Toward A Multilingual and Multimodal Data Repository for COVID-19 Disinformation

Yichuan Li *
Computer Science Department
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
yli29@wpi.edu

Bohan Jiang
Computer Science and Engineering
Arizona State University
Tempe, Arizona, USA
bjiang14@asu.edu

Kai Shu
Department of Computer Science
Illinois Institute of Technology
Chicago, Illinois, United States
kshu@iit.edu

Huan Liu
Computer Science and Engineering
Arizona State University
Tempe, Arizona, USA
huanliu@asu.edu

***Abstract*—**

The COVID-19 epidemic is considered as the global health crisis of the whole society and the greatest challenge mankind faced since World War Two. Unfortunately, the fake news about COVID-19 is spreading as fast as the virus itself. The incorrect health measurements, anxiety, and hate speeches will have bad consequences on people's physical health, as well as their mental health in the whole world. To help better combat the COVID-19 fake news, we propose a new fake news detection dataset MM-COVID[1] **(Multilingual and Multidimensional COVID-19 Fake News Data Repository). This dataset provides the multilingual fake news and the relevant social context. We collect 3981 pieces of fake news content and 7192 trustworthy information from English, Spanish, Portuguese, Hindi, French and Italian, 6 different languages. We present a detailed and exploratory analysis of** MM-COVID **from different perspectives.**

I. Introduction

COVID-19, an infectious disease caused by a newly discovered coronavirus[2], has caused more than 40 million confirmed cases and 1.2 million deaths around the world in 2020 November[3]. Unfortunately, the fake news about Covid-19 has boosted the spreading of the disease and hate speech among people. For example, a couple who followed the half-backed health advice, took chloroquine phosphate to prevent COVID-19 and became ill within 20 minutes[4]; the racist linked the COVID-19 pandemic to Asian and people of Asian descent and the violence attacked Asian people have increased in the United States, United Kingdom, Italy, Greece, France, and Germany[5]. To stop the spreading of COVID-19 fake news, we should first address the problem of fake news detection.

However, identifying these COVID-19 related to fake news is non-trivial. There are several challenges: firstly, the COVID-19 fake news is multilingual. For example, FACTCHECK.org, a fact-checking agency, found that the fake news "COVID-19 is caused bacteria, easily treated with aspirin and coagulant." is firstly seen in Portuguese in Brazil then has the version of English in India and American[6]. The current available fake news datasets and methods are mainly focused on monolingual, omit the correlation between different languages. Thus it is necessary to have a multilingual fake news dataset to utilize rich debunked fake news language to help detect fake news in poor resource language. Second, fake news content merely provides a limited signal for spotting fake news. This is because the fake news is intentionally written to mislead readers and the difficulty in correlating multilingual fake news content. Thus, we need to explore auxiliary features except for fake news content such as social engagements and user profiles on social media. For example, people who post many vaccine conspiracy theories are more likely to transmit COVID-19 fake news. Thus, it is necessary to have a comprehensive dataset that has multilingual fake news content and their related social engagements to facilitate the COVID-19 fake news detection. However, to the best of our knowledge, existing COVID-19 fake news datasets did not cover both aspects.

Therefore, in this paper, we present a fake news dataset MM-COVID which contains fake news content, social engagements, and spatial-temporal information in 6 different languages. The main contribution of this work is to provide a multilingual and multidimensional fake news dataset MM-COVID to facilitate the fake news detection and conduct extensive exploration analysis on MM-COVID from a different perspective to demonstrate the quality of this dataset.

*Most of the work has been done at Arizona State University.

[1] The dataset is available at https://github.com/bigheiniu/X-COVID

[2] https://www.who.int/health-topics/coronavirus

[3] https://coronavirus.1point3acres.com/

[4] https://bit.ly/2IJNeWC

[5] https://bit.ly/3lG8Lhf

[6] https://www.factcheck.org/2020/05/covid-19-isn't-caused-by-bacteria/

TABLE I: Comparison with existing COVID-19 fake news datasets.

Features	**News Content**			**Social Context**				**Spatial-Temporal**	
Dataset	Multilingual	Linguistic	Visual	Tweet	Response	User	Network	Spatial	Temporal
Liar	-	√	-	-	-	-	-	-	-
FakeNewsNet	-	√	√	√	√	√	√	√	√
FakeCovid	√	√	-	-	-	-	-	√	√
ReCOVery	-	√	√	√	√	√	√	√	√
CoAID	-	√	√	√	√	-	-	-	√
CMU-MisCOV19	-	√	-	√	-	-	-	√	√
covid19-datasets	-	√	-	√	-	-	-	-	-
MM-COVID	√	√	√	√	√	√	√	√	√

This rest of this paper is organized as follows. We review the related work in Section II. The detail dataset construction and collection are presented in Section III. The exploring data analysis is illustrated in Section IV.

II. Background and Related Work

The COVID-19 fake news is a global threat now. Different languages of fake news is an explosion on social media. Most of them are intentionally written to mislead readers. To better combat the COVID-19 fake news, a multilingual and comprehensive dataset for developing fake news detection methods is necessary. Although there are many fake news datasets, most of them are either monolingual or only with linguistic features. To relieve the threat of fake news during the pandemic, we propose a dataset MM-COVID, which not only contains multilingual fake news, but also multi-dimensional features including news contents and social engagements. To be clarified, we list the detailed introduction of the related fake news dataset in the following.

- Liar [1]: There are 12.8k annotated short statements with various contexts PolitiFact. Each statement contains the statement content, speaker, context, label, and detailed justification from professional editors.
- FakeNewsNet [2]: This dataset includes the fact-checking article, source article which was debunked or supported in the fact-checking website, and related social engagements. This dataset is collected from PolitiFact[7] and GossipCop[8] with total 23,196 news pieces and 690,732 tweets.
- FakeCovid [3]: There are 5182 pieces of COVID-19 fact-checking news pieces in 40 languages from 105 countries. It get the labeled content from Snopes and Poynter[9].
- ReCOVery [4]: This dataset is used for news credibility classification. It collects the incredible news from the domain listed in NewsGaurd[10]. This dataset contains the news content and related social context in Twitter. There are 2, 029 news pieces and 140, 820 tweets in this dataset.
- CoAID [5]: This dataset contains the labeled news article, short claim, social post and the related social engagements. There are 1, 896 news pieces, 516 social platform posts and 183, 569 related user engagements.
- CMU-MisCOV19 [6]: This is a covid-19 related dataset with 4,573 annotated tweets in English. They classify the users into informed, misinformed and irrelevant groups.
- covid19-datasets [7]: The authors utilize the COVID-19 myth related keywords to collect the fake tweets.

From Table I, we can find that no existing fake news datasets can afford the multilingual fake news and comprehensive news content and social engagements. There are still some limitations to the existing datasets that we want to address in our proposed dataset. For example, FakeCovid labeled news pieces into fake and not fake which contains partly fake, half true, missing evidence, and so on. The news contents in FakeNewsNet contains noise since some of them are collected from Google Search result which often mentions similar but unrelated news pieces. ReCOVery labels each news piece as credible and incredibly based on the news source, rather than the human experts separately label each news pieces. CoAID mostly keeps the title of the fake news and much fake news misses the social engagements.

To address the aforementioned limitations of the existing datasets, we provide a new multilingual and multi-dimensional dataset MM-COVID which covers 6 languages and contains the information from the fake news content to the related social engagements.

III. Data Collection

In this section, we introduce the whole procedure of data collection, including fake news content and social context. The whole process is depicted in Figure 1.

A. News Content Collection

As shown in Figure 1, we need to firstly get the reliable labels from the fact-checking websites, and then retrieve the source content from these websites. We collect the veracity labels from Snopes[11] and Poynter[12] where the domain expert and journalists review the information and provide the fact-checking evaluation results as fake or real. Snopes is an independent publication owned by Snopes Media Group and mainly contains English content.

Poynter is an international fact-checking network (IFCN) alliance unifying 96 different fact-checking agencies like PolitiFact[13], FullFact[14] and etc, in 40 languages.

[7] www.politifact.com

[8] www.gossipcop.com

[9] www.poynter.org

[10] www.newsguardtech.com

[11] www.snopes.com

[12] www.poynter.org/coronavirusfactsalliance/

[13] www.politifact.com

[14] fullfact.org/

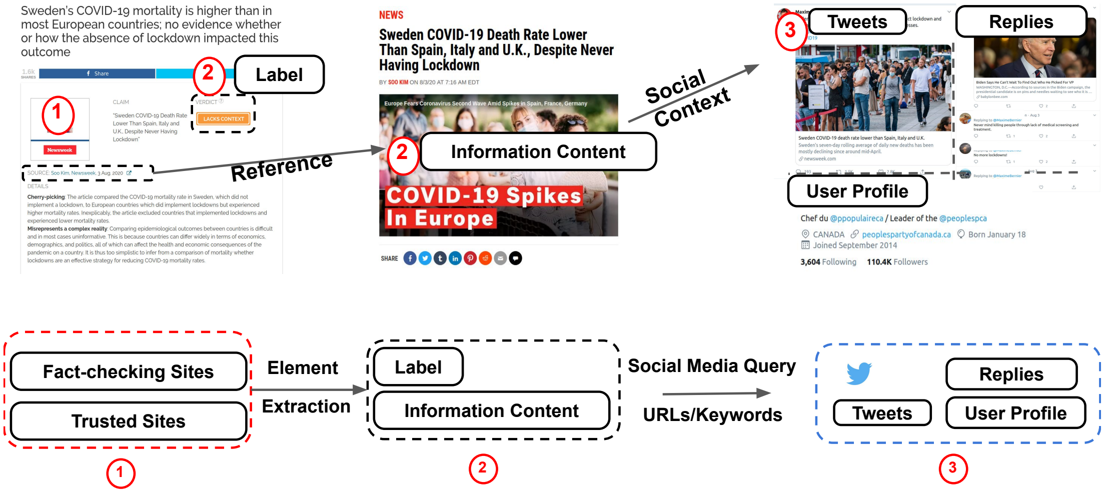

Sweden's COVID-19 mortality is higher than in most European countries; no evidence whether or how the absence of lockdown impacted this outcome
1
2
Label
Reference
NEWS
Sweden COVID-19 Death Rate Lower Than Spain, Italy and U.K., Despite Never Having Lockdown
2
Information Content
COVID-19 Spikes In Europe
Social Context
3
Tweets
Replies
User Profile
Chef du @ppopulaireca / Leader of the @peoplespca
CANADA peoplespartyofcanada.ca Born January 18
Joined September 2014
3,604 Following 110.4K Followers
Fact-checking Sites
Trusted Sites
1
Element Extraction
Label
Information Content
2
Social Media Query
URLs/Keywords
Replies
Tweets
User Profile
3

Fig. 1: The data collection process for MM-COVID with screenshot example.

TABLE II: Description of the features including in the dataset

Category	Features
Fact-checking Reviews	Fact-checking URL, Veracity Label, Debunked Explanations, Twitter Search Query
Source Contents	URL, Language, Location, Release Date, Text Content, Image
Social Engagements	Tweets, Replies, Retweets
Twitter Users	Profiles, Timelines, Location, Followers, Friends

To keep the quantity of each language, we only filter languages like English (en), Spanish (es), Portuguese (pt), Hindi (hi), French (fr), and Italian (it). Because the Poynter website only displays the translated English claims, we set the language for each claim based on the language used in the fact-checking article. After collecting the reliable label, we set heuristic crawling strategies for each fact-checking website to fetch the source content URL from the fact-checking websites. In some cases, the source content URL may be no longer available. To resolve the problem, we check the archived website[15] to see whether the page is archived or not. If not, we will consider the claim as the content of fake news.

Since most news pieces in Poynter and Snopes are fake news, to balance the dataset for each language, we choose several official health websites and collect the COVID-19 related news in these websites as additional real information. To filter unrelated information, we collect the news piece whose title contains any of the keywords *COVID-19*, *Coronavrius* and *SARS-CoV-2*. After we get the source URLs, we utilize the Newspaper3k[16] to crawl the content and its meta-information.

It should be noticed that the source of both fake news and real news include social media posts like Facebook, Twitter, Instagram, WhatsApp, etc, and news article posted in blogger and traditional news agencies.

B. User Social Engagement

As shown in Figure 1, we collect the user social engagements from the social platform based on the news content. Specifically, we form the search query based on the URL, the headline and the first sentence of the source content then use the Twitter advanced search API[17] through twarc[18] to collect the user social engagements. To reduce the search noise, we remove the special character, negative word, utilize the TFIDF [8] to extract the important words, and lastly check the query manually. The social engagements include the tweets which directly mention the news pieces, and the replies and retweets responding to these tweets. After we obtain the related tweets from the advanced search result, we collect the tweets' replies and retweets. Due to the fact that Twitter's API does not support getting replies, we approximately utilize the tweet's ID as the search query, which can only obtain the replies sent in the last week. In the end, we fetch all users' profiles, network connection, and the timeline of who engages in the news dissemination process.

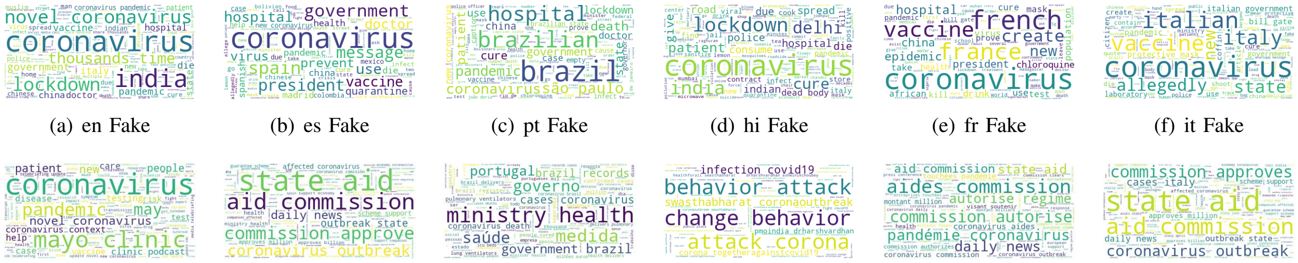
IV. Data Analysis

In this section, we will demonstrate the quality of the MM-COVID through statistical analysis and visualization. Because MM-COVID contains multi-dimensional information which can be used as features to identify the fake news, we separately make comparison among real news and fake news in source content, social context, and language temporal information. The detailed statistical information of our dataset is demonstrated in Table III.

[15] archive.is
[16] https://newspaper.readthedocs.io/en/latest/
[17] https://twitter.com/search-advanced?lang=en
[18] https://github.com/DocNow/twarc

TABLE III: Statistics of MM-COVID

Category	Fake						Real					
	en	es	pt	hi	fr	it	en	es	pt	hi	fr	it
# Source Content	2,168	808	371	336	189	109	2,114	2,405	713	1,023	392	937
# Tweets	32,811	21,911	15,738	1,143	2,821	750	26,565	1,553	268	1,205	166	369
# Replies	25,888	15,222	14,679	1,015	4,459	1,323	18,749	33,939	858	11,381	5,095	7,816
# Retweets	43,048	32,986	20,377	1,677	6,552	1,323	41,270	74,511	2,393	42,477	5,565	17,599
# Twitter Users	37,148	24,644	14,691	1,536	4,760	978	19,225	4,180	203	1,972	86	1,291



(a) en Fake (b) es Fake (c) pt Fake (d) hi Fake (e) fr Fake (f) it Fake

(g) en Real (h) es Real (i) pt Real (j) hi Real (k) fr Real (l) it Real

Fig. 2: Word Cloud of the fake news and real news in different languages. All the tokens are translated into English.

A. Source Content Analysis

Since the malicious users mostly manipulate the text content to mislead the audience, there stay text clues in the fake news content. We reveal these clues through the word cloud and the visualization of semantic representation and make a comparison among the fake news and real news.

In Figure 2, we visualize the most frequent words for each language. Non-English languages are translated into English for comparison. From Figure 2, we can find the fake news often mentions the medical-related words like *doctor*, *hospital* and *vaccine* across languages. This is because these places are the front line of defending Coronavirus, malicious users will transmit this fake news to spread fear and anxiety. The fake news also mentions the country name like *India*, *China*, *Spain*, *Brazil* and et al. While, the real news often mentions the keywords like *test* and *patient*. Besides, we also observe the topic similarity and difference among languages. For example, languages like “es”, “fr”, and “it”, they all talk about the welfare like *commission* and *aid* while other languages do not mentions these phrases. Although there is a topic difference between the fake news and real news, it is not consistent across languages and meanwhile, it cannot be directly applied to a single piece of text [9]. Thus it is necessary to learn a better representation of these contents and include auxiliary features into detection like the social context.

B. Measuring Social Context

Since the social media platform provides direct access to a large amount of information which may contain the COVID-19 related fake news, the propagation networks, transition paths, and the interacted user nodes in the path. They all can provide auxiliary and language invariant information for fake news detection. The monolingual social context integrated fake news models like dEFEND [10] and TCNN-URG [11] have achieved considerable performance improvement compared with only relying on the fake news content. Our dataset contains three different kinds of social context: user profiles, tweet posts, and social network structure. These can provide the opportunity to explore these findings across languages. In the following sections, we will explore the characteristics of these features and discuss the potential utilization of fake news detection.

User Profiles The existing research [12] has proven the correlation between user-profiles and fake news detection. For example, social bots play a disproportionate role in spreading fake news [13], [14]. In this part, we will illustrate the bot-probability of each user.

For each language, we randomly sample 500 users who only respond to the fake news and another 500 users related to real news for the bot detection. For a language that contains less than 500 users, like “pt”, “fr” in real news, we take all the users in these languages. We utilize the state-of-the-art bot detection method Botometer [15] to identify the probability of users being social bots. Botometer makes the prediction based on users’ public profile, timeline, and mentions. From the cumulative distributions listed in Figure 3, we can find that the users who engage in fake news are slightly more likely to be bots. In languages like “hi” and “fr”, the users who have extremely large bot-likelihood (> 0.6) are more likely to interact with the fake news. This observation is also consistent with past fake news research in [2], [16]. However, we also observe that bot-likelihood does not indicate the veracity of the news. For example, in “es” and “pt”, we have the opposite observation, and in “it”, there is no significant difference between the real news and fake news.

Tweet and Response In social media, people will express their emotions and focus on an event through tweets and their responses. These features can benefit the detection of fake news in general [17] [18]. We perform the sentiment analysis on the tweets. Since there is no sentiment classification method cover these 6 languages and emoji is the proxy of the sentiment in the tweets, we reveal the distribution of emojis for tweets

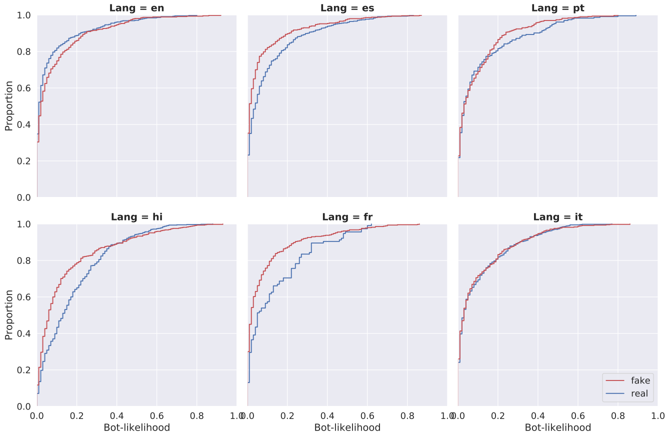
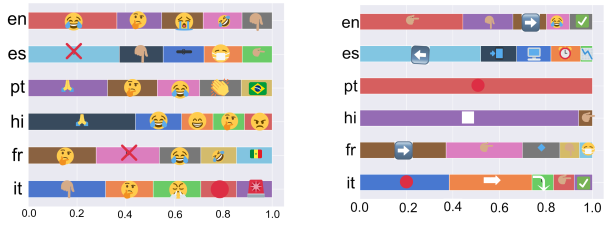


Fig. 3: The cumulative bot-likelihood distribution for users engaged with fake news and real news in all languages.

among different languages in Figure 4. Looking at the emoji of the reply tweets (Figure 4), we observe that there are more emotional emoji in the tweets, like laughing in “en”, “pt”, “hi” and “fr”, and angry in “hi” and “it”. However, in the real news, the direction and enumeration emoji dominate in all languages. These observations indicate that emoji or users’ emotions can benefit from fake news detection.



(a) fake news Tweets

(b) Real News Tweets

Fig. 4: Emoji distribution for tweets in different languages.

To gain insights into user response intensify between the fake news and real news, we reveal the distribution of the count of replies towards them. From Figure 5, we can find that for languages except “en” real news get larger number of replies than the fake news. But in “en”, there is no significant difference between the real news and the fake news. These observations indicate that language also impacts users’ social interactions.

C. Temporal Information

Recent researches have shown that the temporal information of social engagements can improve fake news detection performance [19], [20]. To reveal the temporal patterns difference between real news and fake news, we follow the analysis approaches in [2], [16] that select two news pieces for each language and reveal the count. From Figure 6, we observe that (i) real news in “en”, “es”, “pt”, “hi”, and “fr” have a sudden increase in social engagements. (ii) in the language, on the contrary, there is a steady increase in the real news. These common temporal social engagement patterns allow us to extract the language invariant features for fake news detection.

V. Conclusion and Future Work

To combat the global infodemic, we release a multilingual fake news dataset MM-COVID, which contains the news content, social context, and spatiotemporal information in English, Spanish, Portuguese, Hindi, French, and Italian six different languages. Through our exploratory analysis, we identify several language invariant features for fake news detection. This dataset can facilitate further research in fake news detection, and fake news mitigation. We plan to evaluate existing fake news detection methods on MM-COVID and provide detailed discussion about the potential application of MM-COVID in fake news detection and fake news mitigation. There are several improvements for this dataset in the future work: *(i)* include more languages in the dataset, such as Chinese, Russian, Germany, and Japanese. *(ii)* collect social context from different social platforms like Reddit, Facebook, YouTube, and Instagram, and so on.

Acknowledgments

This work is, in part, supported by NSF (#2029044) and Project A-COGER.

References

[1] W. Y. Wang, “”liar, liar pants on fire”: A new benchmark dataset for fake news detection,” 2017.

[2] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, “Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media,” 2018.

[3] G. K. Shahi and D. Nandini, “Fakecovid – a multilingual cross-domain fact check news dataset for covid-19,” 2020.

[4] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, “Recovery: A multimodal repository for covid-19 news credibility research,” 2020.

[5] L. Cui and D. Lee, “Coaid: Covid-19 healthcare misinformation dataset,” 2020.

[6] S. A. Memon and K. M. Carley, “Characterizing covid-19 misinformation communities using a novel twitter dataset,” *arXiv preprint arXiv:2008.00791*, 2020.

[7] I. Inuwa-Dutse and I. Korkontzelos, “A curated collection of covid-19 online datasets,” *arXiv preprint arXiv:2007.09703*, 2020.

[8] A. Rajaraman and J. D. Ullman, *Data Mining*. Cambridge University Press, 2011, p. 1–17.

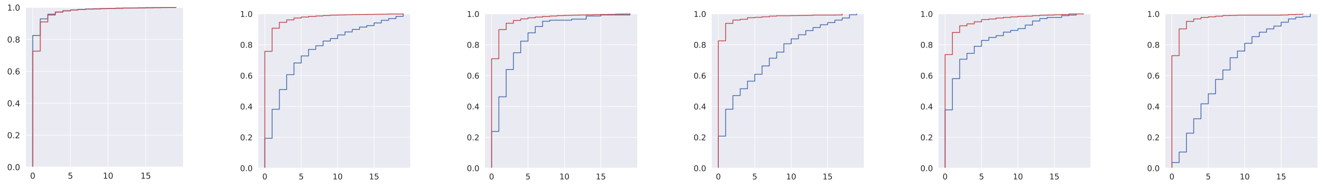
[9] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” 2017.

[10] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, “Defend: Explainable fake news detection,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 395–405. [Online]. Available: https://doi.org/10.1145/3292500.3330935

[11] F. Qian, C. Gong, K. Sharma, and Y. Liu, “Neural user response generator: Fake news detection with collective user intelligence,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, ser. IJCAI’18. AAAI Press, 2018, p. 3834–3840.

[12] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, “The role of user profile for fake news detection,” 2019.

[13] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, “The spread of low-credibility content by social bots,” *Nature Communications*, vol. 9, no. 1, p. 4787, Nov 2018. [Online]. Available: https://doi.org/10.1038/s41467-018-06930-7



(a) **en** (b) **es** (c) **pt** (d) **hi** (e) **fr** (f) **it**

Fig. 5: The accumulated distribution of number of replies for each language. Blue stands for real news related tweets and Red is for fake news related tweets.

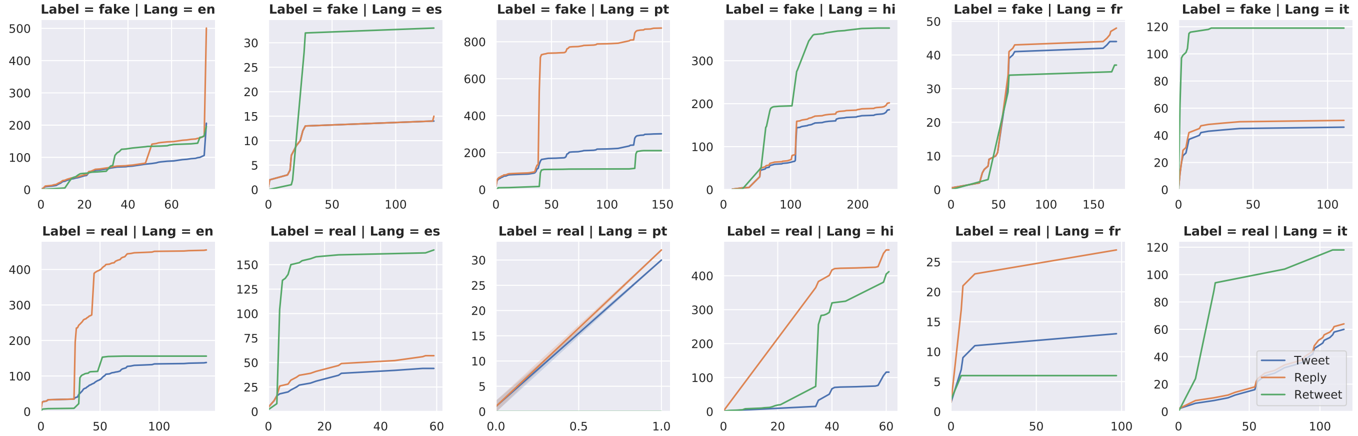


Fig. 6: Temporal patterns of social engagement of fake news and real news in different languages.

[14] Z. K. Stine, T. Khaund, and N. Agarwal, "Measuring the information-foraging behaviors of social bots through word usage," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018, pp. 570–571.

[15] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot," *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 2016. [Online]. Available: http://dx.doi.org/10.1145/2872518.2889302

[16] E. Dai, Y. Sun, and S. Wang, "Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository," 2020.

[17] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, p. 2972–2978.

[18] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 1589–1599. [Online]. Available: https://www.aclweb.org/anthology/D11-1147

[19] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, p. 3818–3824.

[20] K. Shu, D. Mahudeswaran, and H. Liu, "Fakenewstracker: a tool for fake news collection, detection, and visualization," *Computational and Mathematical Organization Theory*, vol. 25, 10 2018.