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]

[1]Computer Science and Engineering, Arizona State University, Tempe, 85281, USA
[2]College of Information Sciences and Technology, Penn State University, University Park, PA, 16802, USA
{kai.shu, dmahudes, huan.liu}@asu.edu, {szw494, dongwon}@psu.edu

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*, i.e., 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, however, the lack of comprehensive and community-driven fake news datasets has become one of major roadblocks. Not only existing datasets 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 paper, to facilitate fake news related research, we present a fake news data repository *FakeNewsNet*, which contains two comprehensive datasets 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 datasets from different perspectives, and discuss the benefits of the FakeNewsNet for potential applications on fake news study on social media.

1 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[1]. As another example, studies from Pew Research Center shows that around 68% of Americans get some of their news on social media in 2018[2] 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 (Shu et al. 2017) 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 (Shu et al. 2017). It is also reported that fake news has an influence on the 2016 US presidential elections[3]. Finally, fake news may cause significant effects on real-world events. For example, "Pizzagate", a piece of fake news from Reddit, leads to a real shooting [4]. 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 datasets which 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 datasets that have news content, social context and spatiotemporal information to facilitate fake news research. However, to the best of our knowledge, existing datasets only cover one or two aspects.

Therefore, in this paper, we construct and publicize a multi-dimensional data repository *FakeNewsNet*[5], which currently contains two datasets with news content, social context, and spatiotemporal information. The dataset is constructed using an end-to-end system,FakeNewsTracker[6] (Shu, Mahudeswaran, and Liu 2018).The constructed FakeNewsNet repository has the potential to boost the study of various open research problems related to fake news study.

[1]https://www.socialmediatoday.com/marketing/how-much-time-do-people-spend-social-media-infographic

[2]http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

[3]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

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

[5]https://github.com/KaiDMML/FakeNewsNet

[6]http://blogtrackers.fulton.asu.edu:3000

First, the rich set of features in the datasets provides an opportunity to experiment with different approaches for fake new 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 dataset (Qian et al.). Third, we can investigate the fake news diffusion process by identifying provenances, persuaders, and developing better fake news intervention strategies (Shu, Bernard, and Liu 2018). 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 paper are:

- We construct and publicize a multi-dimensional 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 datasets from different perspectives to demonstrate the quality of the datasets, understand their characteristics and provide baselines for future fake news detection; and
- We discuss benefits and provides insight for potential fake news studies on social media with FakeNewsNet.

2 Background and Related Work

Fake news detection in social media aims to extract useful features and build effective models from existing social media datasets for detecting fake news in the future. Thus, a comprehensive and large-scale dataset with multi-dimension information in online fake news ecosystem is important. The multi-dimension 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. Though there exist several datasets for fake news detection, the majority of them only contains linguistic features. Few of them contains both linguistic and social context features. To facilitate research on fake news, we provide a data repository which 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 datasets below and compare them with the FakeNewsNet repository in Table 1.

BuzzFeedNews [7]: This dataset comprises a complete sample of news published in Facebook from 9 news agencies over a week close to the 2016 U.S. election from September 19 to 23 and September 26 and 27. Every post and the linked article were fact-checked claim-by-claim by 5 BuzzFeed journalists. It contains 1,627 articles 826 mainstream, 356 left-wing, and 545 right-wing articles.

LIAR [8]: This dataset (Wang 2017) is collected from fact-checking website PolitiFact. It has 12.8 K human labeled short 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 [9]: This dataset 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 [10]: This is a large-scale crowd-sourced dataset (Mitra and Gilbert) of around 60 million tweets that cover 96 days starting from Oct. 2015. The tweets are related to over 1,000 news events. Each event is assessed for credibilities by 30 annotators from Amazon Mechanical Turk.

BuzzFace [11]: This dataset (Santia and Williams) is collected by extending the BuzzFeed dataset with comments related to news articles on Facebook. The dataset contains 2263 news articles and 1.6 million comments discussing news content.

FacebookHoax [12]: This dataset (Tacchini et al. 2017) comprises information related to posts from the facebook pages related to scientific news (non- hoax) and conspiracy pages (hoax) collected using Facebook Graph API. The dataset contains 15,500 posts from 32 pages (14 conspiracy and 18 scientific) with more than 2,300,000 likes.

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

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

3 Dataset Integration

In this section, we introduce the dataset integration process for the FakeNewsNet repository. We demonstrate (see Figure 1) how we can collect news contents with reliable ground

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

[8] https://www.cs.ucsb.edu/ william/software.html

[9] https://github.com/bs-detector/bs-detector

[10] http://compsocial.github.io/CREDBANK-data/

[11] https://github.com/gsantia/BuzzFace

[12] https://github.com/gabll/some-like-it-hoax

Table 1: Comparison with existing fake news detection datasets

Features / Dataset	News Content		Social Context				Spatiotemporal Information	
	Linguistic	*Visual*	*User*	*Post*	*Response*	*Network*	*Spatial*	*Temporal*
BuzzFeedNews	✓							
LIAR	✓							
BS Detector	✓							
CREDBANK	✓		✓	✓			✓	✓
BuzzFace	✓			✓	✓			✓
FacebookHoax	✓		✓	✓	✓			
FakeNewsNet	✓	✓	✓	✓	✓	✓	✓	✓

truth labels, how we obtain additional social context and spatialtemporal information.

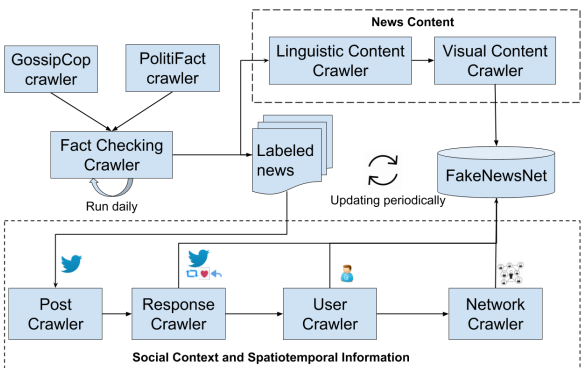


Figure 1: The flowchart of dataset integration process for FakeNewsNet. It mainly describes the collection of news content, social context and spatiotemporal information.

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* [13] and *GossipCop* [14]. In PolitiFact, journalists and domain experts review the political news and provide fact-checking evaluation results to claim news articles as fake [15] or real [16]. 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 i) check if the removed page was archived and automatically retrieve content at the Wayback Machine [17]; and ii) make use of Google web search in automated fashion 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 to 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 less than 5, which is mainly because the purpose of GossipCop is to showcase more fake stories. In order to collect true entertainment news pieces, we crawl the news articles from E! Online[18], 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 less than 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 the Wayback Machine archive to obtain the news source information. The headline of the GossipCop story articles are generally written to reflect the actual fact and may not be used directly. For example, one of the headlines, "Jennifer Aniston NOT Wearing Brad Pitts Engagement Ring, Despite Report" mentions the actual fact instead of the original news articles title. We utilize some heuristics to extract proper headlines such as i) using the text in quoted string; ii) removing negative sentiment words. For example, some headlines include quoted string which are exact text from the original news source. In this case, we extract the named entities through Stanford's CoreNLP tool (Manning et al. 2014) from the headline and quoted strings from the headline to form the search query. For example, in the 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 entites mostly provides the context of the original news. As another example, the headline are written in the negative sense to correct the false information, e.g., "Jennifer Aniston NOT Wearing Brad Pitts Engagement Ring, Despite Report". So we remove negative sentiment words retrieved from SentiWordNet(Baccianella, Esuli, and Sebastiani 2010) and some hand-picked words from the headline to form the search query, e.g., "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 API provided by social media platforms

[13] https://www.politifact.com/

[14] https://www.gossipcop.com/

[15] available at https://www.politifact.com/subjects/fake-news/

[16] available at https://www.politifact.com/truth-o-meter/rulings/true/

[17] https://archive.org/web/

[18] https://www.eonline.com/

Table 2: Statistics of the FakeNewsNet repository

	Category	Features	PolitiFact		GossipCop	
			Fake	Real	Fake	Real
News Content	*Linguistic*	# News articles	432	624	5,323	16,817
		# News articles with text	420	528	4,947	16,694
	Visual	# News articles with images	336	447	1,650	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	31692	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
Spatiotemporal Information	*Spatial*	# User profiles with locations	217,379	719,331	429,547	220,264
		# Tweets with locations	3,337	12,692	12,286	2,451
	Temporal	# Timestamps for news pieces	296	167	3,558	9,119
		# Timestamps for response	171,301	669,641	381,600	200,531

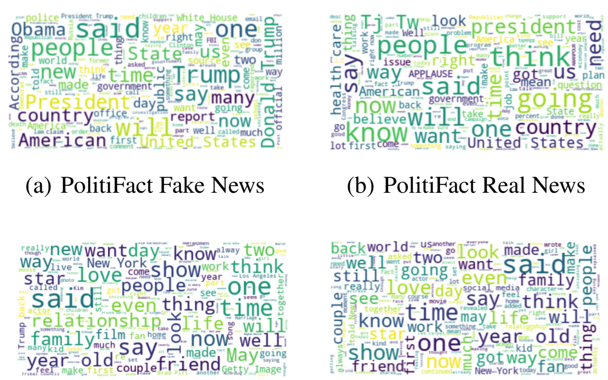
such as the Twitter's Advanced Search API [19]. The search queries for collecting user engagements are formed from the headlines of news articles, with special characters removed from the search query to filter out the noise. After we obtain the social media posts that directly spread news pieces, we further fetch the user *response* towards 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.

4 Data Analysis

FakeNewsNet has multi-dimensional 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



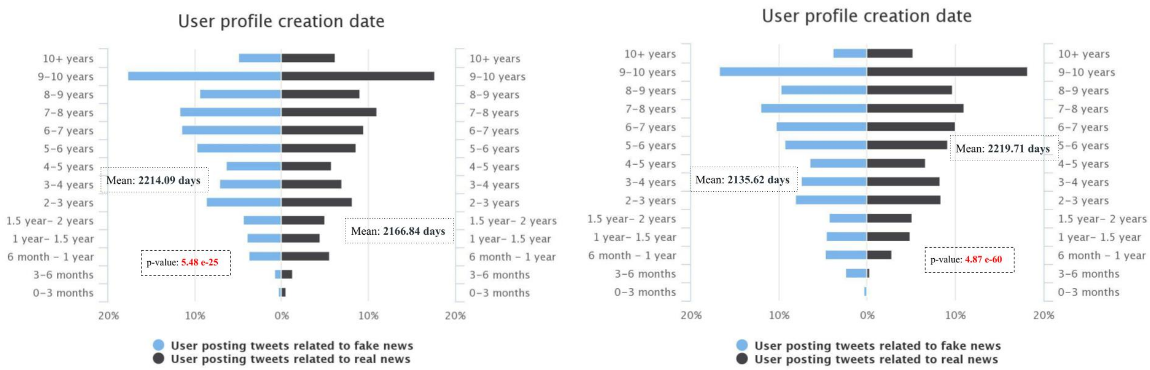
(a) PolitiFact Fake News (b) PolitiFact Real News

(c) GossipCop Fake News (d) GossipCop Real News

Figure 2: The word cloud of new body text for fake and real news on PolitiFact and GossipCop.

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 to detect fake news. First, we analyze the topic distribution of fake and real news articles. From figures 2(a) and 2(b), we can observe that the fake and real news of the PolitiFact dataset is mostly related to the political campaign. In case of GossipCop dataset from figures 2(c) and 2(d), we observe that the fake and real news are mostly related to gossip about the relationship among

[19]https://twitter.com/search-advanced?lang=en



(a) PolitiFact dataset

(b) GossipCop dataset

Figure 3: The distribution of user profile creation dates on PolitiFact and GossipCop datasets

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 (Shu et al. 2017), 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 datasets. 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 6,048 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.

Measuring Social Context

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, 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 (Shu, Wang, and Liu 2018). Research has also shown that fake news pieces are likely to be created and spread by non-human accounts, such as social bots or cyborgs (Shu et al. 2017; Shao et al. 2017). 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 are different or not. We compute the time range 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 who posting fake news is significantly different from those who post real news, with the p-value$<$ 0.05 under statistical t-test. In addition, we notice that it's 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; while we see opposite case in Gossipcop dataset.

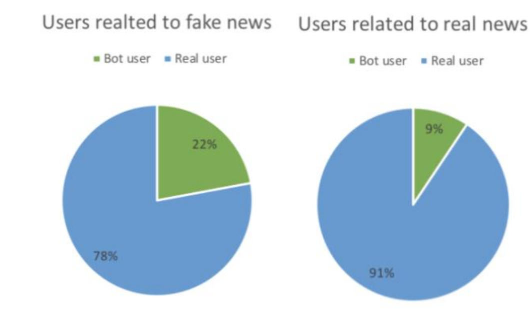
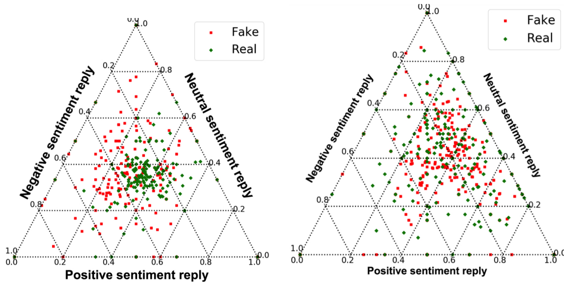


Figure 4: Comparison of bot scores on users related to fake and real news on PolitiFact dataset.

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 one of the state-of-the-art bot detection algorithm Botometer (Davis et al.) API[20]. The Botometer takes a Twitter username as input and utilizes various features extracted from meta-data and output a probability score in $[0, 1]$, indicating how likely the user is a social 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.

[20]https://botometer.iuni.iu.edu/



(a) PolitiFact dataset (b) GossipCop dataset

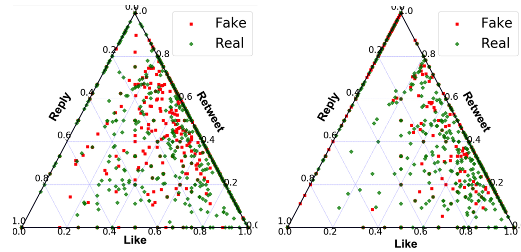
Figure 5: Ternary plot of the ratio of the positive, neutral and negative sentiment replies for fake and real news.

For example, almost 22% of users involved in fake news are bots, while only around 9% of users are predicted as bot users for real news. Similar results were observed with different thresholds on bot scores based on both datasets. This indicates that there are bots in Twitter for spreading fake news, which is consistent with the observation in (Shao et al. 2017). In addition, most users that spread fake news (around 78%) are still more likely to be humans than bots (around 22%), which is also in consistence with the findings in (Vosoughi, Roy, and Aral 2018).

Post and Response People express their emotions or opinions towards fake news through social media posts, such as skeptical opinions, sensational reactions, etc. These features are important signals to study fake news and disinformation in general (Jin et al. ; Qazvinian et al.).

We perform sentiment analysis on the replies of user posts that spreading fake news and real news using one of the state-of-the-art unsupervised prediction tool called VADER [21] (Gilbert). It is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. 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 5(a). 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 dataset shown in Figure 5(b), 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 the likes, retweets, and replies of the tweets, which can help gain insights on user interaction networks 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 (Kim and Dennis 2017). For example, the behaviors of likes and retweets are more emotional while replies are more rational.



(a) PolitiFact dataset (b) GossipCop dataset

Figure 6: Ternary plot of the ratio of likes, retweet and reply of tweets related to fake and real news

We plot the ternary triangles which illustrate the ratio of replies, retweets, and likes from the second order engagements towards the posts that spread fake news or real news pieces. From Figure 6, we observe that the: i) fake news pieces tend to have fewer replies and more retweets; ii) 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 datasets to understand the social factors of user engagements and underly social science as well.

Network 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 (Shu et al. 2017). 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 (Del Vicario et al. 2016).

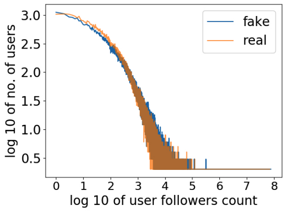
We look at the social network statistics of all the users that spread fake news or real news. The social network features such as followers count and followee count can be used to estimate the scope of how the fake news can spread in social media. We plot the distribution of follower count and followee count of users in Figure 7. We can see that: i) the follower and followee count of the users generally follows power law distribution, which is commonly observed in social network structures; ii) there is a spike in the followee count distribution of both users and this is because of the restriction imposed by Twitter[22] on users to have at most 5000 followees when the number of following is less than 5000.

Characterizing Spatiotemporal Information

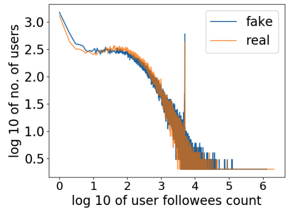
Recent research has shown users' temporal responses can be modeled using deep neural networks to help detection fake news (Ruchansky, Seo, and Liu), and deep generative models can generate synthetic user engagements to help early fake news detection (Liu and Wu). The spatiotemporal information in FakeNewsNet depicts the temporal user en-

[21] https://github.com/cjhutto/vaderSentiment

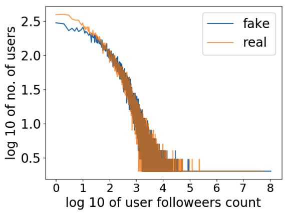
[22] https://help.twitter.com/en/using-twitter/twitter-follow-limit



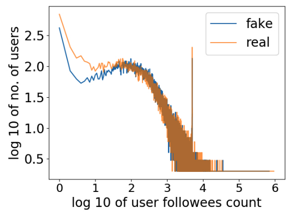
(a) Follower count of users in PolitiFact dataset



(b) Followee count of users in PolitiFact dataset

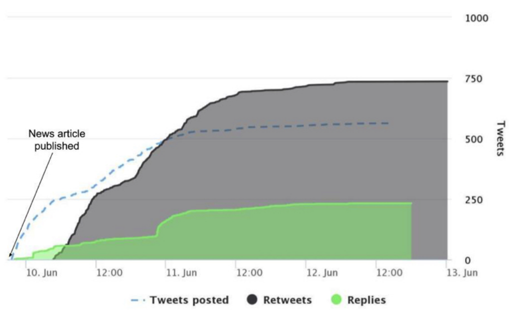


(c) Follower count of users in Gossipcop dataset

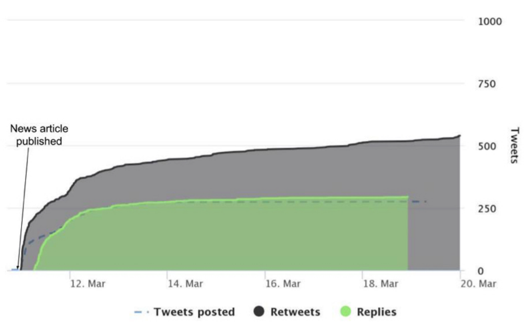


(d) Followee count of users in PolitiFact dataset

Figure 7: The distribution of the count of followers and followees related to fake and real news

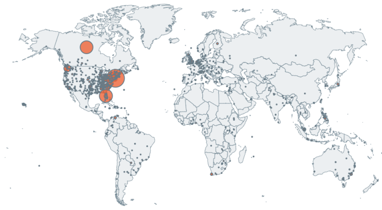


(a) Temporal user engagements of fake news

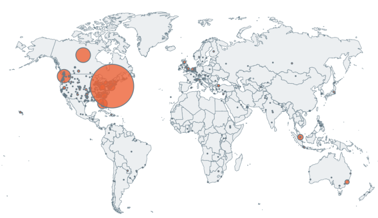


(b) Temporal user engagements of real news

Figure 8: The comparison of temporal user engagements of fake and real news



(a) Spatial distribution of users posting fake news



(b) Spatial distribution of users posting real news

Figure 9: Spatial distribution of users posting tweets related to fake and real news.

gagements 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, retweets, are different for fake news and real news with similar topics, e.g., fake news "*TRUMP APPROVAL RATING Better than Obama and Reagan at Same Point in their Presidencies*" from June 9, 2018 to 13 June, 2018 and real news "*President Trump in Moon Township Pennsylvania*" from March 10, 2018 to 20 March, 2018. As shown in Figure 8, we can observe that: i) 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; ii) 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 (Shu, Bernard, and Liu 2018).

Next, we demonstrate the geo-location distribution of users engaging in fake and real news (See Figure 9 for Politifact dataset). We show the locations explicitly provided by users in their profiles, and we can see that users in the PolitiFact dataset who posting 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 utilize existing approaches for inferring the locations (Zubiaga et al.). It would be interesting to explore how users are geo-located distributes using FakeNewsNet repository from different perspectives.

Fake News Detection Performance

In this subsection, we utilize the PolitiFact and GossipCop datasets from FakeNewsNet repository to perform fake news detection. We use 80% of data for training and 20% for testing. For evaluation metrics, we use accuracy, precision, recall and F1 score. We deployed several state-of-the-art baselines for fake news detection,

- **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 (SVM), logistic regression (LR), Naive Bayes (NB), and CNN. For SVM, LR, and NB, 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 [23]. We also evaluate the classification of news articles using Social article fusion (SAF /S) (Shu, Mahudeswaran, and Liu 2018) model that utilizes autoencoder for learning features from news articles to classify new articles as fake or real.
- **Social context:** In order to evaluate the social context, we utilize the variant of SAF model (Shu, Mahudeswaran, and Liu 2018), i.e., SAF /A, which utilize the temporal pattern of the user engagements to detect fake news.
- **News content and social context:** Social Article Fusion(SAF) model that combines SAF /S and SAF /A is used. This model uses autoencoder with LSTM cells of 2 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 2 layers.

The experimental results are shown in Table 4. We can see that: i) For news content-based methods, SAF /S perform better in terms of accuracy, recall, and F1 score while logistic regression has better precision than others. SAF /A provides a similar result around 66.7% accuracy as SAF /S but has higher precision. The compared baselines models provide reasonably good performance results for the fake news detection where accuracy is mostly around 65% on PolitiFact; ii) we observe that SAF relatively achieves better accuracy than both SAF /S and SAF /A for both dataset. For example, SAF has around 5.65% and 3.68% performance improvement than SAF /S and SAF /A on PolitiFact. This indicates that user engagements can help fake news detection in addition to news articles on PolitiFact dataset.

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

5 Data Structure

In this section, we describe in details of the structure of FakeNewsNet. We will introduce the data format and provided API interfaces that allows for efficient slicing of the data.

[23]https://github.com/dennybritz/cnn-text-classification-tf

Data Format

Each directory will possess the associated autogenerated news ID as its name and contain the following structure: news content.json file, tweets folder, retweets folder. Finally, user_profiles folder and user_timeline_tweets folder contains the user profile information about all the users involved in tweet provided in the dataset.

- *news content.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:
  - *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.
  - *publish date* indicate the date that news article is published.
- *tweets folder* contains the metadata of the list of tweets associated with the news article collected as separate files for each tweet. Each file in this folder contains the tweet objects returned by Twitter API.
- *retweets folder* includes a list of files containing the retweets of tweets associated with the news article. Each file is named as <tweet id>.json and have a list of retweet objects associated with a particular tweet collected using Twitter API.
- *user_profiles folder* includes files containing all the metadata of the users in the dataset. 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.

API Interface

The full dataset is massive and the actual content cannot be directly distributed because of Twitter's policy [24]. To help readers to better process the data, we have created an API [25] that allows the users to download specific subsets of data. The API is provided in the form of multiple Python scripts which are well-documented and CSV file with news content URLs and associated tweet ids. In order to initiate the API, the user must simply run the main.py file with the required configuration. The API makes use of Twitter Access tokens fetch information related to tweets. Since FakeNewsNet includes multiple data sources, API provides options to select dataset of interest. Additionally, API facilitates user to download specific subsets of dataset like linguistic content only, visual content only, only tweet information only, retweet information only, user information only and social network only.

[24]https://developer.twitter.com/en/developer-terms/agreement-and-policy
[25]https://github.com/KaiDMML/FakeNewsNet

Table 3: Fake news detection performance on FakeNewsNet

Model	PolitiFact				GossipCop			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
SVM	0.580	0.611	0.717	0.659	0.497	0.511	0.713	0.595
Logistic regression	0.642	0.757	0.543	0.633	0.648	0.675	0.619	0.646
Naive Bayes	0.617	0.674	0.630	0.651	0.624	0.631	0.669	0.649
CNN	0.629	0.807	0.456	0.583	0.723	0.751	0.701	0.725
Social Article Fusion /S	0.654	0.600	0.789	0.681	0.689	0.671	0.738	0.703
Social Article Fusion /A	0.667	0.667	0.579	0.619	0.635	0.589	0.882	0.706
Social Article Fusion	0.691	0.638	0.789	0.706	0.689	0.656	0.792	0.717

6 Potential Applications

FakeNewsNet contains information from multi-dimensions which 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 dataset 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 multi-dimension 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, a study has shown that the clickbait's headlines usually can serve as a good indicator for recognizing fake news articles (Chen, Conroy, and Rubin 2015; Shu et al.). In FakeNewsNet, we provide various attributes of news articles such as publishers, headlines, body texts, and videos. These information can be used to extract different linguistic features (Hosseinimotlagh and Papalexakis 2018) and visual features to further build detection models for clickbaits or fake news. For example, style-based approaches try to detect fake news by capturing the manipulators in the writing style of news contents (Potthast et al. 2017; Wang 2017). In addition, Knowledge-based approaches aim to use external sources to fact-check proposed claims in news content (Shu et al. 2017). Since we directly collect news articles from fact-checking websites such as PolitiFact and GossipCop, we provide the information of detail description and explanations from the fact-checkers, which are useful for us 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 (Shu, Wang, and Liu). Generally, there are three major aspects of the social media context: users, generated posts, and networks. Since fake news pieces are likely to be created and spread by non-human accounts, such as social bots or cyborgs (Shao et al. 2017). Thus, capturing users profiles and characteristics by user-based features can provide useful information for fake news detection. FakeNewsNet includes all the meta-data for user profiles. In addition, people express their emotions or opinions towards fake news through social media posts, such as skeptical opinions, sensational reactions, etc. We collect all the user posts for the news pieces, as well as the second engagements (see Table 2) such as reposts, comments, likes, which can be utilized to extract abundant features, e.g., sentiment scores as in Figure 5, to captures 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 (see Table 2).

Third, early fake news detection aims to give early alerts of fake news during the dissemination process before it reaches a broad audience (Liu and Wu). 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 (Yang et al.). Recent approaches utilize advanced deep generative models to generate synthetic user comments to help improve fake news detection performance (Qian et al.). FakeNewsNet contains all these types of information, which provides potentials to further explore early fake news detection models.

In addition, FakeNewsNet contains two datasets of different domains, i.e., 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 peoples 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 (Castillo et al.), 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 (Shao et al.). Thus, FakeNewsNet has collected the related temporal user engagements which 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 (e.g., see Figure 8), and further investigate distinct temporal patterns to detect fake news (Ruchansky, Seo, and Liu). Moreover, statistical time series models such as temporal point process can be used to characterize different stages of user activities of news engagements (Farajtabar et al. 2017). FakeNewsNet enables the temporal modeling from real-world datasets, which is otherwise impossible from synthetic datasets.

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 for publishing fake news pieces; *persuaders*: who spread fake news with supporting opinions; and *clarifiers*: who propose skeptically and opposing viewpoints towards fake news and try to clarify them. Identifying key users on social media is important to mitigate the effect of fake news. For example, the provenances can help answer questions such as whether the piece of news has been modified during its propagation. In addition, its necessary to identify influential persuaders to limit the spread scope of fake news by blocking the information flow from them to their followers on social media (Shu, Bernard, and Liu 2018). FakeNewsNet provides rich information about users who were posting, liking, commenting on fake news and real news pieces (see Figure 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: i) *Influence Minimization*: minimizing the spread scope of fake news during dissemination process; ii) *Mitigation Campaign*: Limiting the impact of fake news by maximizing the spread of true news. FakeNewsNet allows researchers to build a diffusion network of users with spatiotemporal information and thus 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 datasets.

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 are social media accounts that are controlled by a computer algorithm. 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 the user who spread fake news and identify the characteristics of the bot account used in the fake news diffusion process through FakeNewsNet. Using the feature like the user profile metadata and the historical tweets of the user who spread fake news along with the social network one could analyze the differences in characteristics of the users to clusters the users as malicious or not. Through a preliminary study in Figure 4, we have shown that bot users are more likely to exist in fake news spreading process. Even though existing work have studied the bot detection in general, but 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 further, explore the mutual benefits of studying fake news detection or bot detection.

7 Conclusion and Future Work

In this paper, we provide a comprehensive repository FakeNewsNet collected which contains information from 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 tasks over several state-of-the-art baselines. FakeNewsNet has the potential to facilitate many promising research directions such as fake news detection, mitigation, evolution, malicious account detection, etc.

There are several interesting options for future work. First, we will extend the FakeNewsNet repository to other reliable news sources such as other fact-checking websites or curated data collections. Second, we will improve the selection strategy used for web search results to reduce noise in the data collection process. We will also integrate FakeNewsNet repository with front-end software such as FakeNewsTracker (Shu, Mahudeswaran, and Liu 2018), and build an end-to-end system for fake news study.

References

[Baccianella, Esuli, and Sebastiani 2010] Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, 2200–2204.

[Castillo et al.] Castillo, C.; El-Haddad, M.; Pfeffer, J.; and Stempeck, M. Characterizing the life cycle of online news stories using social media reactions. In *CHI'14*.

[Chen, Conroy, and Rubin 2015] Chen, Y.; Conroy, N. J.; and Rubin, V. L. 2015. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, 15–19. ACM.

[Davis et al.] Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. Botornot: A system to evaluate social bots. In *WWW'16*.

[Del Vicario et al. 2016] Del Vicario, M.; Vivaldo, G.; Bessi, A.; Zollo, F.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports* 6:37825.

[Farajtabar et al. 2017] Farajtabar, M.; Yang, J.; Ye, X.; Xu, H.; Trivedi, R.; Khalil, E.; Li, S.; Song, L.; and Zha, H. 2017. Fake news mitigation via point process based intervention. *arXiv preprint arXiv:1703.07823*.

[Gilbert] Gilbert, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM'14*.

[Hosseinimotlagh and Papalexakis 2018] Hosseinimotlagh, S., and Papalexakis, E. E. 2018. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles.

[Jin et al.] Jin, Z.; Cao, J.; Zhang, Y.; and Luo, J. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI'16*.

[Kim and Dennis 2017] Kim, A., and Dennis, A. R. 2017. Says who?: How news presentation format influences perceived believability and the engagement level of social media users.

[Liu and Wu] Liu, Y., and Wu, Y.-f. B. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *AAAI'18*.

[Manning et al. 2014] Manning, C.; Surdeanu, M.; Bauer, J.; Finkel, J.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL'14*, 55–60.

[Mitra and Gilbert] Mitra, T., and Gilbert, E. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM'15*.

[Potthast et al. 2017] Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

[Qazvinian et al.] Qazvinian, V.; Rosengren, E.; Radev, D. R.; and Mei, Q. Rumor has it: Identifying misinformation in microblogs. In *EMNLP'11*.

[Qian et al.] Qian, F.; Gong, C.; Sharma, K.; and Liu, Y. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI'18*.

[Ruchansky, Seo, and Liu] Ruchansky, N.; Seo, S.; and Liu, Y. Csi: A hybrid deep model for fake news detection. In *CIKM'17*.

[Santia and Williams] Santia, G. C., and Williams, J. R. Buzzface: A news veracity dataset with facebook user commentary and egos. In *ICWSM'18*.

[Shao et al.] Shao, C.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. Hoaxy: A platform for tracking online misinformation. In *WWW'16*.

[Shao et al. 2017] Shao, C.; Ciampaglia, G. L.; Varol, O.; Flammini, A.; and Menczer, F. 2017. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*.

[Shu, Bernard, and Liu 2018] Shu, K.; Bernard, H. R.; and Liu, H. 2018. Studying fake news via network analysis: Detection and mitigation. *CoRR* abs/1804.10233.

[Shu et al.] Shu, K.; Wang, S.; Le, T.; Lee, D.; and Liu, H. Deep headline generation for clickbait detection. In *ICDM'18*.

[Shu et al. 2017] Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19(1):22–36.

[Shu, Mahudeswaran, and Liu 2018] Shu, K.; Mahudeswaran, D.; and Liu, H. 2018. Fakenewstracker: a tool for fake news collection, detection, and visualization. *Computational and Mathematical Organization Theory* 1–12.

[Shu, Wang, and Liu] Shu, K.; Wang, S.; and Liu, H. Beyond news contents: The role of social context for fake news detection. In *WSDM'19*.

[Shu, Wang, and Liu 2018] Shu, K.; Wang, S.; and Liu, H. 2018. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE.

[Tacchini et al. 2017] Tacchini, E.; Ballarin, G.; Della Vedova, M. L.; Moret, S.; and de Alfaro, L. 2017. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

[Vosoughi, Roy, and Aral 2018] Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

[Wang 2017] Wang, W. Y. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

[Yang et al.] Yang, S.; Shu, K.; Wang, S.; Gu, R.; Wu, F.; and Liu, H. Unsupervised fake news detection on social media: A generative approach. In *AAAI'19*.

[Zubiaga et al.] Zubiaga, A.; Voss, A.; Procter, R.; Liakata, M.; Wang, B.; and Tsakalidis, A. Towards real-time, country-level location classification of worldwide tweets. *TKDE'17*.