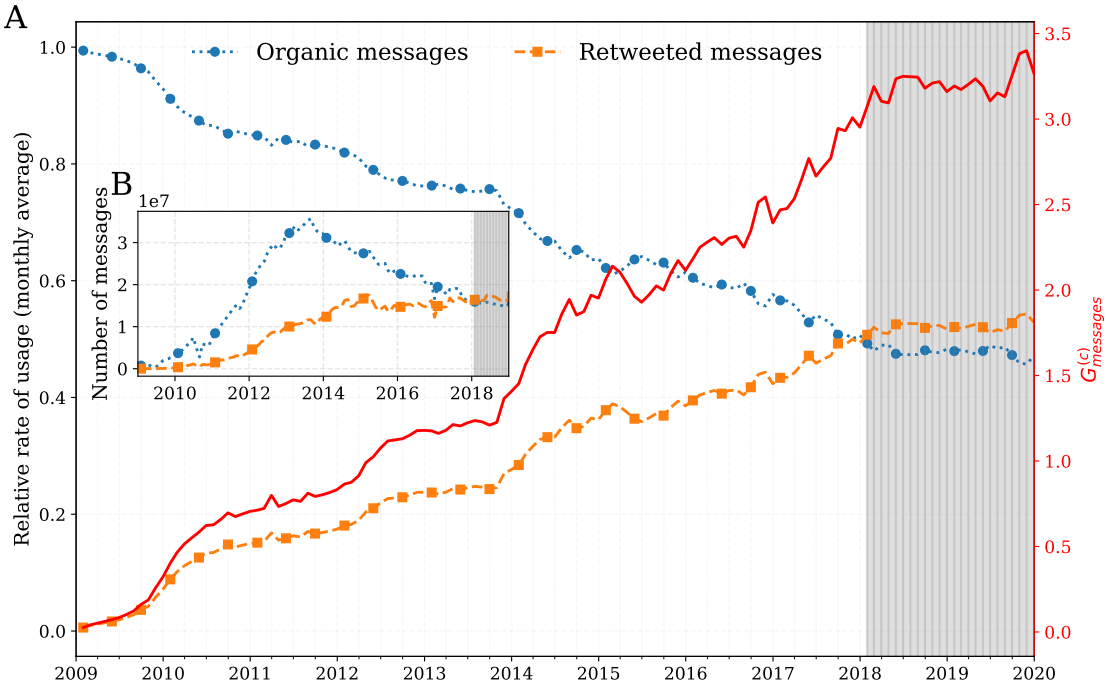

FIG. S5. **Timeseries for organic messages (blue), retweeted messages (orange), and gain (red) for all languages.** We show monthly average of all messages on the platform as a function of time. The blue and orange lines show a monthly average relative rate of usage for organic messages (*tweets*, *replies*, *comments*), and retweeted messages respectively. The red line highlights the steady rise of the gain of retweeted messages $G^{(c)}_{\text{messages}}$ as defined in Section V C. Inset **(B)** shows the number of organic messages compared to retweeted messages. The areas shaded in light grey starting from early 2018 highlights an interesting shift on the platform where the number of retweeted messages exceed the number of organic messages in our dataset. An interactive version of the figure for all languages is available in an online appendix: http://compstorylab.org/share/papers/alshaabi2020a/gain_timeseries.html

covariance matrix is given by $\Sigma = \sigma^T R \sigma$, where $\sigma \sim$ Log-Normal$(0, 1)$ (the isotropic multivariate lognormal

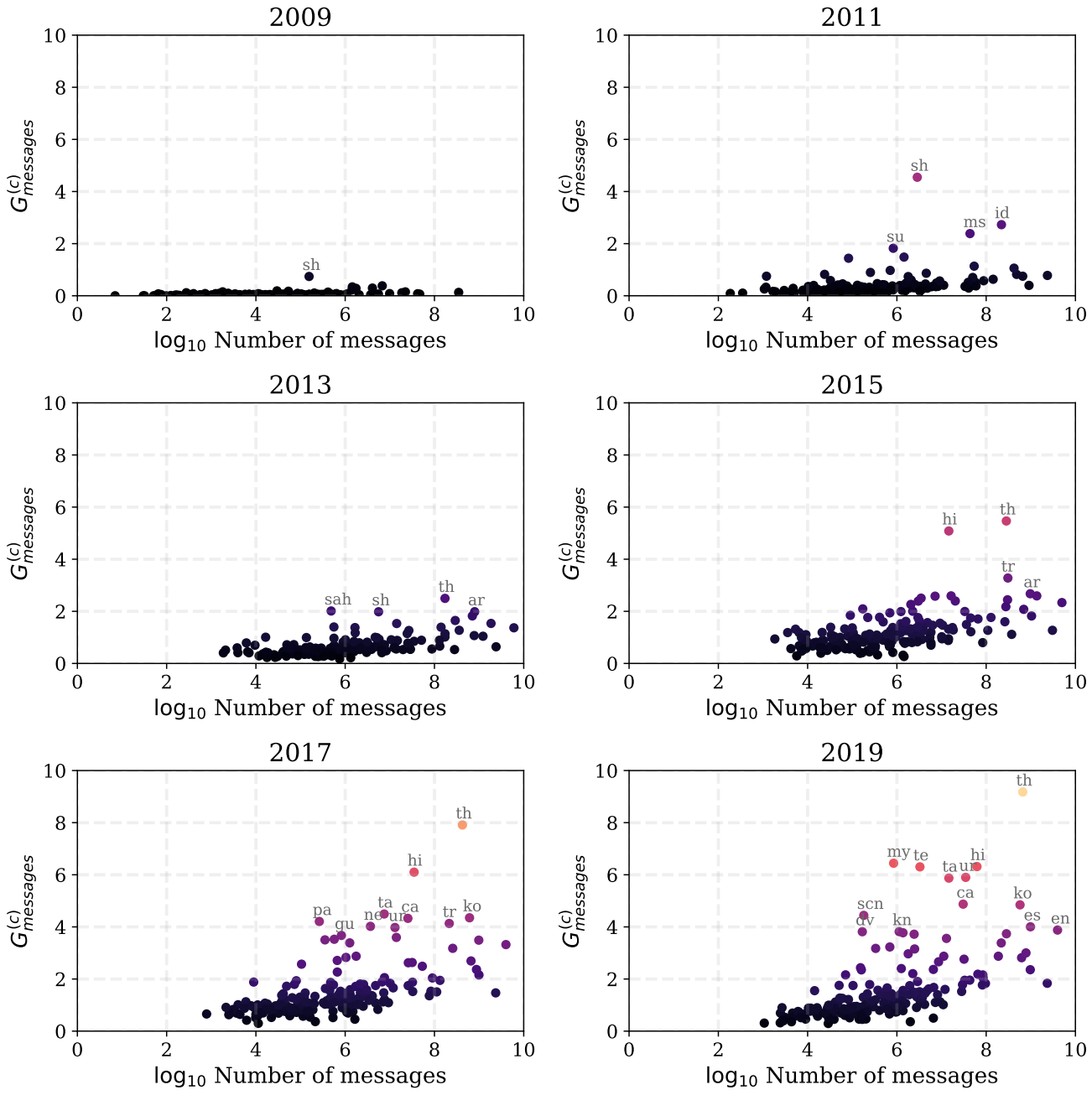

FIG. S6. **Yearly average of gains as a function of number of messages by language.** Evidently, we see the same sort of exponential relationship between number of messages and gain of retweeted messages as the one seen in ratio-space. Unlike ratio, using gain allows us to account for the usage of every language in the dataset with respect to other languages on the platform.

distribution) and the correlation matrix $R \sim \text{LKJ}(\eta)$, with $\eta = 2$. The density of the LKJ distribution is given as $p(R) \propto |R|^{\eta-1}$. We construct pseudo-observations for this process using the expected values of each of the parameters at each timestep. That is, we fit the parameters of the random walk using the pseudo-observations

$$E[z_t] = \int dz_t \; z_t \; p(z_t | \log_{10} N_t, R^{(c)}_{\text{messages}}(t))$$

generated by the posterior. We did this to reduce the time complexity of model fitting.

After fitting the random walk model, we are able to

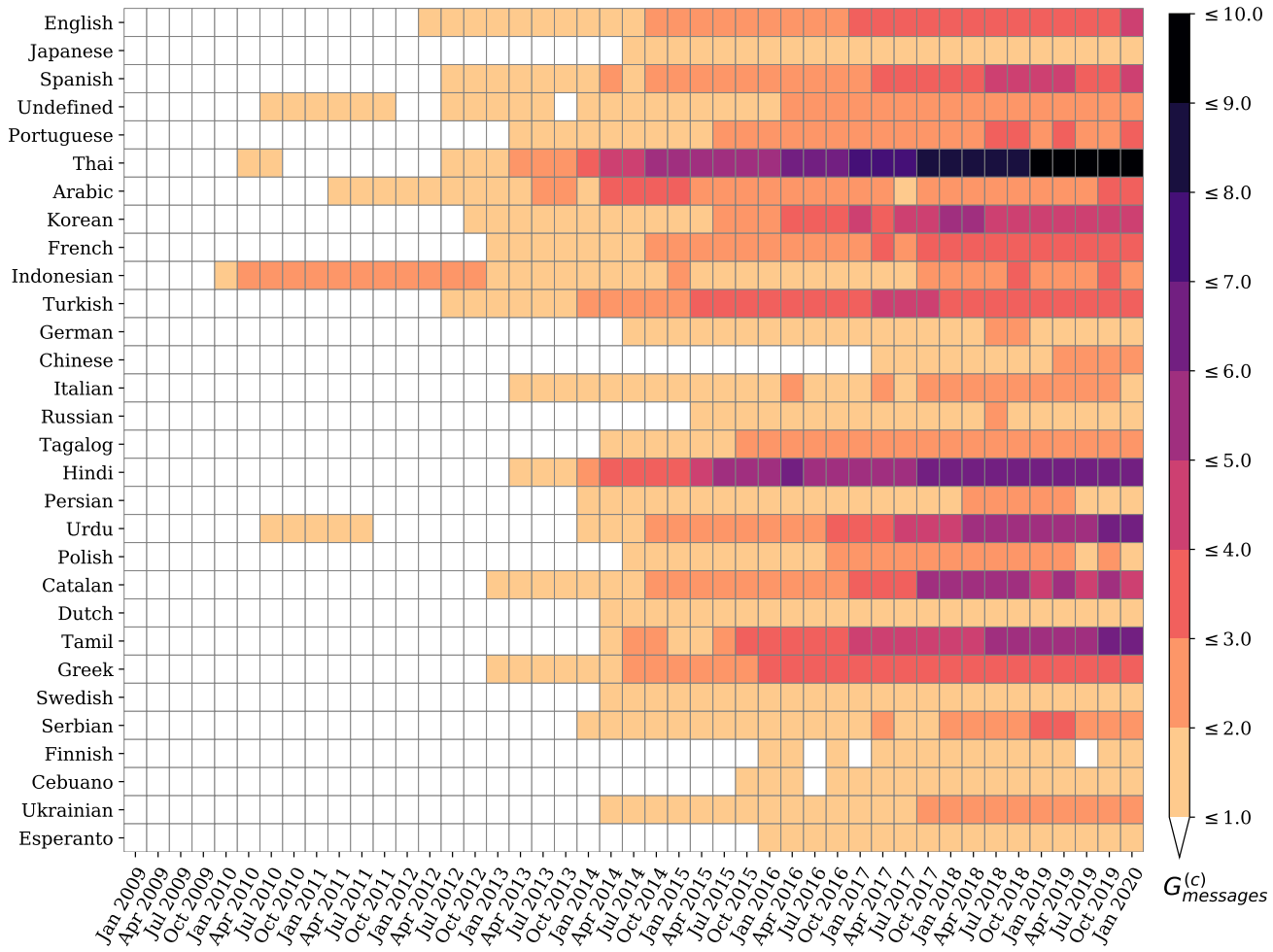

FIG. S7. **Timelapse of gains.** Here we show average gain as a function of time for the top 30 ranked languages of 2019. Every cell represents average gain of every quarter of the year. Colored cells indicate a gain higher than 1 whereas gain values below 1 are colored in white.

forecast the statistical relationship between $\log_{10} N_t$ and $R^{(c)}_{\text{messages}}(t)$ for $t = 2020$. We evolve the random walk one step (one year) forward in time, and then use the predicted values of μ_t, τ_t, and α_t to generate a synthetic dataset of $\log_{10} N_t$. We then apply the GLM to this dataset using the predicted values of β_{0t}, β_{1t}, and b_t.

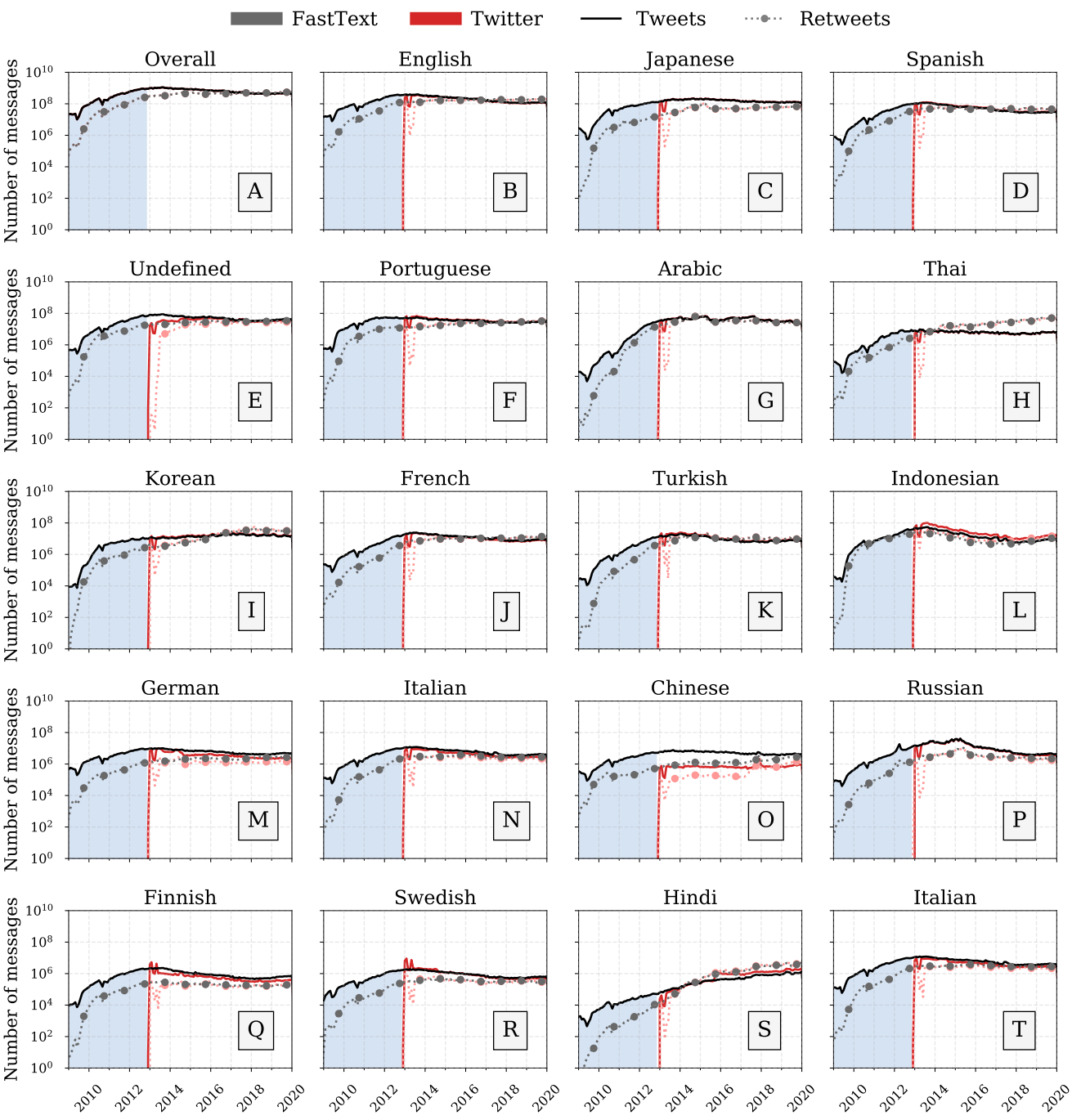

FIG. S8. **A comparison of language time series classified by `FastText` LID and Twitter LID for the top 15 languages in 2019.** We show the total number of messages (tweets + retweets) captured monthly by `FastText` (in black) and Twitter (in red) for the last decade. The areas highlighted in light shades of blue represent messages with missing language labels that are not captured by the Twitter LID model. Total number of tweets is indicted by solid lines, while the total number of retweets is indicated by dotted lines. The first subplot in the upper left corner shows the total number of messages of all languages on Twitter. An online appendix of all languages is available here: http://compstorylab.org/share/papers/alshaabi2020a/fasttext_twitter_timeseries.html

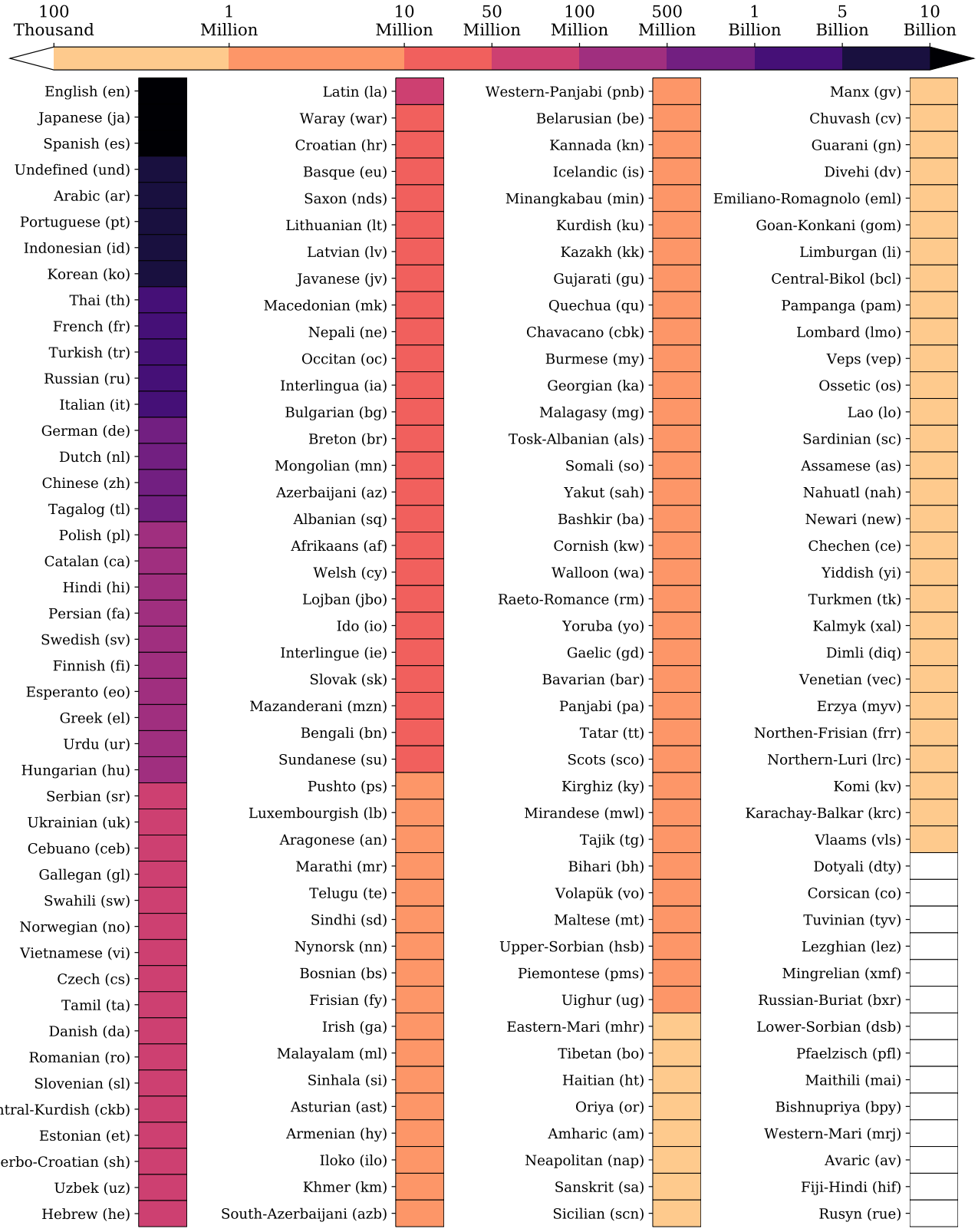

FIG. S9. **Overall dataset statistics.** Number of messages captured in our dataset as classified by `FastText` LID algorithm between 09/09/2008 and 12/31/2019, which sums up to a total of 118, 256, 393, 395 messages throughout that period (languages are sorted by popularity).

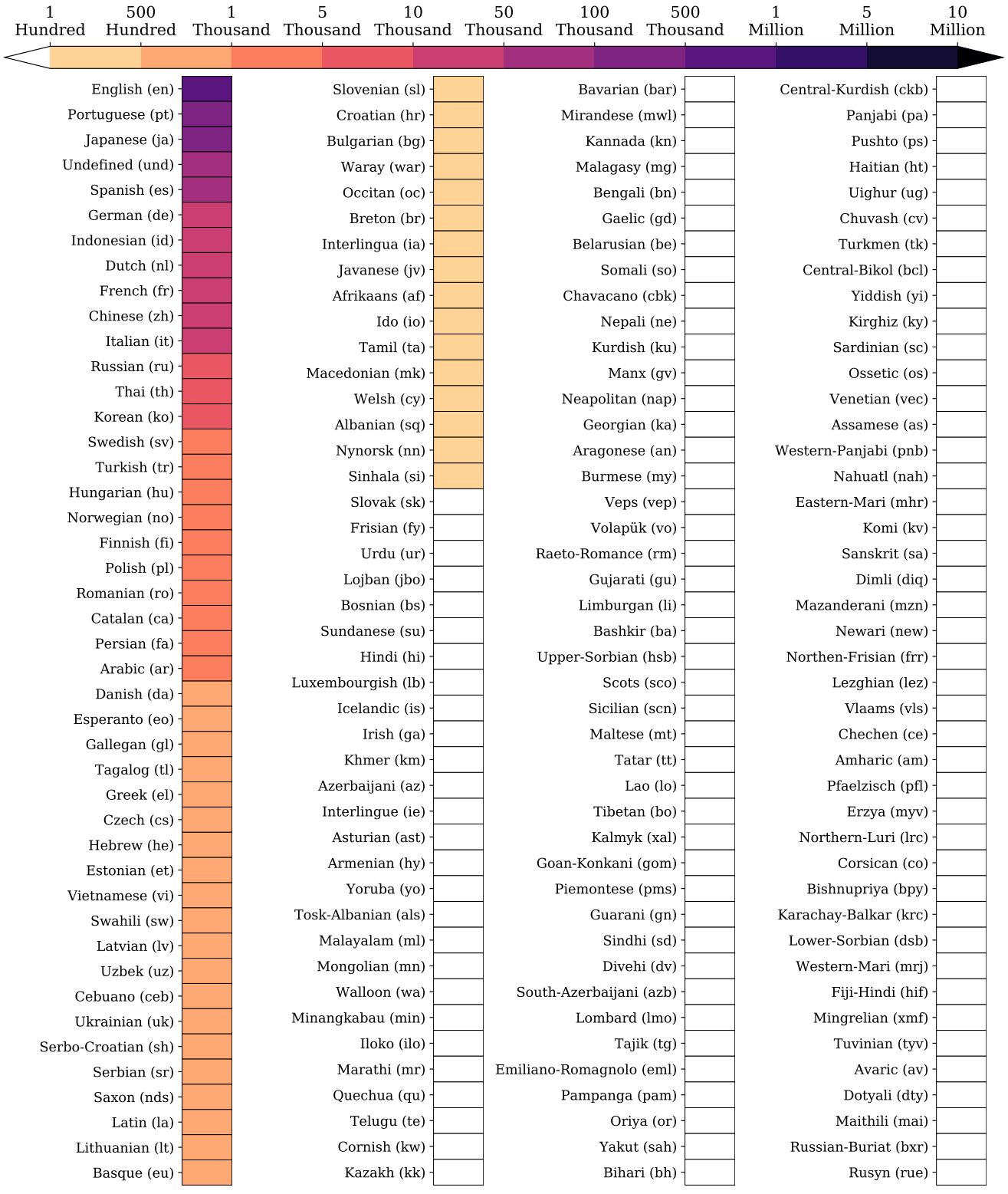

FIG. S10. **Dataset statistics (2009).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2009.

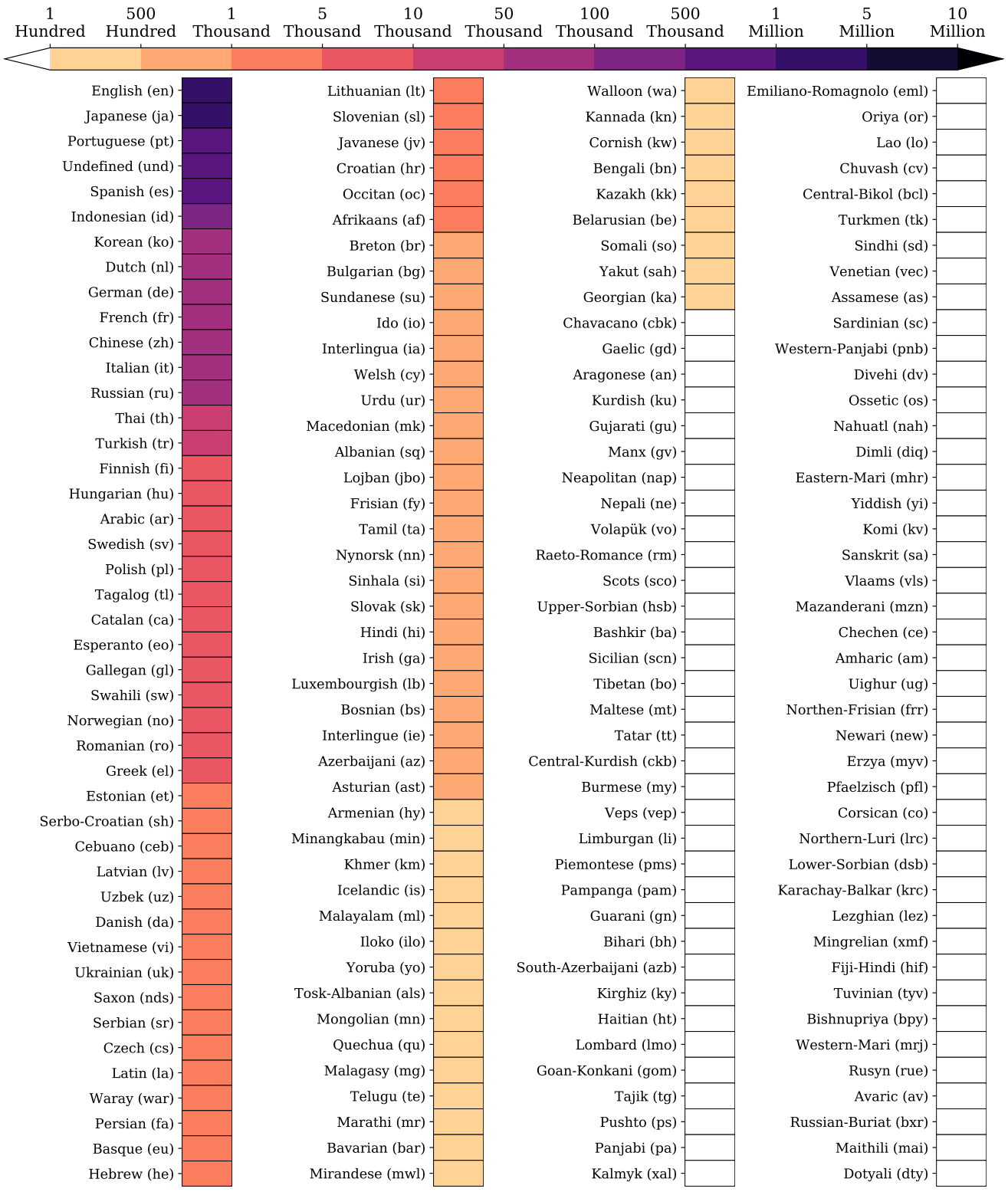

FIG. S11. **Dataset statistics (2010).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2010.

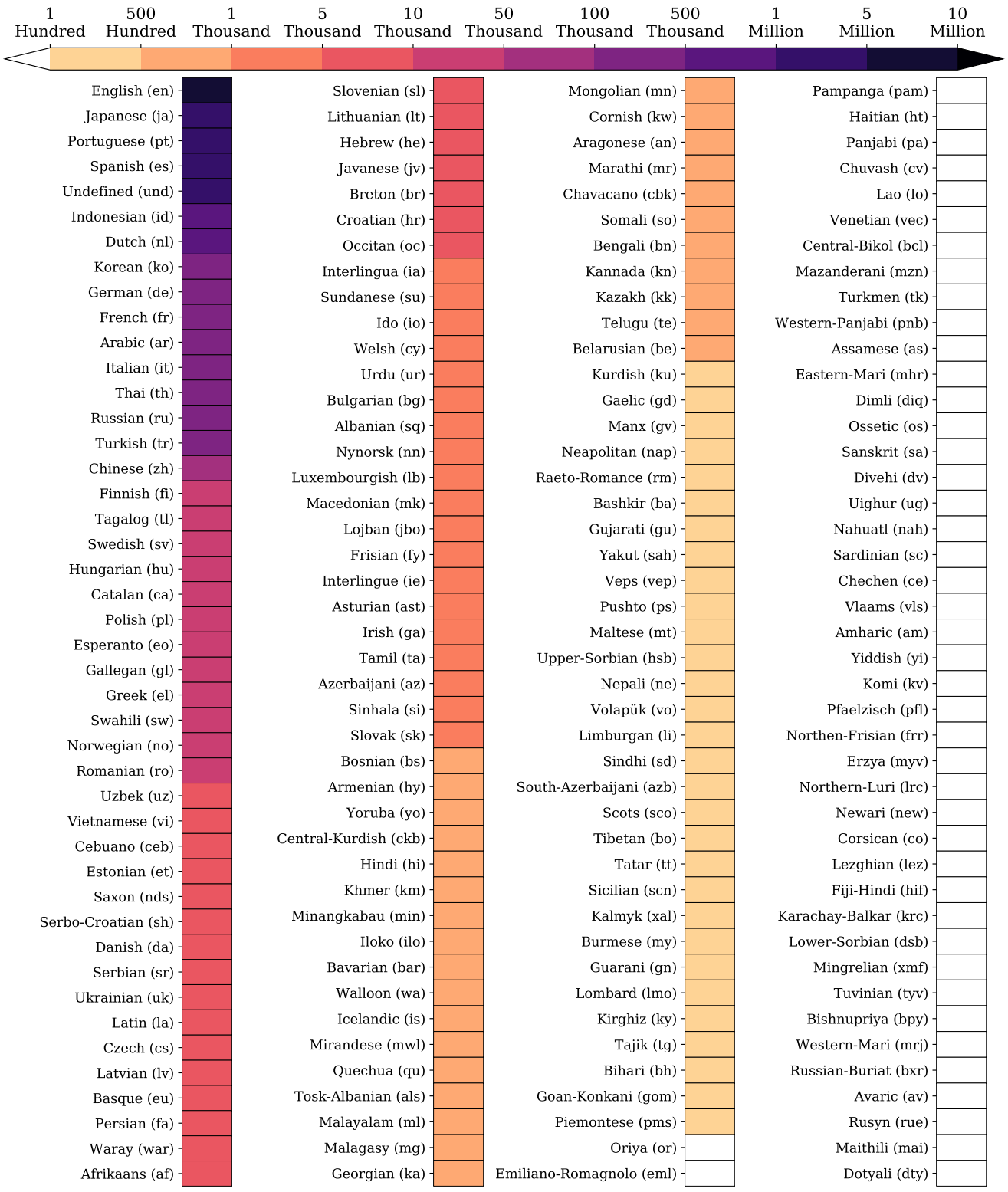

FIG. S12. **Dataset statistics (2011).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2011.

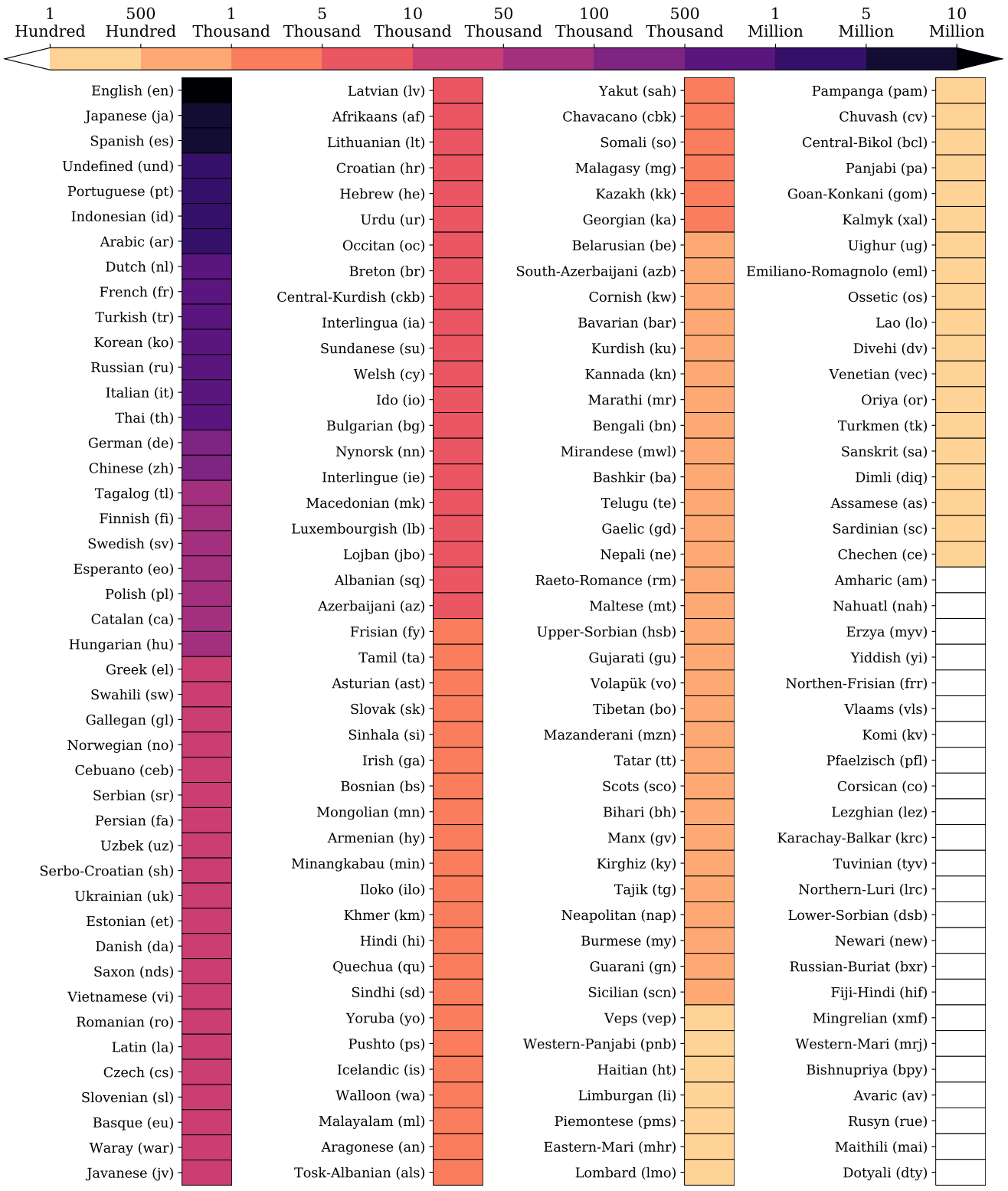

FIG. S13. **Dataset statistics (2012).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2012.

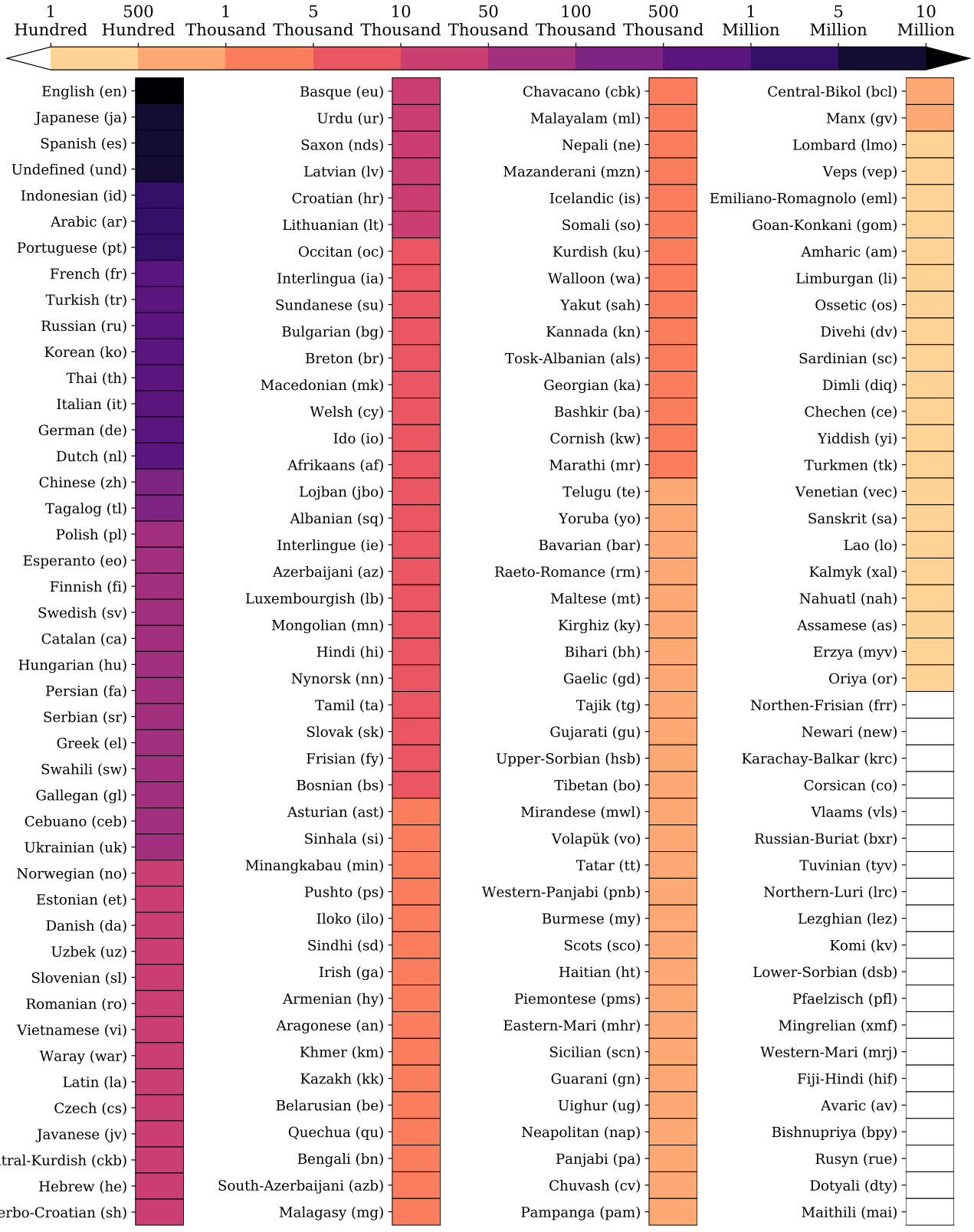

FIG. S14. **Dataset statistics (2013).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2013.

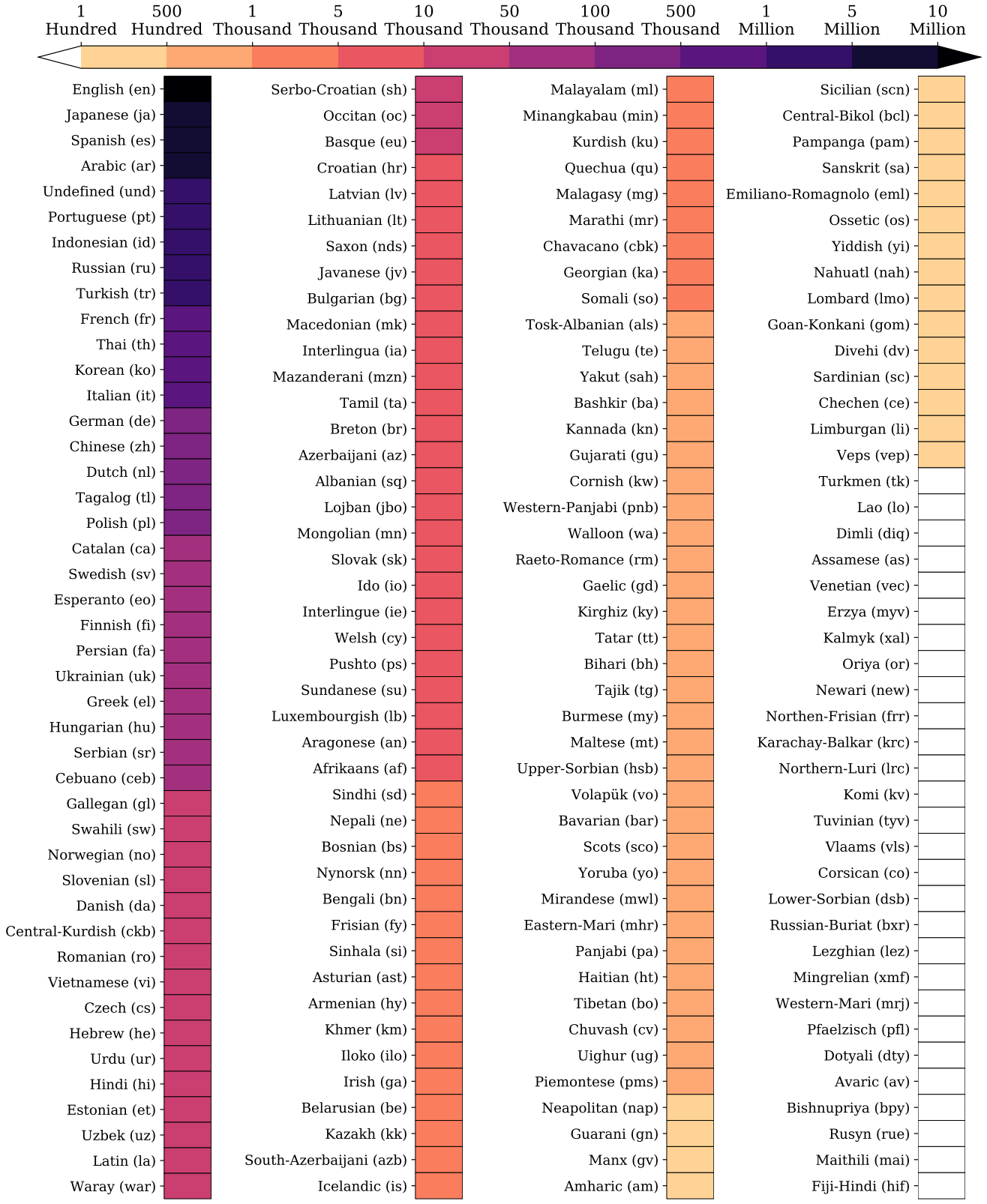

FIG. S15. **Dataset statistics (2014).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2014.

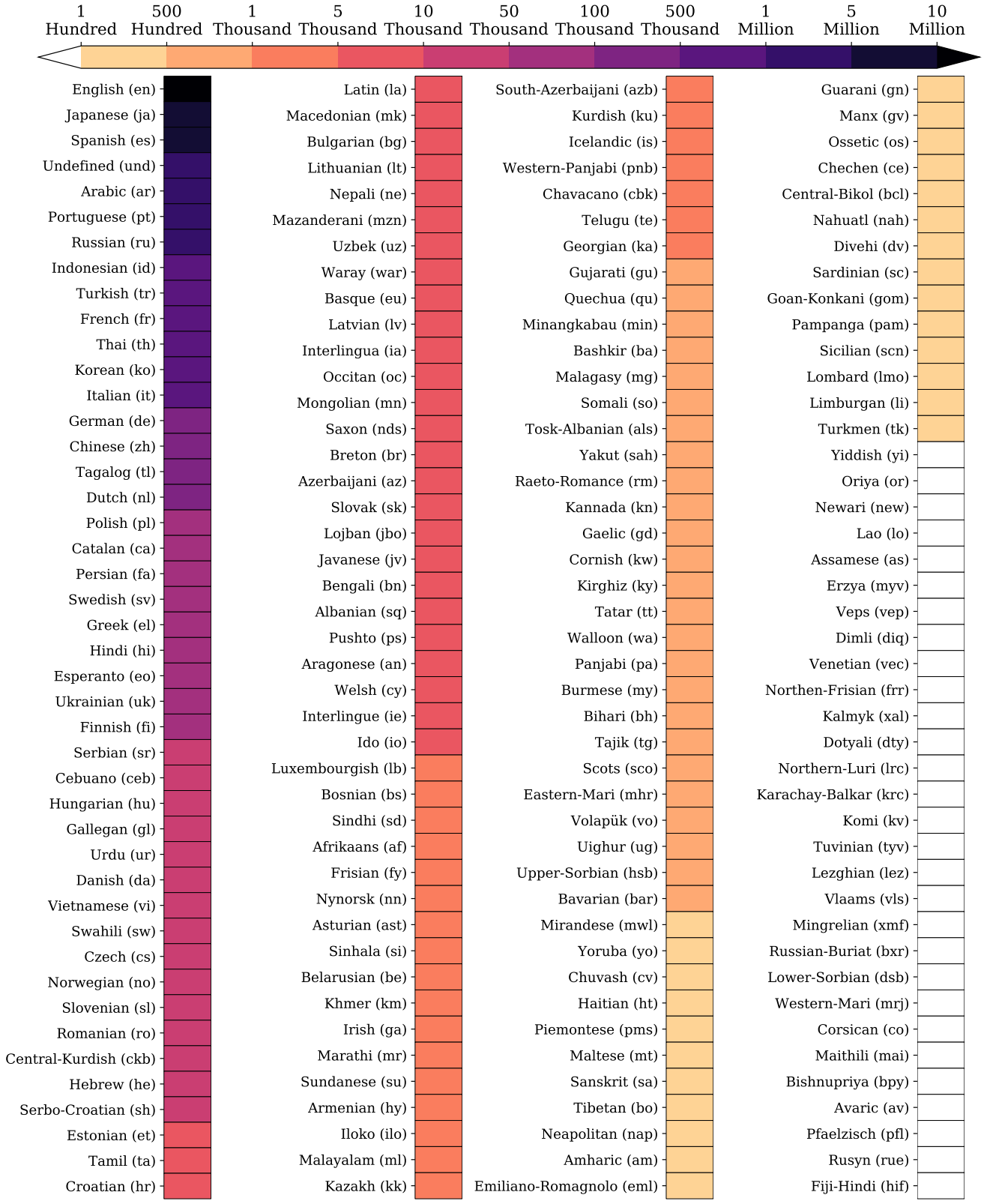

FIG. S16. **Dataset statistics (2015).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2015.

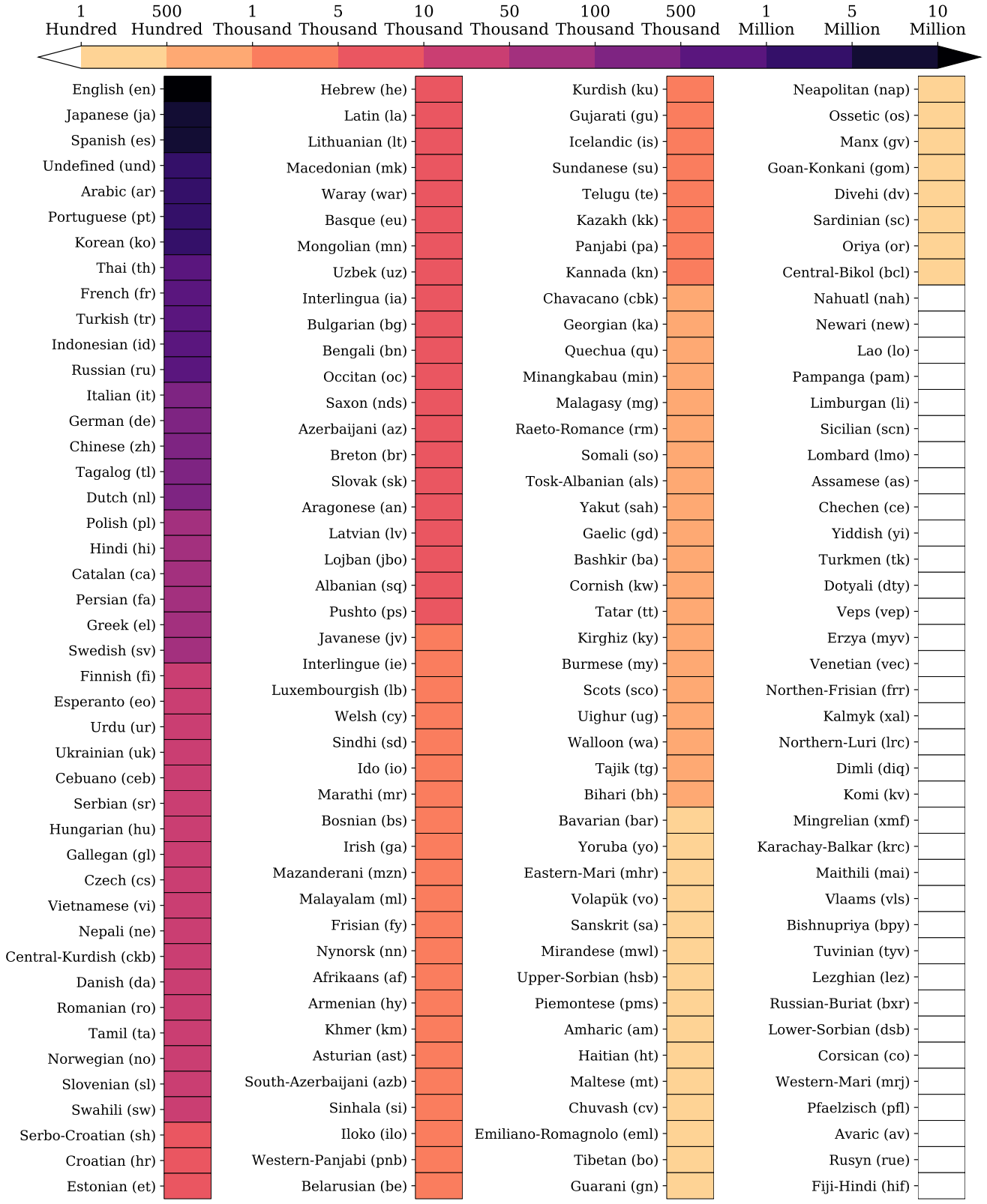

FIG. S17. **Dataset statistics (2016).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2016.

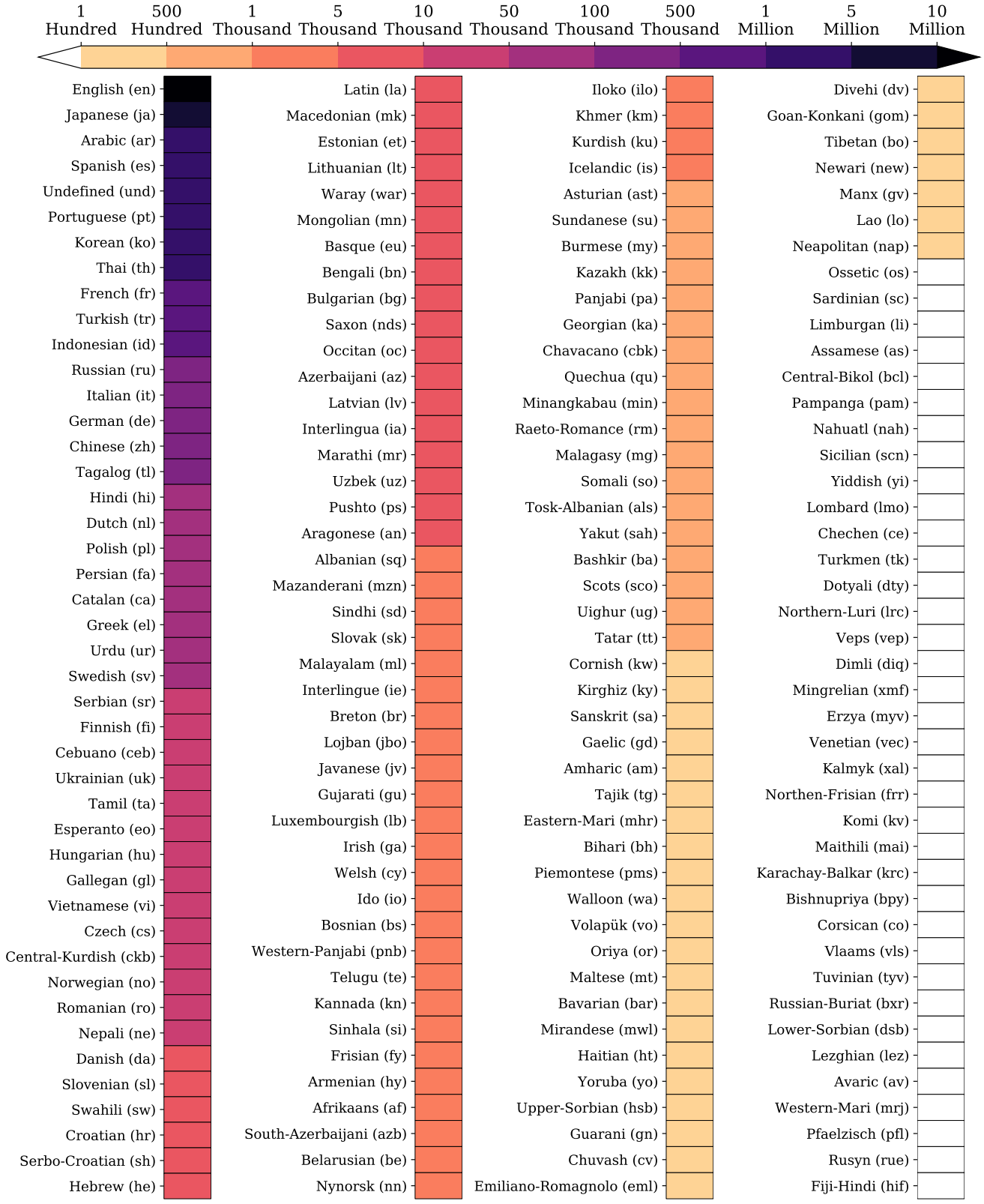

FIG. S18. **Dataset statistics (2017).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2017.

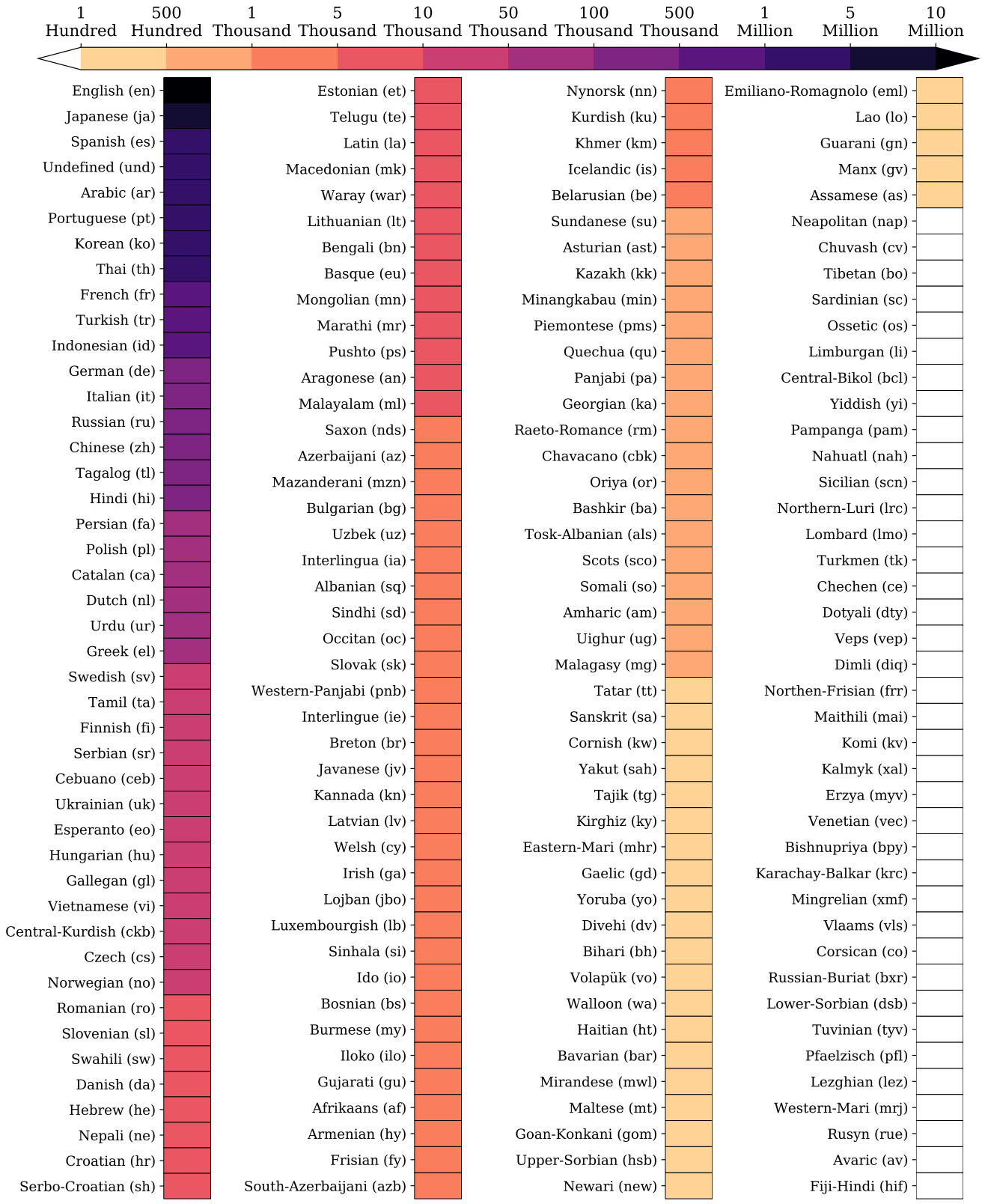

FIG. S19. **Dataset statistics (2018).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2018.

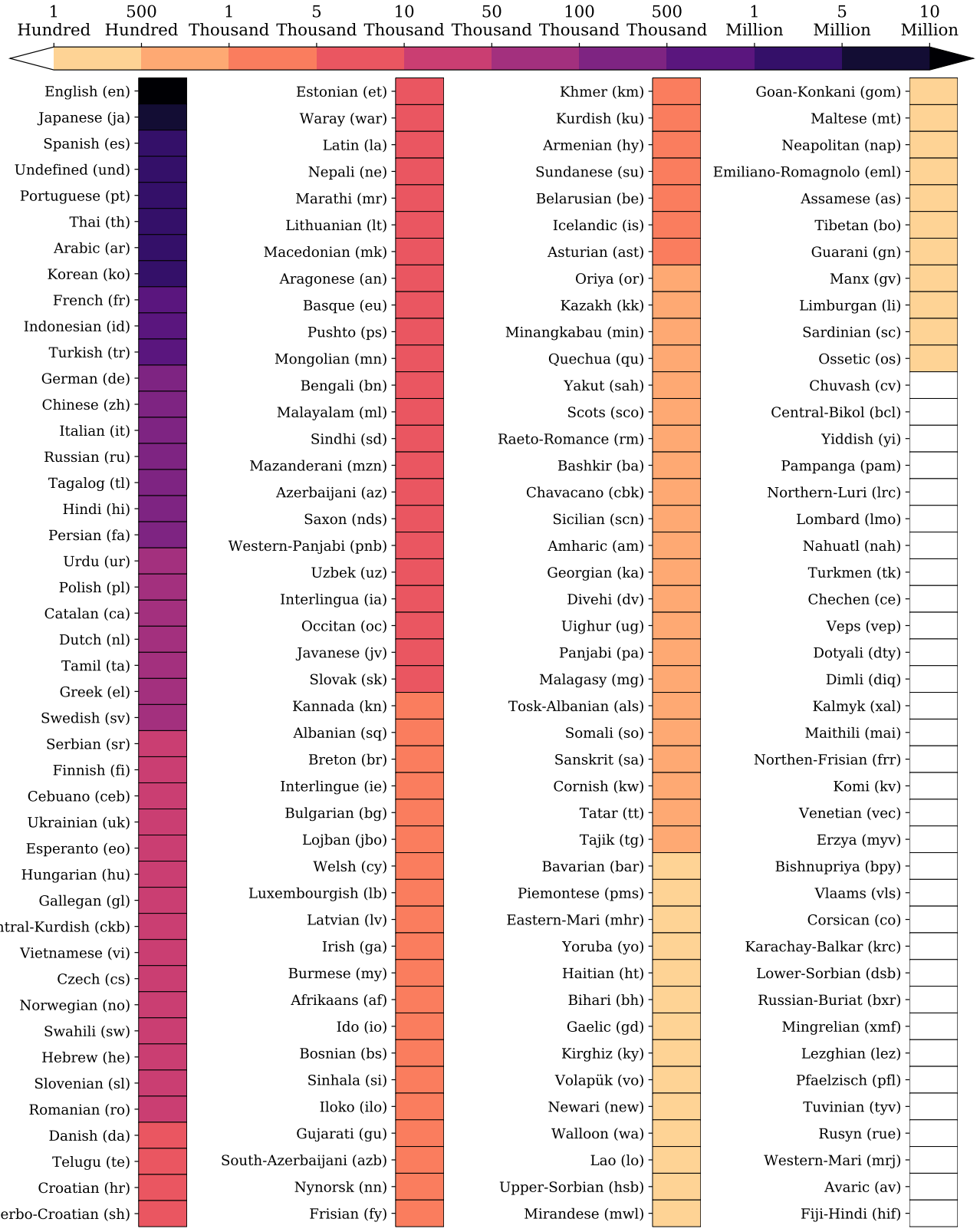

FIG. S20. **Dataset statistics (2018).** Average number of messages captured in our dataset as classified by `FastText` LID algorithm for 2019.