Big Data
Volume 8, Number 3, 2020

DOI: 10.1089/big.2020.0062

ORIGINAL ARTICLE

FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media

Kai Shu,[1,*] Deepak Mahudeswaran,[1] Suhang Wang,[2] Dongwon Lee,[2] and Huan Liu[1]

Abstract

Social media has become a popular means for people to consume and share the news. At the same time, however, it has also enabled the wide dissemination of *fake news*, that is, news with intentionally false information, causing significant negative effects on society. To mitigate this problem, the research of fake news detection has recently received a lot of attention. Despite several existing computational solutions on the detection of fake news, the lack of comprehensive and community-driven fake news data sets has become one of major roadblocks. Not only existing data sets are scarce, they do not contain a myriad of features often required in the study such as *news content*, *social context*, and *spatiotemporal information*. Therefore, in this article, to facilitate fake news-related research, we present a fake news data repository *FakeNewsNet*, which contains two comprehensive data sets with diverse features in *news content*, *social context*, and *spatiotemporal information*. We present a comprehensive description of the FakeNewsNet, demonstrate an exploratory analysis of two data sets from different perspectives, and discuss the benefits of the FakeNewsNet for potential applications on fake news study on social media.

Keywords: fake news; disinformation; misinformation; data repository

Introduction

Social media has become a primary source of news consumption nowadays. Social media is cost-free, easy to access, and can fast disseminate posts. Hence, it acts as an excellent way for individuals to post and/or consume information. For example, the time individuals spend on social media is continually increasing.* As another example, studies from Pew Research Center shows that ∼68% of Americans get some of their news on social media in 2018† and this has shown a constant increase since 2016. Since there is no regulatory authority on social media, the quality of news pieces spread in social media is often lower than traditional news sources. In other words, social media also enables the widespread of fake news. Fake news[1] means the false information that is spread deliberately to deceive people. Fake news affects the individuals as well as society as a whole. First, fake news can disturb the authenticity balance of the news ecosystem. Second, fake news persuades consumers to accept false or biased stories. For example, some individuals and organizations spread fake news in social media for financial and political gains.[2,3] It is also reported that fake news has an influence on the 2016 U.S. presidential elections.‡ Finally, fake news may cause significant effects on real-world events. For example, "Pizzagate,"

*https://www.socialmediatoday.com/marketing/how-much-time-do-people-spend-social-media-infographic
†http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018
‡https://www.independent.co.uk/life-style/gadgets-and-tech/news/tumblr-russian-hacking-us-presidential-election-fake-news-internet-research-agency-propaganda-bots-a8274321.html

[1]*Department of Computer Science and Engineering, Arizona State University, Tempe, Arizona, USA.*
[2]*College of Information Sciences and Technology, Penn State University, University Park, Pennsylvania, USA.*

*Address correspondence to: *Kai Shu, Department of Computer Science and Engineering, Arizona State University, Brickyard Suite 561BB (CIDSE), 699 South Mill Avenue, Tempe, AZ 85281, USA*, E-mail: kai.shu@asu.edu

a piece of fake news from Reddit, leads to a real shooting.* Thus, fake news detection is a critical issue that needs to be addressed.

Detecting fake news on social media presents unique challenges. First, fake news pieces are intentionally written to mislead consumers, which makes it not satisfactory to spot fake news from news content itself. Thus, we need to explore information in addition to news content, such as user engagements and social behaviors of users on social media. For example, a credible user's comment that "This is fake news" is a strong signal that the news may be fake. Second, the research community lacks data sets that contain spatiotemporal information to understand how fake news propagates over time in different regions, how users react to fake news, and how we can extract useful temporal patterns for (early) fake news detection and intervention. Thus, it is necessary to have comprehensive data sets that have news content, social context, and spatiotemporal information to facilitate fake news research. However, to the best of our knowledge, existing data sets only cover one or two aspects.

Therefore, in this article, we construct and publicize a multidimensional data repository *FakeNewsNet*,† which currently contains two data sets with news content, social context, and spatiotemporal information. The data set is constructed using an end-to-end system, FakeNewsTracker.[4,‡] The constructed FakeNewsNet repository has the potential to boost the study of various open research problems related to fake news study. First, the rich set of features in the data sets provides an opportunity to experiment with different approaches for fake news detection, understand the diffusion of fake news in social network, and intervene in it. Second, the temporal information enables the study of early fake news detection by generating synthetic user engagements from historical temporal user engagement patterns in the data set.[5] Third, we can investigate the fake news diffusion process by identifying provenances, persuaders, and developing better fake news intervention strategies.[6] Our data repository can serve as a starting point for many exploratory studies for fake news, and provide a better shared insight into disinformation tactics. We aim to continuously update this data repository, expand it with new sources and features, as well as maintain completeness. The main contributions of the article are as follows:

- We construct and publicize a multidimensional data repository for various facilitating fake news detection-related researches such as fake news detection, evolution, and mitigation.
- We conduct an exploratory analysis of the data sets from different perspectives to demonstrate the quality of the data sets, understand their characteristics, and provide baselines for future fake news detection.
- We discuss benefits and provide insight for potential fake news studies on social media with FakeNewsNet.

Background and Related Work

Fake news detection in social media aims to extract useful features and build effective models from existing social media data sets for detecting fake news in the future. Thus, a comprehensive and large-scale data set with multidimensional information in online fake news ecosystem is important. The multidimensional information not only provides more signals for detecting fake news but can also be used for researches such as understanding fake news propagation and fake news intervention. Although there exist several data sets for fake news detection, the majority of them only contains linguistic features. Few of them contain both linguistic and social context features. To facilitate research on fake news, we provide a data repository that includes not only news contents and social contents but also spatiotemporal information. For a better comparison of the differences, we list existing popular fake news detection data sets hereunder and compare them with the FakeNewsNet repository in Table 1.

BuzzFeedNews

This data set comprises a complete sample of news published in Facebook from nine news agencies over a week close to the 2016 U.S. election from September 19–23 and September 26 and 27.§ Every post and the linked article were fact-checked claim-by-claim by five BuzzFeed journalists. It contains 1627 articles—826 mainstream, 356 left-wing, and 545 right-wing articles.

LIAR

This data set[7] is collected from fact-checking website PolitiFact.** It has 12.8K human labeled short

*https://www.rollingstone.com/politics/politics-news/anatomy-of-a-fake-news-scandal-125877/

†https://github.com/KaiDMML/FakeNewsNet

‡http://blogtrackers.fulton.asu.edu:3000/#/about

§https://github.com/BuzzFeedNews/2016-10-facebook-fact-check/tree/master/data

**https://www.cs.ucsb.edu/william/software.html

Table 1. Comparison with existing fake news detection data sets

Features / Data set	News content		Social context				Spatiotemporal information	
	Linguistic	*Visual*	*User*	*Post*	*Response*	*Network*	*Spatial*	*Temporal*
BuzzFeedNews	✓							
LIAR	✓							
BS detector	✓							
CREDBANK	✓		✓	✓			✓	✓
BuzzFace	✓			✓	✓			✓
FacebookHoax	✓		✓	✓	✓			
FakeNewsNet	✓	✓	✓	✓	✓	✓	✓	✓

statements collected from PolitiFact and the statements are labeled into six categories ranging from completely false to completely true as pants on fire, false, barely true, half-true, mostly true, and true.

BS detector

This data set is collected from a browser extension called BS detector developed for checking news veracity.* It searches all links on a given web page for references to unreliable sources by checking against a manually compiled list of domains. The labels are the outputs of the BS detector, rather than human annotators.

CREDBANK

This is a large-scale crowd-sourced data set[8] of $\sim$60 million tweets that cover 96 days starting from October 2015.† The tweets are related to >1000 news events. Each event is assessed for credibility by 30 annotators from Amazon Mechanical Turk.

BuzzFace

This data set[9] is collected by extending the BuzzFeed data set with comments related to news articles on Facebook.‡ The data set contains 2263 news articles and 1.6 million comments.

FacebookHoax

This data set[10] comprises information related to posts from the Facebook pages related to scientific news (nonhoax) and conspiracy pages (hoax) collected using Facebook Graph API.§ The data set contains 15,500 posts from 32 pages (14 conspiracy and 18 scientific) with >2,300,000 likes.

We provide a comparison in Table 1 to show that no existing public data set can provide all possible features of news content, social context, and spatiotemporal information. Existing data sets have some limitations that we try to address in FakeNewsNet. For example, BuzzFeedNews only contains headlines and text for each news piece and covers news articles from very few news agencies. LIAR data set contains mostly short statements instead of entire news articles with meta attributes. BS detector data are collected and annotated by using a developed news veracity checking tool, rather than using human expert annotators. CREDBANK data set was originally collected for evaluating tweet credibility and the tweets in the data set are not related to the fake news articles and hence cannot be effectively used for fake news detection. BuzzFace data set has basic news contents and social context information, but it does not capture the temporal information. The FacebookHoax data set consists very few instances about conspiracy theories and scientific news.

To address the disadvantages of existing fake news detection data sets, the proposed FakeNewsNet repository collects multidimensional information from news content, social context, and spatiotemporal information from different types of news domains such as political and entertainment sources.

Data Set Integration

In this section, we introduce a process that integrates data sets to construct the FakeNewsNet repository. We demonstrate (Fig. 1) how we can collect news contents with reliable ground truth labels, how we obtain additional social context and spatiotemporal information.

News content

To collect reliable ground truth labels for fake news, we utilize fact-checking websites to obtain news contents for fake news and true news such as *PolitiFact*** and *GossipCop*.†† In PolitiFact, journalists and domain

*https://github.com/bs-detector/bs-detector
†http://compsocial.github.io/CREDBANK-data/
‡https://github.com/gsantia/BuzzFace
§https://github.com/gabll/some-like-it-hoax

**https://www.politifact.com/
††https://www.gossipcop.com/

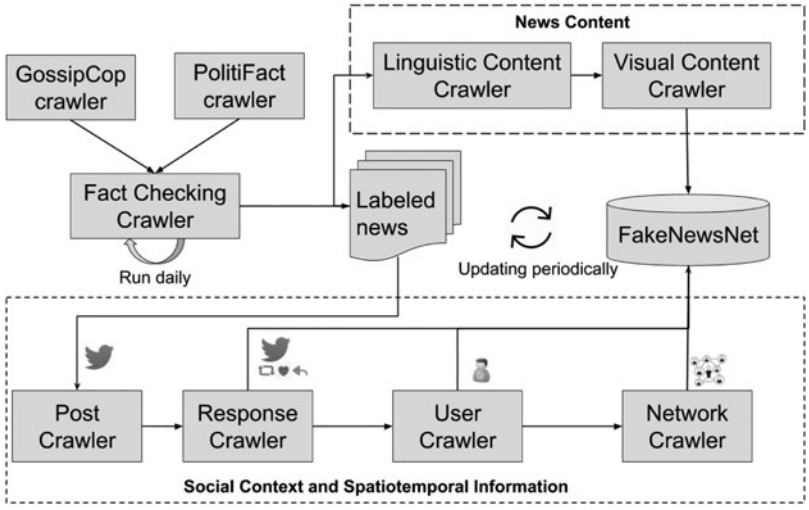


FIG. 1. The flowchart of data set integration process for FakeNewsNet. It mainly describes the collection of news content, social context, and spatiotemporal information.

experts review the political news and provide fact-checking evaluation results to claim news articles as fake* or real.† We utilize these claims as ground truths for fake and real news pieces. In PolitiFact's fact-checking evaluation result, the source URLs of the web page that published the news articles are provided, which can be used to fetch the news contents related to the news articles. In some cases, the web pages of source news articles are removed and are no longer available. To tackle this problem, we (1) check if the removed page was archived and automatically retrieve content at the Wayback Machine;‡ and (2) make use of Google web search in automated manner to identify news article that is most related to the actual news. GossipCop is a website for fact-checking entertainment stories aggregated from various media outlets. GossipCop provides rating scores on the scale of 0–10 to classify a news story as the degree from fake to real. From our observation, almost 90% of the stories from GossipCop have scores <5, which is mainly because the purpose of GossipCop is to showcase more fake stories. To collect true entertainment news pieces, we crawl the news articles from E! Online,§ which is a well-known trusted media website for publishing entertainment news pieces. We consider all the articles from E! Online as real news sources. We collect all the news stories from GossipCop with rating scores <5 as the fake news stories.

Since GossipCop does not explicitly provide the URL of the source news article, so similarly we search the news headline in Google or archive to obtain the news source information. The headlines of GossipCop news articles are generally written to reflect the fact and so may not be used directly. For example, one of the headlines, "Jennifer Aniston NOT Wearing Brad Pitts Engagement Ring, Despite Report" mentions the fact instead of the original news articles title. We utilize some heuristics to extract proper headlines such as (1) using the text in quoted string and (2) removing negative sentiment words. For example, some headlines include quoted strings that are exact text from the original news source. In this case, we extract the named entities from the headline using CoreNLP tool[11] and quoted strings to form the search query. For example, in headline Jennifer Aniston, Brad Pitt NOT "Just Married" Despite Report, we extract named entities, including Jennifer Aniston, Brad Pitt, and quoted strings, including Just Married, and form the search query as "Jennifer Aniston Brad Pitt Just Married" because the quoted text in addition with named entities mostly provides the context of the original news. As another example, the headlines are written in the negative sense to correct the false information, for example, "Jennifer Aniston NOT Wearing Brad Pitts Engagement Ring, Despite Report." So we remove negative sentiment words retrieved from

*https://www.politifact.com/subjects/fake-news/
†https://www.politifact.com/truth-o-meter/rulings/true/
‡https://archive.org/web/
§https://www.eonline.com/

SentiWordNet[12] and some hand-picked words from the headline to form the search query, for example, "Jennifer Aniston Wearing Brad Pitts Engagement Ring."

Social context

The user engagements related to the fake and real news pieces from fact-checking websites are collected using search application programming interfaces (API) provided by social media platforms such as the Twitter's Advanced Search API.* The search queries for collecting user engagements are formed from the headlines of news articles, with special characters removed from the search query to filter out the noise. We search for tweets using queries containing all the words in the headline to ensure the relevance of the resultant tweets. In addition, the URLs mentioned in the tweets collected are further used as search queries to collect additional tweets, so that we try to reduce the bias of data collection only using keywords. After we obtain the social media posts that directly spread news pieces, we further fetch the user *response* toward these posts such as replies, likes, and reposts. In addition, when we obtain all the users engaging in news dissemination process, we collect all the metadata for user profiles, user posts, and the social network information.

Spatiotemporal information

The spatiotemporal information includes spatial and temporal information. For spatial information, we obtain the locations explicitly provided in user profiles. The temporal information indicates that we record the timestamps of user engagements, which can be used to study how fake news pieces propagate on social media, and how the topics of fake news are changing over time. Since fact-checking websites periodically update newly coming news articles, so we dynamically collect these newly added news pieces and update the FakeNewsNet repository as well. In addition, we keep collecting the user engagements for all the news pieces periodically in the FakeNewsNet repository such as the recent social media posts, and second-order user behaviors such as replies, likes, and retweets. For example, we run the news content crawler and update Tweet collector per day. The spatiotemporal information provides useful and comprehensive information for studying fake news problem from a temporal perspective.

*https://twitter.com/search-advanced?lang=en

Data Analysis

FakeNewsNet has multidimensional information related to news content, social context, and spatiotemporal information. In this section, we first provide some preliminary quantitative analysis to illustrate the features of FakeNewsNet. We then perform fake news detection using several state-of-the-art models to evaluate the quality of the FakeNewsNet repository. The detailed statistics of FakeNewsNet repository is illustrated in Table 2.

Assessing news content

Since fake news attempts to spread false claims in news content, the most straightforward means of detecting it is to find clues in a news article. First, we analyze the topic distribution of fake and real news articles. From Figure 2a and b, we can observe that the fake and real news of the PolitiFact data set is mostly related to the political campaign. In case of GossipCop data set from Figure 2c and d, we observe that the fake and real news are mostly related to gossip about the relationship among celebrities. In addition, we can see the topics for fake news and real news are slightly different in general. However, for specific news, it is difficult to only use topics in the content to detect fake news,[1] which necessitates the need to utilize other auxiliary information such as social context.

We also explore the distribution of publishers who publish fake news on both data sets. We find out that there are in total 301 publishers publishing 432 fake news pieces, among which 191 of all publishers only publish 1 piece of fake news, and 40 publishers publish at least 2 pieces of fake news such as theglobalheadlines .net and worldnewsdailyreport.com. For Gossipcop, there are in total 209 publishers publishing 6048 fake news pieces, among which 114 of all publishers only publish 1 piece of fake news, and 95 publishers publish at least 2 pieces of fake news such as hollywoodlife.com and celebrityinsider.org. The reason may be that these fact-checking websites try to identify those check-worthy breaking news events regardless of the publishers, and fake news publishers can be shut down after they were reported to publish fake news pieces.

Comparing social contexts of fake and real news

Social context represents the news proliferation process over time, which provides useful auxiliary information to infer the veracity of news articles. Generally, there are three major aspects of the social media context that we want to represent: user profiles, user posts,

Table 2. Statistics of the FakeNewsNet repository

		PolitiFact		*GossipCop*	
Category	*Features*	*Fake*	*Real*	*Fake*	*Real*
News content					
Linguistic	# News articles	432	624	5323	16,817
	# News articles with text	420	528	4947	16,694
Visual	# News articles with images	336	447	1650	16,767
Social context					
User	# Users posting tweets	95,553	249,887	265,155	80,137
	# Users involved in likes	113,473	401,363	348,852	145,078
	# Users involved in retweets	106,195	346,459	239,483	118,894
	# Users involved in replies	40,585	18,6675	106,325	50,799
Post	# Tweets posting news	164,892	399,237	519,581	876,967
Response	# Tweets with replies	11,975	41,852	39,717	11,912
	# Tweets with likes	31,692	93,839	96,906	41,889
	# Tweets with retweets	23,489	67,035	56,552	24,955
Network	# Followers	405,509,460	1,012,218,640	630,231,413	293,001,487
	# Followees	449,463,557	1,071,492,603	619,207,586	308,428,225
	Average # followers	1299.98	982.67	1020.99	933.64
	Average # followees	1440.89	1040.21	1003.14	982.80
SpaTemp. Information					
Spatial	# User profiles with locations	217,379	719,331	429,547	220,264
	# Tweets with locations	3337	12,692	12,286	2451
Temporal	# Timestamps for news	296	167	3558	9119
	# Timestamps for response	171,301	669,641	381,600	200,531

and network structures. Next, we perform an exploratory study of these aspects on FakeNewsNet and introduce the potential usage of these features to help fake news detection.

User profiles. User profiles on social media have been shown to be correlated with fake news detection.[13] Research has also shown that fake news pieces are likely to be created and spread by nonhuman accounts, such as social bots or cyborgs.[1,14] We will illustrate some user profile features in FakeNewsNet repository.

First, we explore whether the creation time of user accounts for fake news and true news is different or not. We compute time ranges of account register time with the current date and the results are shown in Figure 3. We can see that the account creation time distribution of users posting fake news is significantly different from those who post real news, with the p-value <0.05 under t-test. Also, we notice that it is not necessary that users with an account created long time or shorter time post fake/real news more often. For example, the mean creation time for users posting fake news (2214.09) is less than that for real news (2166.84) in Politifact, whereas we see opposite case in Gossipcop data set.

Next, we take a deeper look into the user profiles and assess the social bots effects. We randomly selected 10,000 users who posted fake and real news and performed bot detection using Botometer,[15] one of the state-of-the-art bot detection algorithm. Botometer* takes Twitter username as input and utilizes various features extracted from metadata and outputs a probability score in [0,1], indicating how likely the user is a bot. We set the threshold of 0.5 on the bot score returned from the Botometer results to determine bot accounts. Figure 4 shows the ratio of the bot and human users involved in tweets related to fake and real news. We can see that bots are more likely to post tweets related to fake news than real users. For example, almost 22% of users involved in fake news are bots, whereas only ~9% of users are predicted as bot users for real news. Similar results were observed with different thresholds on bot scores based on both data sets. This indicates that there are bots in Twitter for spreading fake news, which is consistent with the observation in Shao et al.[14] In addition, most users who spread fake news (~78%) are still more likely to be humans than bots (~22%), which is also in consistence with the findings in Vosoughi et al.[16]

Post and response. People express their emotions or opinions toward fake news through social media posts, such as skeptical opinions and sensational

*https://botometer.iuni.iu.edu/

FIG. 2. **(a–d)** The word cloud of new body text for fake and real news on PolitiFact and GossipCop.

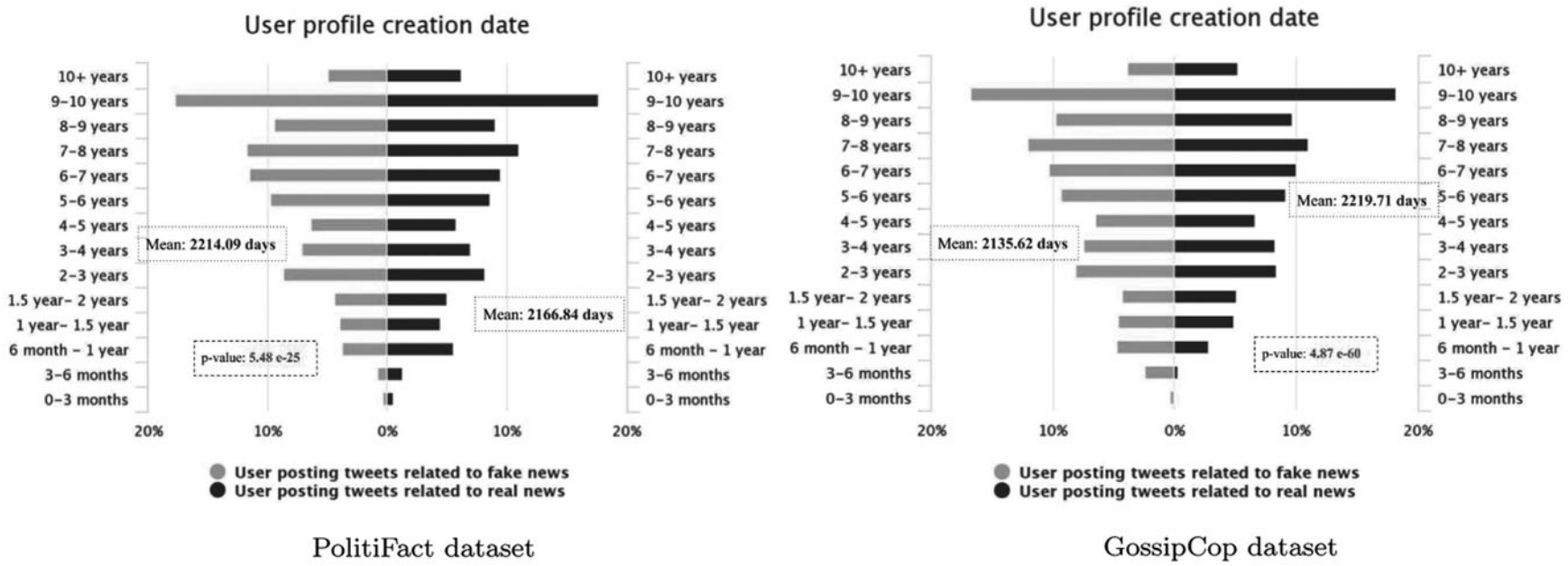


FIG. 3. The distribution of user profile creation dates on PolitiFact and GossipCop data sets.

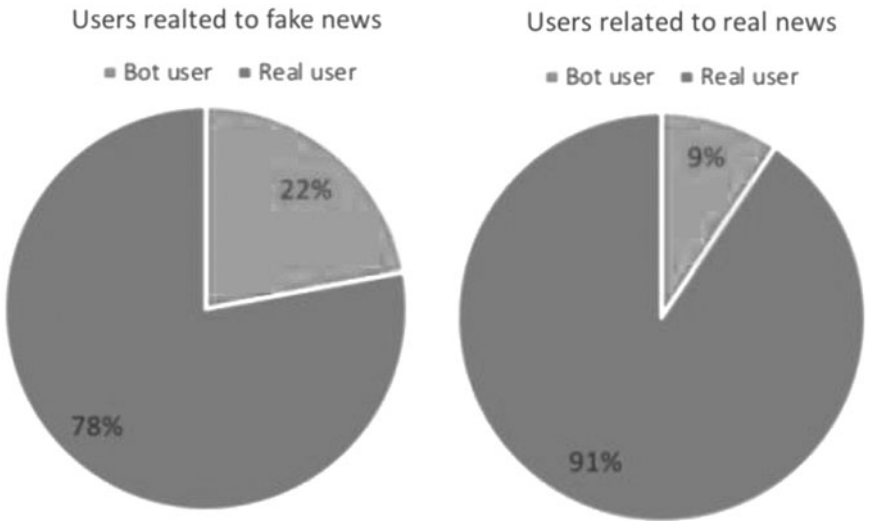


FIG. 4. A comparison of bot scores on users related to fake and real news on PolitiFact data set.

reactions. These features are important signals to study fake news and disinformation in general.[17,18]

We perform sentiment analysis on the replies of user posts that spread fake news and real news using one of the state-of-the-art unsupervised sentiment prediction tool called VADER.[19,*] Figure 5 shows the relationship between positive, neutral, and negative replies for all news articles. For each news piece, we obtain all the replies for this news piece and predict the sentiment as positive, negative, or neutral. Then we calculate the ratio of positive, negative, and neutral replies for the news. For example, if a news piece has the sentiment distribution of replies as [0.5, 0.5, 0.5], it occurs in the middle of the very center of the triangle in Figure 5a. We can also see that the real news have more number of neutral replies over positive and negative replies, whereas fake articles have a bigger ratio of negative sentiments. In case of sentiment of the replies of the Gossipcop data set shown in Figure 5b, we cannot observe any significant differences between fake and real news. This could be because of the difficulty in identifying fake and real news related to entertainment by common people.

We analyze the distribution of likes, retweets, and replies of tweets, which can help gain insights on user interaction related to fake and real news. Social science studies have theorized the relationship between user behaviors and their perceived beliefs on the information on social media.[20] For example, the behaviors of likes and retweets are more emotional, whereas replies are more rational.

We plot the ternary triangles that illustrate the ratio of replies, retweets, and likes from the second-order engagements toward the posts that spread fake news or real news pieces. From Figure 6, we observe that the (1) fake news pieces tend to have fewer replies and more retweets; (2) real news pieces have more ratio of likes than fake news pieces, which may indicate that users are more likely to agree on real news. The differences in the distribution of user behaviors between fake news and real news have potentials to study users' beliefs characteristics. FakeNewsNet provides real-world data sets to understand the social factors of user engagements and underlying social science as well.

Networks. Users tend to form different networks on social media in terms of interests, topics, and relations, which serve as the fundamental paths for information diffusion.[1] Fake news dissemination processes tend to form an echo chamber cycle, highlighting the value of extracting network-based features to represent these types of network patterns for fake news detection.[21]

We look at the social network statistics of all the users who spread fake news or real news. The social network features such as followers count and followee count can be used to estimate the scope of how the fake news can spread in social media. We plot the distribution of follower count and followee count of users

*https://github.com/cjhutto/vaderSentiment

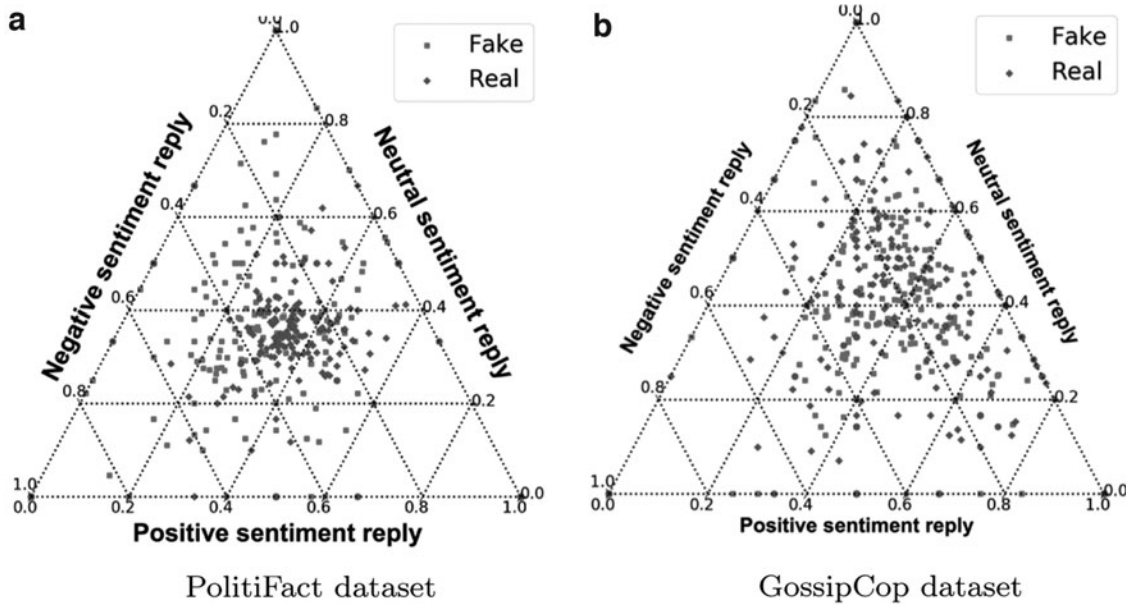


FIG. 5. **(a, b)** Ternary plots of the ratio of the positive, neutral, and negative sentiment replies for fake and real news.

in Figure 7. We can see that (1) the follower and followee count of the users generally follows power law distribution, which is commonly observed in social network structures; (2) there is a spike in the followee count distribution of both users and this is because of the restriction imposed by Twitter* on users to have at most 5000 followees when the number of following is <5000.

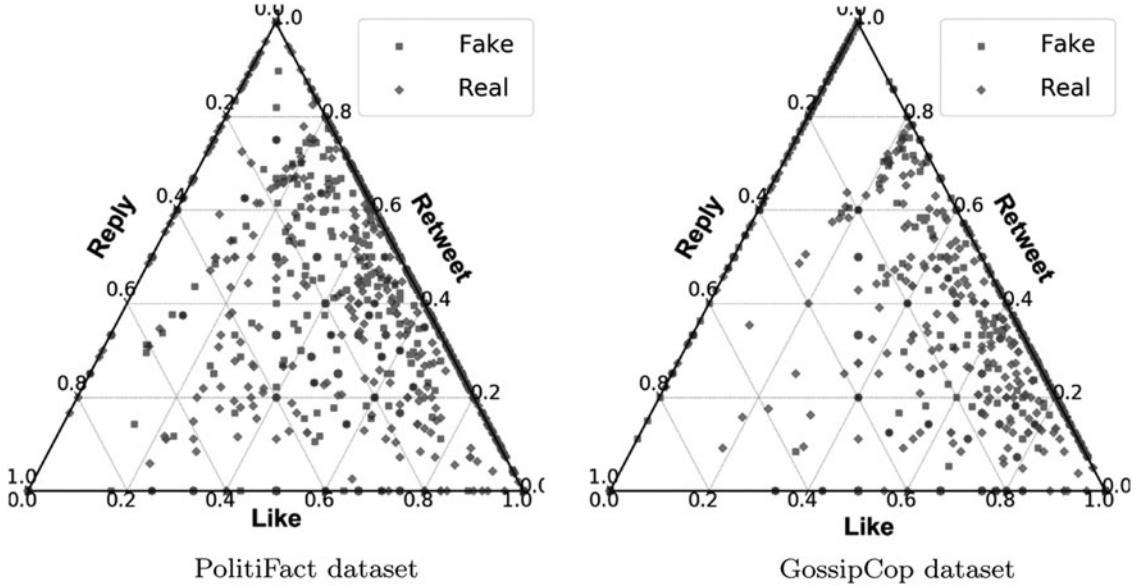
Characterizing spatiotemporal information

Recent research has shown users' temporal responses can be modeled using deep neural networks to help detection fake news,[22] and deep generative models can generate synthetic user engagements to help early fake news detection.[23] The spatiotemporal information in FakeNewsNet depicts the temporal user engagements for news articles, which provides the necessary information to further study the utility of using spatiotemporal information to detect fake news.

First, we investigate if the temporal user engagements such as posts, replies, and retweets are different for fake news and real news with similar topics, for example, fake news "*TRUMP APPROVAL RATING Better than Obama and Reagan at Same Point in their Presidencies*" from June 9 to 13, 2018 and real news "*President Trump in Moon Township Pennsylvania*" from March 10 to 20, 2018. As shown in Figure 8, we can observe that (1) for fake news, there is a sudden increase in the number of retweets and it does remain constant beyond a short time, whereas in the case of real news there is a steady increase in the number of retweets; (2) fake news pieces tend to receive fewer replies than real news. We have similar observations in Table 2, and replies count for 5.76% among all tweets for fake news, and 7.93% for real news. The differences of diffusion patterns for temporal user engagements have the potential to determine the threshold time for early fake news detection. For example, if we can predict the sudden increase of user engagements, we should use the user engagements before the time point and detect fake news accurately to limit the affect size of fake news spreading.[6]

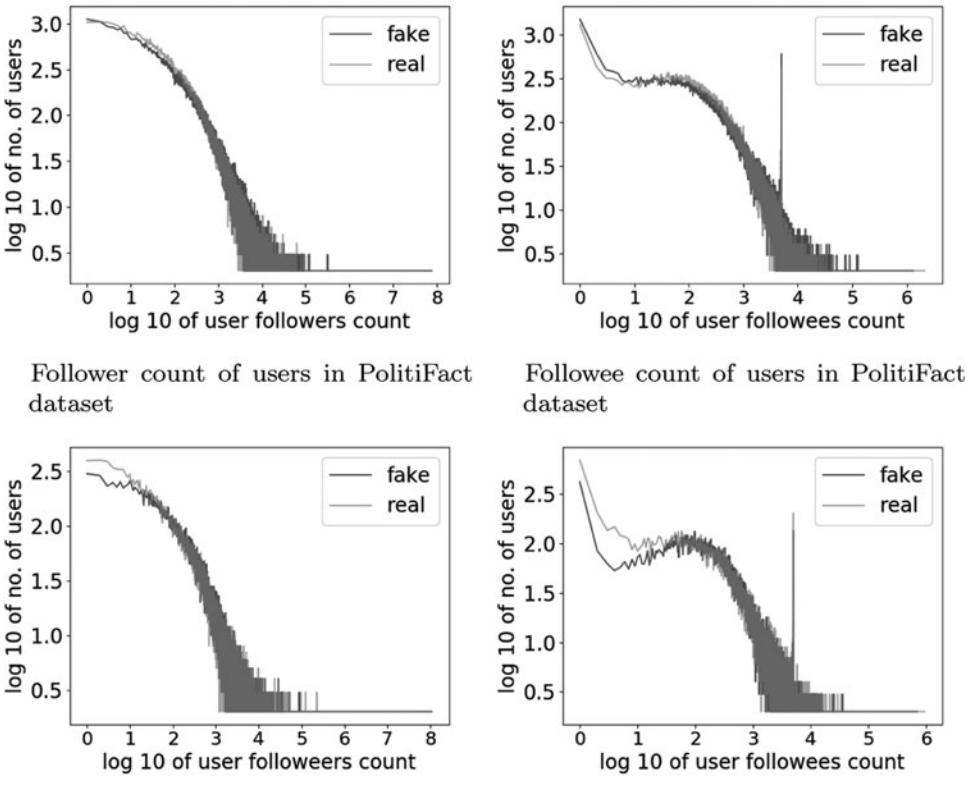
Next, we demonstrate the geolocation distribution of users engaging in fake and real news (see Fig. 9 for Politifact data set). We show the locations explicitly provided by users in their profiles, and we can see that users in the PolitiFact data set who post fake news have a different distribution than those posting real news. Since it is usually sparse of locations provided by users explicitly, we can further consider the location information attached with Tweets, and even

*https://help.twitter.com/en/using-twitter/twitter-follow-limit



PolitiFact dataset GossipCop dataset

FIG. 6. Ternary plots of the ratio of likes, retweet, and reply of tweets related to fake and real news.



Follower count of users in PolitiFact dataset

Followee count of users in PolitiFact dataset

Follower count of users in Gossipcop dataset

Followee count of users in PolitiFact dataset

FIG. 7. The distribution of the count of followers and followees related to fake and real news.

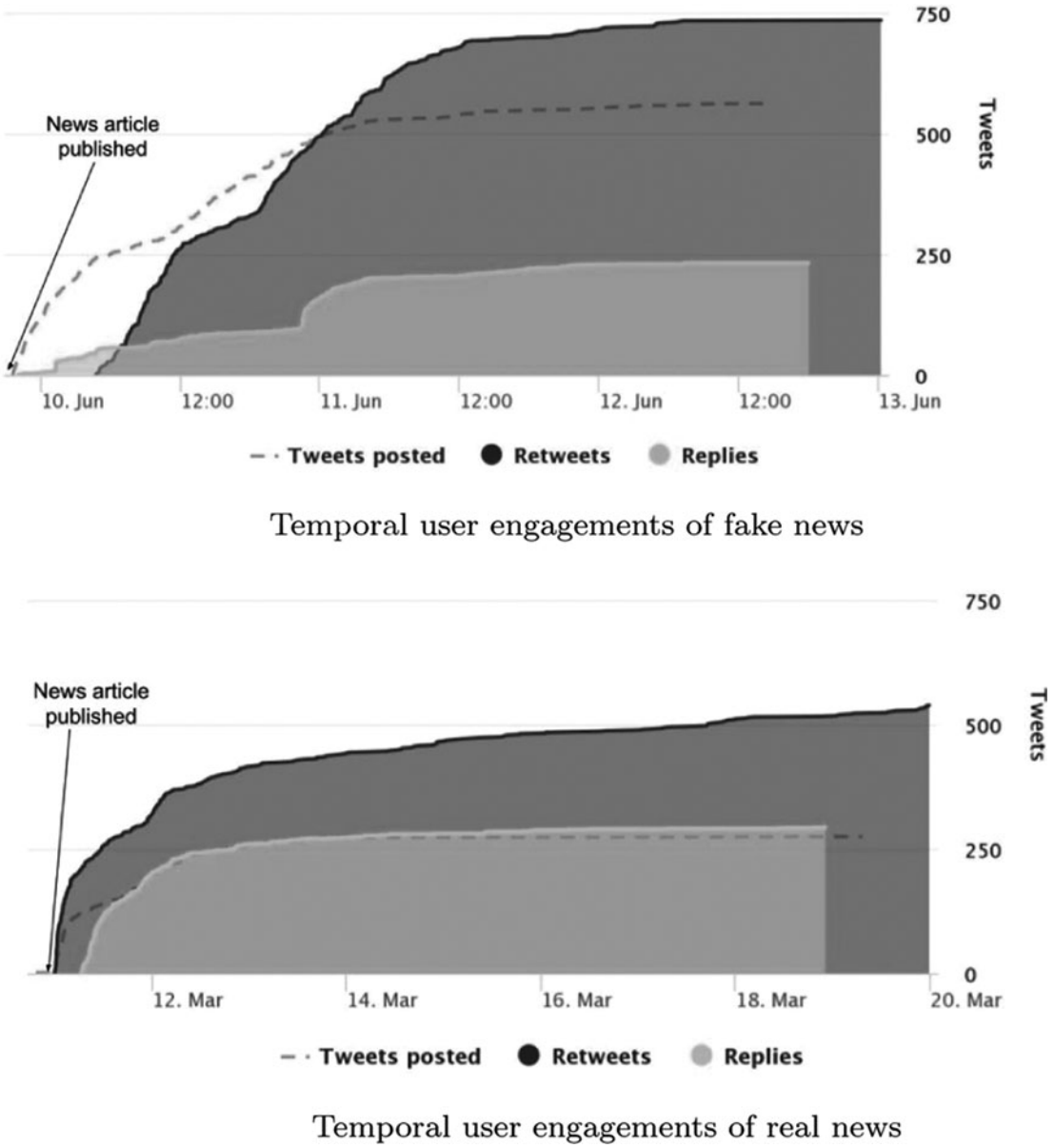


FIG. 8. The comparison of temporal user engagements of fake and real news.

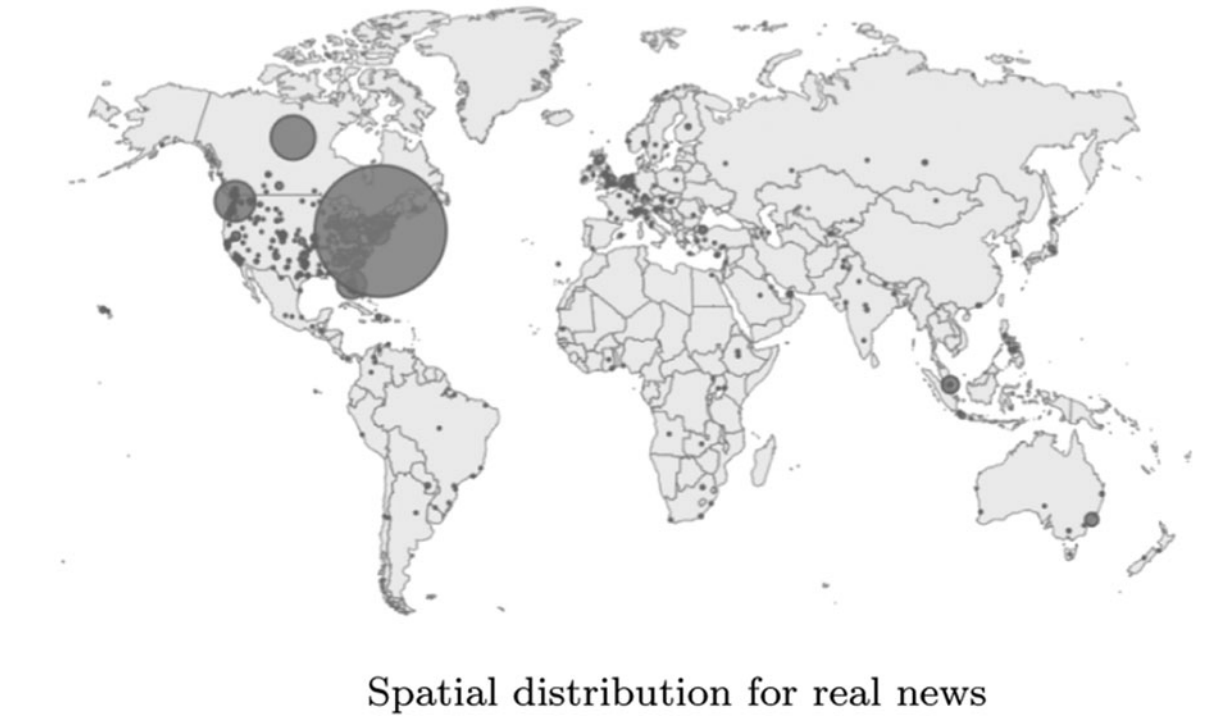
utilize existing approaches for inferring the locations.[24] It would be interesting to explore how users are geolocated distributes using FakeNewsNet repository from different perspectives.

Fake news detection performance

In this subsection, we utilize the PolitiFact and GossipCop data sets from FakeNewsNet repository to perform fake news detection. We use 80% of data for training and 20% for testing. For evaluation, we use accuracy and F1 score.

- *News content:* To evaluate the news contents, the text contents from source news articles are represented as a one-hot encoded vector and then we apply standard machine learning models, including support vector machines (SVMs), logistic regression (LR), Naive Bayes (NB), and Convolutional Neural Network (CNN). For SVM, LR, and NB,

Spatial distribution for fake news



Spatial distribution for real news

FIG. 9. Spatial distribution of users posting tweets related to fake and real news in PolitiFact data set.

we used the default settings provided in the scikit-learn and do not tune parameters. For CNN we use the standard implementation with default setting.* We also evaluate the classification of news articles using social article fusion (SAF/S)[4] model that utilizes autoencoder for learning features from news articles to classify new articles as fake or real.

- *Social context:* To evaluate the social context, we utilize the variant of SAF model,[4] that is, SAF/A, which utilizes the temporal pattern of the user engagements to detect fake news.
- *News content and social context:* SAF model that combines SAF/S and SAF/A is used. This model uses autoencoder with long short-term memory (LSTM) cells of two layers for encoder as well as decoder and also temporal pattern of the user engagements are also captured using another network of LSTM cells with two layers.

The experimental results are shown in Table 3. We can see that (1) among news content-based methods, SAF/S performs better in terms of accuracy and F1 score. SAF/A provides a similar result ~66.7% accuracy as SAF/S. The compared baselines models provide reasonably good performance results for the fake news detection where accuracy is mostly ~65% on PolitiFact; (2) we observe that SAF relatively achieves better accuracy than both SAF/S and SAF/A for both data set. For example, SAF has ~5.65% and 3.60% performance improvement than SAF/S and SAF/A on PolitiFact in terms of accuracy. This indicates that user engagements can help fake news detection in addition to news articles on PolitiFact data set.

In summary, FakeNewsNet provides multiple dimensions of information that has the potential to benefit researchers to develop novel algorithms for fake news detection.

Data Structure

In this section, we describe in details of the structure of FakeNewsNet. We will introduce the data format and provide API interfaces that allows for efficient downloading of data set under the policy of social media platforms.

API interfaces

The full data set is massive and the actual content cannot be directly distributed because of Twitter's sharing policy.† The data set‡ is referenced using DOI§ and adheres FAIR Data Principles.** The APIs are provided in the form of multiple Python scripts that are well-documented and comma-separated values file with news content URLs and associated tweet id's are provided as well. To initiate the download, the user needs to simply run the *main.py* file with the required configuration. The APIs make use of Twitter Access tokens to fetch information related to tweets. These APIs can help to download specific subsets of data set such as linguistic content, tweet information, retweet information, user information and social network. Since Twitter does not provide APIs to download replies and likes of tweets, web scrapping tools can be used. For reviewing purposes, we include the comprehensive data sets through this link.††

Table 3. Fake news detection performance on FakeNewsNet

	PolitiFact		*GossipCop*	
Model	*Acc.*	*F1*	*Acc.*	*F1*
Support vector machines	0.580	0.659	0.497	0.595
Logistic regression	0.642	0.633	0.648	0.646
Naive Bayes	0.617	0.651	0.624	0.649
CNN	0.629	0.583	0.723	0.725
SAF/S	0.654	0.681	0.689	0.703
SAF/A	0.667	0.619	0.635	0.706
SAF	0.691	0.706	0.689	0.717

CNN, Convolutional Neural Network; SAF, social article fusion.

Data format

The news pieces from different platforms/domains are stored in different directories. For example, *gossipcop/fake* directory will contain fake news samples from gossipcop data set. Each directory will possess the associated autogenerated news ID as its name and contain the following structure: *news_article.json* file, *tweets* folder, *retweets* folder, *replies* folder, and *likes* folder.

- *news_article.json* includes all the meta information of the news articles collected using the provided news source URLs. This is a JSON object with attributes, including the following:

  text is the text of the body of the news article.
  images is a list of the URLs of all the images in the news article web page.

*https://github.com/dennybritz/cnn-text-classification-tf

†https://developer.twitter.com/en/developer-terms/agreement-and-policy
‡To access the dataset, we have published code implementation available at https://github.com/KaiDMML/FakeNewsNet that allows the users to download specific subsets of data.
§https://doi.org/10.7910/DVN/UEMMHS
**https://www.force11.org/group/fairgroup/fairprinciples
††https://www.dropbox.com/sh/nx4w5125t5us7pf/AACg2H1BJ_iPHiaI-zAzNSAka?dl=0

publish date indicates the publishing date of that article.

- *tweets folder* contains the metadata of the list of tweets associated with the news article. Each file in this folder contains the tweet objects returned by the Twitter API.
- *retweets folder* includes a list of files containing the retweets of tweets posting the news articles. Each file is named as <tweet id>.json and have a list of retweet objects collected using Twitter API.
- *replies folder* contains files, including replies and conversation threads of tweets sharing the news such as reply text, user details, and reply timestamps.
- *likes folder* comprises files containing a list of IDs for users who have liked each of the tweets sharing the news article.

In addition, we store the metadata of all users, including profiles, historical tweets, followers, and followees through the following folders. Each of these folders contains files named as <user id>.json indicating a particular user details. Note that we only show the metadata of 5000 users in the provided *link* due to the space limitation.

- *user_profiles folder* includes files containing all the metadata of the users in the data set. Each file is this directory is a JSON object collected from Twitter API containing information about the user, including profile creation time, geolocation of the user, profile image URL, followers count, followees count, number of tweets posted, and number of tweets favorited.
- *user_timeline_tweets folder* includes JSON files containing the list of at most 200 recent tweets posted by the user. This includes the complete tweet object with all information related to tweet.
- *user_followers folder* includes JSON files containing a list of user IDs of users following a particular user.
- *user_following folder* includes JSON files containing a list of user IDs a particular user follows.

Potential Applications

FakeNewsNet contains information from multidimensions that could be useful for many applications. We believe FakeNewsNet would benefit the research community for studying various topics such as (early) fake news detection, fake news evolution, fake news mitigation, malicious account detection.

Fake news detection

One of the challenges for fake news detection is the lack of labeled benchmark data set with reliable ground truth labels and comprehensive information space, based on which we can capture effective features and build models. FakeNewsNet can help the fake news detection task because it has reliable labels annotated by journalists and domain experts, and multidimensional information from news content, social context, and spatiotemporal information.

First, news contents are the fundamental sources to find clues to differentiate fake news pieces. For example, studies have shown that news contents can be modeled with tensor embedding in a semisupervised or unsupervised manner to detect fake news.[25,26] In addition, news representation can be obtained with deep neural networks to improve fake news detection.[27,28] In FakeNewsNet, we provide various attributes of news articles such as publishers, headlines, body texts, and images/videos. This information can be used to extract different linguistic features and visual features to further build detection models for clickbaits or fake news. Since we directly collect news articles from fact-checking websites such as PolitiFact and GossipCop, we provide detailed explanations from the fact-checkers, which are useful to learn common and specific perspectives of in what aspects the fake news pieces are formed.

Second, user engagements represent the news proliferation process over time, which provides useful auxiliary information to infer the veracity of news articles.[29] Generally, there are three major aspects of the social context: users, generated posts, and networks. Since fake news pieces are likely to be created and spread by nonhuman accounts, such as bots.[14] Thus, capturing users' profiles and characteristics can provide useful information for fake news detection. Also, people express their emotions or opinions toward fake news through social media posts and thus we collect all the user posts for news pieces, as well as engagements such as reposts, comments, likes, which can be utilized to extract abundant features to capture fake news patterns. Moreover, fake news dissemination processes tend to form an echo chamber cycle, highlighting the value of extracting network-based features to represent these types of network patterns for fake news detection. We provide a large-scale social network of all the users involving in the news dissemination process.

Third, early fake news detection aims to give early alerts of fake news during the dissemination process before it reaches a broad audience.[23] Therefore, early fake

news detection methods are highly desirable and socially beneficial. For example, capturing the pattern of user engagements in the early phases could be helpful to achieve the goal of unsupervised detection. Recent approaches utilize advanced deep generative models to generate synthetic user comments to help improve fake news detection performance.[5] FakeNewsNet contains all these types of information, which provides potentials to further explore early fake news detection models. In addition, FakeNewsNet contains two data sets of different domains, that is, political and entertainment, which can help to study common and different patterns for fake news under different topics.

Fake news evolution

The fake news diffusion process also has different stages in terms of people's attention and reactions as time goes by, resulting in a unique life cycle. For example, breaking news and in-depth news demonstrate different life cycles in social media,[30] and social media reactions can help predict future visitation patterns of news pieces accurately even at an early stage. We can have a deeper understanding of how particular stories "go viral" from normal public discourse by studying the fake news evolution process. First, tracking the life cycle of fake news on social media requires recording essential trajectories of fake news diffusion in general.[31] Thus, FakeNewsNet has collected the related temporal user engagements that can keep track of these trajectories. Second, for a specific news event, the related topics may keep changing over time and be diverse for fake news and real news. FakeNewsNet is dynamically collecting associated user engagements and allows us to perform comparison analysis (Fig. 8), and further investigate distinct temporal patterns to detect fake news.[22] Moreover, statistical time series models such as temporal point process can be used to characterize different stages of user activities of news engagements.[32] FakeNewsNet enables the temporal modeling from real-world data sets, which is otherwise impossible from synthetic data sets.

Fake news mitigation

Fake news mitigation aims to reduce the negative effects brought by fake news. During the spreading process of fake news, users play different roles such as *provenances*: the sources or originators who publish fake news pieces; *persuaders*: who spread fake news with supporting opinions; and *clarifiers*: who propose skeptically and opposing viewpoints toward fake news and try to clarify them. Identifying key users on social media is important to mitigate the effect of fake news.[33] For example, provenances can help answer questions such as whether the piece of news has been modified during its propagation. In addition, it is necessary to identify influential persuaders to limit the spread scope of fake news by blocking the information flow from them to their followers.[6] FakeNewsNet provides rich information about users who post, like, comment on fake and real news pieces (Fig. 6), which enables the exploration of identifying different types of users.

To mitigate the effect of fake news, network intervention aims to develop strategies to control the widespread dissemination of fake news before it goes viral. Two major strategies of network intervention are (1) *Influence Minimization*: minimizing the spread scope of fake news during dissemination process; (2) *Mitigation Campaign*: Limiting the impact of fake news by maximizing the spread of true news. FakeNewsNet allows researchers to build a diffusion network with spatiotemporal information and can facilitate the deep understanding of minimizing the influence scopes. Furthermore, we may able to identify the fake news and real news pieces for a specific event from FakeNewsNet and study the effect of mitigation campaigns in real-world data sets.

Malicious account detection

Studies have shown that malicious accounts that can amplify the spread of fake news include social bots, trolls, and cyborg users. Social bots can give a false impression that information is highly popular and endorsed by many people, which enables the echo chamber effect for the propagation of fake news. We can study the nature of users who spread fake news and identify the characteristics of bot accounts used in fake news diffusion process through FakeNewsNet. Using features such as user profile metadata and historical tweets of users who spread fake news along with social network one could analyze the differences in characteristics of users to cluster them as malicious or not. Through a preliminary study in Figure 4, we have shown that bot users are more likely to exist in the fake news spreading process. Although existing works have studied bot detection in general, few studies investigate the influences of social bots for fake news spreading. FakeNewsNet could potentially facilitate the study of understanding the relationship between fake news and social bots, and furthermore, explore the mutual benefits of studying fake news detection or bot detection.

Conclusion and Future Work

In this article, we provide a comprehensive repository FakeNewsNet that contains news content, social context, and spatiotemporal information. We propose a principled strategy to collect relevant data from different sources. Moreover, we perform a preliminary exploration study on various features on FakeNewsNet and demonstrate its utility through fake news detection task over several state-of-the-art baselines. FakeNewsNet has the potential to facilitate many promising research directions such as fake news detection, mitigation, evolution, and malicious account detection.

There are several interesting options for future work. First, FakeNewsNet repository can be extended to other reliable news sources such as other fact-checking websites or curated data collections. Second, the selection strategy can be used for web search results to reduce noise in the data collection process. Third, FakeNewsNet repository can be integrated with front-end software and build an end-to-end system for fake news study.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

This material is in part supported by the National Science Foundation awards #1909555, #1614576, #1742702, #1820609, and #1915801.

References

1. Shu K, Sliva A, Wang S, et al. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsl. 2019;19: 22–36.
2. Kumar S, and Shah N. False information on web and social media: A survey. arXiv 2018;arXiv:1804.08559.
3. Kumar S, West R, Leskovec J. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In: Proceedings of the 25th International Conference on World Wide Web, WWW '16, Republic and Canton of Geneva. CHE, 2016, pp. 591–602.
4. Shu K, Mahudeswaran D, Liu H. Fake NewsTracker: a tool for fake news collection, detection, and visualization. Comput Math Organ Th 2019; 25:60–71.
5. Qian F, Gong CY, Sharma K, Liu Y. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. IJCAI-18, pp. 3834–3840.
6. Shu K, Bernard HR, Liu H. Studying fake news via net-work analysis: Detection and mitigation. arXiv 2018;arXiv:1804.10233.
7. Wang WY. "liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv 2017;arXiv:1705.00648.
8. Mitra T, Gilbert E. CREDBANK: A large-scale social media corpus with associated credibility annotations. In: Ninth International AAAI Conference on Web and Social Media, 2015.
9. Santia GC, Williams JR. BuzzFace: A news veracity dataset with Facebook user commentary and egos. In: Twelfth International AAAI Conference on Web and Social Media, 2018.
10. Tacchini E, Ballarin G, Della Vedova ML, et al. Some like it hoax: Automated fake news detection in social networks. arXiv 2017;arXiv: 1704.07506.
11. Manning C, Surdeanu M, Bauer J, et al. The Stanford CoreNLP natural language processing toolkit. In: ACL'14, 2014. pp. 55–60.
12. Baccianella S, Esuli A, Sebastiani F. Sentiwordnet3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC'10, 2010. pp. 2200–2204.
13. Shu K, Wang S, Liu H. Understanding user profiles on social media for fake news detection. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Miami, FL: IEEE, April 10–12, 2018. pp. 430–435.
14. Shao C, Ciampaglia GL, Varol O, et al. The spread of fake news by social bots. arXiv 2017;arXiv:1707.07592.
15. Davis CA, Varol O, Ferrara E, et al. Botornot: A system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web, 2016. pp. 273–274.
16. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science. 2018;359:1146–1151.
17. Jin Z, Cao J, Zhang Y, Luo J. News verification by exploiting conflicting social viewpoints in microblogs. In: Thirtieth AAAI conference on artificial intelligence, 2016.
18. Qazvinian V, Rosengren E, Radev DR, Mei Q. Rumor has it: Identifying misinformation in microblogs. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011. pp. 1589–1599.
19. Hutto Eric Gilbert CJ. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media, 2014.
20. Kim A, Dennis AR. Says who?: How news presentation format influences perceived believability and the engagement level of social media users. In: Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.
21. Del Vicario M, Vivaldo G, Bessi A, et al. Echo chambers: Emotional contagion and group polarization on Facebook. Sci Rep. 2016;6: 37825.
22. Ruchansky N, Seo S, Liu Y. Csi: A hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017. pp. 797–806.
23. Liu YP, Brook Wu Y-F. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
24. Zubiaga A, Voss A, Procter R, et al. Towards real-time, country-level location classification of worldwide tweets. IEEE Trans Knowl Data Eng. 2017;29:2053–2066.
25. Guacho GB, Abdali S, Shah N, Papalexakis EE. Semi-supervised content-based detection of misinformation via tensor embeddings. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018. IEEE, pp. 322–325.
26. Hosseinimotlagh S, Papalexakis EE. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In: Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2), 2018.
27. Karimi H, Roy P, Saba-Sadiya S, Tang J. Multi source multi-class fake news detection. In: Proceedings of the 27th International Conference on Computational Linguistics, 2018. pp. 1546–1557.
28. Karimi H, Tang J. Learning Hierarchical Discourse-level Structure for Fake News Detection. arXiv 2019;arXiv:1903.07389.

29. Shu K, Wang S, Liu H. Beyond news contents: The role of social context for fake news detection. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019. pp. 312–320.
30. Castillo C, El-Haddad M, Pfeffer J, Stempeck M. Characterizing the life cycle of online news stories using social media reactions. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 2014. pp. 211–223.
31. Shao C, Ciampaglia GL, Flammini A, Menczer F. Hoaxy: A platform for tracking online misinformation. In: Proceedings of the 25th international conference companion on worldwide web, 2016. pp. 745–750.
32. Farajtabar M, Yang J, Ye X, et al. Fake news mitigation via point process based intervention. arXiv 2017;arXiv:1703.07823.
33. Alassad M, Hussain MN, Agarwal N. Finding fake news key spreaders in complex social networks by using bi-level decomposition optimization method. In: International Conference on Modelling and Simulation of Social-Behavioural Phenomena in Creative Societies, 2019. pp. 41–54.

Cite this article as: Shu K, Mahudeswaran D, Wang S, Lee D, Liu H (2020) FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data* 8:3, 171–188, DOI: 10.1089/big.2020.0062.

Abbreviations Used

API = application programming interface
CNN = Convolutional Neural Network
LR = logistic regression
LSTM = long short-term memory
NB = Naive Bayes
SAF = social article fusion
SVM = support vector machine