A First Step Towards Combating Fake News over Online Social Media

Kuai Xu[1(✉)], Feng Wang[1], Haiyan Wang[1], and Bo Yang[2]

[1] Arizona State University, Glendale, USA
{kuai.xu,fwang25,haiyan.wang}@asu.edu
[2] Jiangxi University of Finance and Economics, Nanchang, China
jxncluoming@hotmail.com

Abstract. Fake news has recently leveraged the power and scale of online social media to effectively spread misinformation which not only erodes the trust of people on traditional presses and journalisms, but also manipulates the opinions and sentiments of the public. Detecting fake news is a daunting challenge due to subtle difference between real and fake news. As a first step of fighting with fake news, this paper characterizes hundreds of popular fake and real news measured by shares, reactions, and comments on Facebook from two perspectives: Web sites and content. Our site analysis reveals that the Web sites of the fake and real news publishers exhibit diverse registration behaviors and registration timing. In addition, fake news tends to disappear from the Web after a certain amount of time. The content characterizations on the fake and real news corpus suggest that simply applying term frequency - inverse document frequency (tf-idf) and Latent Dirichlet allocation (LDA) topic modeling is inefficient in detecting fake news, while exploring document similarity with the term and word vectors is a very promising direction for predicting fake and real news. To the best of our knowledge, this is the first effort to systematically study the Web sites and content characteristics of fake and real news, which will provide key insights for effectively detecting fake news on social media.

1 Introduction

The last decade has witnessed the rapid growth and success of online social networks, which has disrupted traditional media by fundamentally changing how, who, when, and where on the distribution of the latest news stories. Unlike traditional newspapers or magazines, anyone can spread any information at any time on many open and always-on social media platforms without real-world authentications and accountability, which has resulted in unprecedented circulation and spreadings of fake news, social spams, and misinformation [1–5].

Driven by the political or financial incentives, the creators of fake news generate and submit these well-crafted news stories on online social media, and subsequently recruit social bots or paid spammers to push the news to a certain popularity [6–8]. The recommendation and ranking algorithms on social media,

S. Chellappan et al. (Eds.): WASA 2018, LNCS 10874, pp. 521–531, 2018.
https://doi.org/10.1007/978-3-319-94268-1_43

if failed to immediately detect such fake news, likely surface such news to many other innocent users who are interested in the similar topics and content of the news, thus leading to a viral spreading process on social media. These rising social spams [9], click baits [10] and fake news [1], mixed with real news and credible content, create challenges and difficulties for regular Internet users to distinguish credible and fake content.

Towards effectively detecting, characterizing, and modeling Internet fake news on online social media [11], this paper proposes a new framework which systematically characterizes the Web sites and reputations of the publishers of the fake and real news articles, analyzes the similarity and dissimilarity of the fake and real news on the most important terms of the news articles via tf-idf and LDA topic modeling, as well as explores document similarity analysis via Jaccard similarity measures between fake, real and hybrid news articles.

The contributions of this paper are three-fold:

- We systematically characterize the Web sites and reputations of the publishers of the fake and real news articles on their registration patterns, Web site ages, and the probabilities of news disappearance from the Internet.
- We analyze the similarity and dissimilarity of the fake and real news on the most important terms of the news articles via term frequency - inverse document frequency (tf-idf) and Latent Dirichlet allocation (LDA) topic modeling.
- We explore document similarity between fake, real or hybrid news articles via Jaccard similarity to distinguish, classify and predict fake and real news.

The remainder of this paper is organized as follows. Section 2 describes the background of the fake news problem over online social media and describes data-sets used in this study. Section 3 characterizes the Web sites and reputations of the publishers of the fake and real news articles, while Sect. 4 focuses on analyzing the similarity and dissimilarity of the fake and real news on the most important terms of the news articles. In Sect. 5, we show the promising direction of leveraging document similarity to distinguish fake and real news by measuring their document similarity. Section 6 summarizes related work in detecting and analyzing fake news and highlights the difference between this effort with existing studies. Finally, Sect. 7 concludes this paper and outlines our future work.

2 Background and Data-Sets

As online social media such as Facebook and Twitter continue to play a central role in disseminate news articles to billions of Internet users, fake and real news share the same distribution channels and diffusion networks. The creators of fake news, motivated by a variety of reasons including financial benefits and political campaigns, are very innovative in writing the news stories and attractive titles that convince thousands of regular people to read, like, comment, forward. Such high engagement in a short time period can make the news go viral with little challenges or doubts on authenticity, verification or fact checking.

In this paper we explore the research data shared from a recent study in [12]. The data consists three data-sets, each of which includes hundreds of fake and real news stories over a 3-month time-span from dozens of fake news sites as well as well-respected major news outlets including New York Times, Washington Post, NBC News, USA Today, and Wall Street Journal. These three data-sets are referred to as *dataset 1*, *dataset 2*, and *dataset 3* throughout the rest of this paper. For each fake or real news article, the data includes the tile of the story, the Web URL of the news story, the publisher of the news and the total engagement, measured by the total number of shares, likes, comments, and other reactions of the news received on Facebook.

3 Characterizing Fake and Real News

In this section, we study a variety of subjective features on the publishers of real and fake news such as the registration behaviors of publishers' Web sites, the sites ages of the publishers, and the probability of the news disappearance on the Internet.

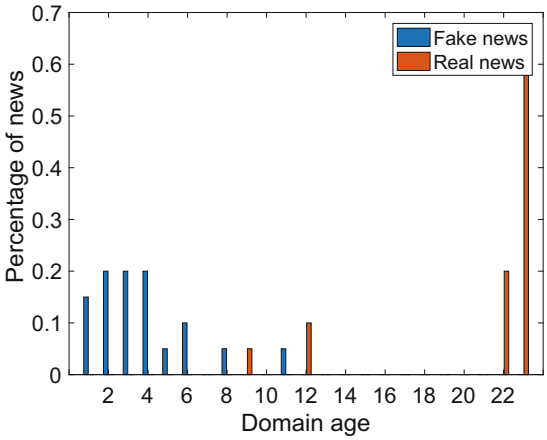
3.1 Web Site Registration Behavior of the Publishers

The real or fake news publishers typically have to go through the domain registration process, which allows anonymous domain registrar to serve as a proxy for publishers who prefer to hind their identities. If a publisher chooses to remain anonymous, the Internet whois database will show the proxy, e.g., Domains By Proxy, LLC as the registration organization. Most popular and well known newspaper typically choose to use the real organization name during the registration process. For example, the registration organization for wsj.com is Dow Jones & Company, Inc, which owns Wall Street Journal newspaper.

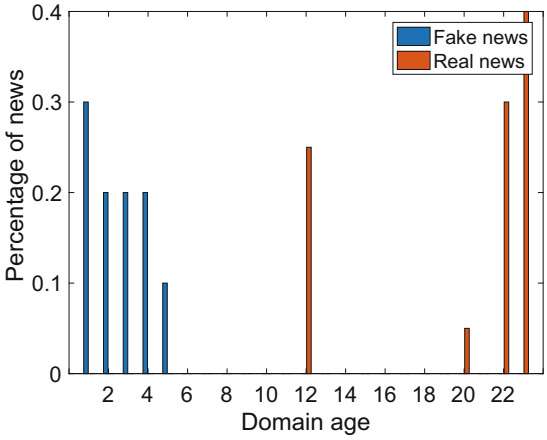
Our findings show that the majority of the fake news publishers register their Web sites via proxy services to remain anonymous, while all the real news publisher use their real identifies during the domain registration process. As shown in Table 1, over 78% of the domains publishing fake news are registered via proxy services to hide their true identities of the domain owners, while less than 2% of the domains publishing real news are registered in such a fashion. Thus we believe such patterns can become a powerful feature for machine learning models to distinguish fake and real news.

Table 1. Domain registration with proxy service for hiding domain owners' identify

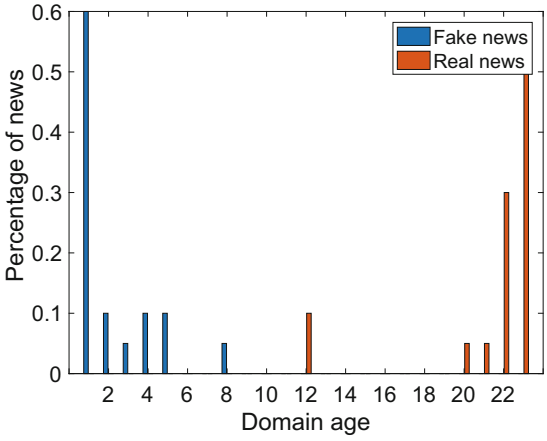
Category	Dataset 1	Dataset 2	Dataset 3	Average
Fake news	90%	65%	80%	78.3%
Real news	5%	0%	0%	1.7%



(a) dataset 1



(b) dataset 2



(c) dataset 3

Fig. 1. The Web site age distribution of the fake news publishers vs. real news publishers

3.2 Internet Site Ages of the Publishers

Beside the domain registration behavior, we also study the ages of the domains for the fake and real news in three data-sets. For each data-set, we characterize the domain age distribution for the fake and real news, respectively. As illustrated in Fig. 1, all data-sets exhibit consistent observations which reveal the very short domain ages for fake news, and the long domain ages for real news. This result is not surprising in that the credible newspapers registered their domains in early 1990s when Internet and Web start to attract attentions, while the fake news driven publishers often temporarily register the sites for the purpose of spreading fake news in a very short time of period.

3.3 Probability of News Disappearance

Credible news agency tends to maintain high quality sites that keep the published news for a long time. However, fake news sites often take the news offline after achieving the short-term goals of misleading the readers. Our analysis on the fake and real news corpus confirm such common practice.

Table 2. Page not found due to news disappearing.

Category	02/16–04/16	05/16–07/16	08/16–10/16	Average
Fake news	40%	70%	55%	55%
Real news	0%	0%	0%	0%

As shown in Table 2, the three data-sets of fake news corpus exhibit consistent news disappearing patterns, while the real news corpus has zero news that are taken offline. Thus we believe news disappearance could become a valuable feature for differentiating or modeling fake and real news.

In summary, our preliminary results on these popular fake and real news reveal substantial difference between fake and real news on the quality of the new pages, as well as the reputations of the publishing domains reflected by the domain ages as well as the interesting usage of the registration proxies.

4 Topics and Content of Fake and Real News

In this section, we first identify the most important topics of each fake or real news article via tf-idf analysis [13]. Subsequently, we explore the probabilistic LDA topic model to understand the difference or similarity of topics between labeled fake and real news.

Table 3. Top terms ranked by tf-idf values in fake, real and hybrid news corpus

Fake news corpus	Real news corpus	Hybrid news corpus
Violence	Trump	Comey
Trade	Nation	Transgender
Palin	Melania	Putin
Nuclear	Intelligence	Fraud
Mexico	FBI	Obama
Isis	Corrupt	Nuclear
Goods	Conway	Corrupt
Country	Conservative	Melania
Comey	Hillary	Isis
Canada	Wikileaks	Trump

4.1 Important Topics Identifications via tf-idf Analysis

tf-idf (term frequency - inverse document frequency) is a widely used statistical technique for extracting the most important term, t, or word, w, of a document in a document corpus, D. The tf-idf value of a term t, in the document d, tf-idf(t, d), is a product of the term frequency $tf(t, d)$ and the inverse document frequency $idf(t, D)$, i.e.,

$$tf\text{-}idf(t, d) = tf(t, d) * idf(t, D). \tag{1}$$

Table 3 shows that the most important terms extracted from fake, real and hybrid news corpus via tf-idf analysis are quite similar, thus relying on these terms alone is inefficient for detecting fake news.

4.2 Latent Dirichlet Allocation Topic Modeling

Topic models are widely used for understanding the content of documents based on word usage. In this paper, we explore Latent Dirichlet Allocation (LDA), a probabilistic topic model, to capture the topics of fake, real and hybrid news corpus respectively. The goal of LDA topic modeling on fake and real news is to understand the difference or similarity of topics between labeled fake and real news.

Tables 4, 5 and 6 illustrate the three topics with 5 most frequent terms for each corpus. As shown in these tables, the fake and real news share strong similarity in the overall topics, thus topic model alone is not an effective approach to detect or differentiate fake or real news in the real world.

Table 4. LDA topics for fake news corpus

Top 1	Top 2	Top 3
Trump	President	Candidate
Clinton	News	Black
Comey	State	Will
Hillary	American	One
Donald	Time	Said

Table 5. LDA topics for real news corpus

Top 1	Top 2	Top 3
Trump	Facebook	Republican
Clinton	Romney	Democratic
Donald	People	Authoritarian
President	Source	Politician
People	See	Party

Table 6. LDA topics for hybrid fake and real news corpus

Top 1	Top 2	Top 3
People	Trump	Trump
Authoritarian	Clinton	Donald
Politician	Republican	People
Party	Democratic	Make
American	President	Will

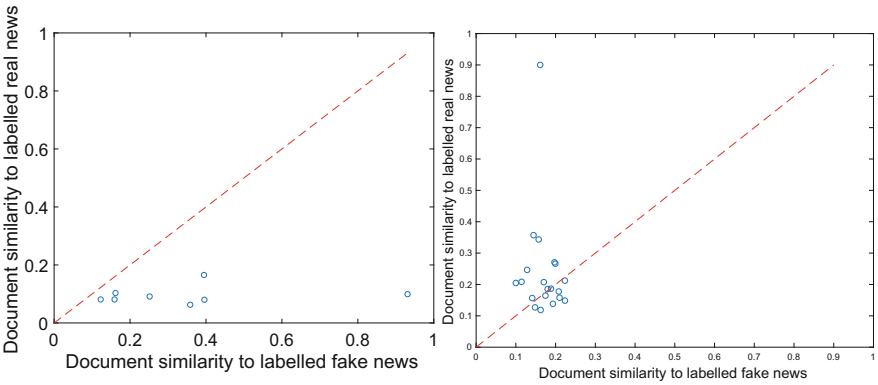
5 Document Similarity Analysis for News Predictions

As the LDA topics are inefficient to distinguish fake and real news, our followup analysis to explore document similarity between fake, real or hybrid news articles. First, we randomly divide the labeled fake and real news into training sets and test sets with a spit ratio of 67% for training and 33% for test.

For each fake or real news $\mathtt{n}$ in the test corpus, we measure the document similarity between $\mathtt{n}$ and every news in the fake news training set $\mathcal{F}$ and the real news training set $\mathcal{R}$. In particularly, we calculate Jaccard similarity $J(doc_1, doc_2)$, a widely used similarity measure between two documents doc_1 and doc_2 with the following equation

$$J(doc_1, doc_2) = \frac{doc_1 \cap doc_2}{doc_1 \cup doc_2}, \tag{2}$$

where doc_1 and doc_2 are represented with the vectors, typically sparse, of terms in the documents.



(a) Prediction on fake news (b) Prediction on real news

Fig. 2. The prediction on fake and real news based on labeled fake and real news corpus

Figure 2(a) shows the fake news in the test set have a much higher average document similarity with the news in the fake news training set $\mathcal{F}$ than with those in $\mathcal{R}$. However, Fig. 2(b) shows the real news in the test set have surprising similar document similarity with the news in the real news training set $\mathcal{R}$ and with those in $\mathcal{F}$. Thus as shown in Fig. 2(a), document similarity can potentially detect fake news. One of our future work is to systematically quantify the precision and recall of detecting both fake and real news in a large-scale news corpus.

In summary, our preliminary analysis on the topics and content of fake and real news reveals that it is very challenging to simply exploring the tf-idf and LDA topic modeling to effectively detecting fake news. However, our study also shows the promising aspect of leveraging document similarity to distinguish fake and real news by measuring the document similarity of the news under tests with the known fake and real news corpus.

6 Related Work

In recent years, several algorithms [1,2,8,14–20] have been proposed to detect the dissemination of information, misinformation or fake news. For example [2] exploits the diffusion patterns of information to automatically classify and detect misinformation, hoaxes or fake news, while [8] proposes linguistic approaches, network approaches, and a hybrid approach combining linguistic cues and network-based behavior insights for identifying fake news. In addition, [1] reviews the data mining literature on characterizing and detecting fake news on social media.

Similarly, [14] proposes a SVM-based algorithm for predicting misleading news with predictive features such as absurdity, grammar, punctuation, humor, and negative affect, and [15] uses logistic regression to distinguish credible news

from fake news based on n-gram linguistic, embedding, capitalization, punctuation, pronoun use, sentiment polarity features. A recent effort in [16] formulates the fake news mitigation as the problem of optimal point process intervention in a network, and combines reinforcement learning with a point process network activity model for mitigating fake news in social networks.

In addition, [21] classifies the task of fake news detection into three different types: serious fabrications, large-scale hoaxes, and humorous fakes, and discusses the challenges of detecting each type of fake news. To address the lack of labeled data-sets for fake news detection, [22] introduces a real-world data-set consisting of 12,836 statements with real or fake labels. In [17], the authors locates the hidden paid posters who get paid for posting fake news via modeling the behavioral patterns of paid posters.

7 Conclusions and Future Work

As fake news and disinformation continue to grow in online social media, it becomes imperative to gain in-depth understanding on the characteristics of fake and real news articles for better detecting and filtering fake news. Towards effectively combating fake news, this paper characterizes hundreds of very popular fake and real news from a variety of perspectives including the domains and reputations of the news publishers, as well as the important terms of each news and their word embeddings. Our analysis shows that the fake and real news exhibit substantial differences on the reputations and domain characteristics of the news publishers. On the other hands, the difference on the topics and word embedding shows little or subtle difference between fake and real news. Our future work is centered on exploring the word2vec algorithm [23], a computationally-efficient predictive model based on neural networks for learning the representations of words in the high-dimensional vector space, to learn word embedding of the important words or terms discovered via the aforementioned tf-idf analysis. Rather than comparing the few important words of each new article, word2vec will allow us to compare the entire vector and embeddings of each word for broadly capturing the similarity and dissimilarity of the content in the fake or real news.

Acknowledgements. This work was supported in part by National Science Foundation Algorithms for Threat Detection (ATD) Program under the grant DMS #1737861.

References

1. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. **19**(1), 22–36 (2017)
2. Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S., de Alfaro, L.: Automated fake news detection in social networks. Technical report UCSC-SOE-17-05, School of Engineering, University of California, Santa Cruz (2017)
3. Mitchell Waldrop, M.: News feature: the genuine problem of fake news. Proc. Natl. Acad. Sci. **114**(48), 12631–12634 (2017)

4. He, Z., Cai, Z., Wang, X.: Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks. In: Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS), June 2015
5. He, Z., Cai, Z., Yu, J., Wang, X., Sun, Y., Li, Y.: Cost-efficient strategies for restraining rumor spreading in mobile social networks. IEEE Trans. Veh. Technol. **66**(3), 2789–2800 (2017)
6. Shao, C., Ciampaglia, G.L., Varol, O., Flammini, A., Menczer, F.: The spread of fake news by social bots. arXiv.org, July 2017
7. Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X., Vlachos, A.: Fake news stance detection using stacked ensemble of classifiers. In: Proceedings of EMNLP Workshop: Natural Language Processing meets Journalism, September 2017
8. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. In: Proc. Assoc. Inf. Sci. Technol **52**(1) (2015)
9. Markines, B., Cattuto, C., Menczer, F.: Social spam detection. In: Proceedings of International Workshop on Adversarial Information Retrieval on the Web, April 2009
10. Chen, Y., Conroy, N.J., Rubin, V.L.: Misleading online content: recognizing clickbait as false news. In: Proceedings of ACM on Workshop on Multimodal Deception Detection, November 2015
11. Ruchansky, N., Seo, S., Liu, Y.: CSI: a hybrid deep model for fake news detection. In: Proceedings of ACM on Conference on Information and Knowledge Management, November 2017
12. This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook. https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook
13. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York City (1983)
14. Rubin, V.L., Conroy, N.J., Chen, Y., Cornwell, S.: Fake news or truth? Using satirical cues to detect potentially misleading news. In: Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 2016
15. Hardalov, M., Koychev, I., Nakov, P.: In search of credible news. In: Dichev, C., Agre, G. (eds.) AIMSA 2016. LNCS (LNAI), vol. 9883, pp. 172–180. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44748-3_17
16. Farajtabar, M., Yang, J., Ye, X., Xu, H., Trivedi, R., Khalil, E., Li, S., Song, L., Zha, H.: Fake news mitigation via point process based intervention. In: Proceedings of International Conference in Machine Learning (ICML) (2017)
17. Chen, C., Wu, K., Srinivasan, V., Zhang, X.: Battling the Internet water army: detection of hidden paid posters. In: Proceedings of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), August 2013
18. Wang, F., Wang, H., Xu, K., Wu, J., Jia, X.: Characterizing information diffusion in online social networks with linear diffusive model. In: Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS), Philadelphia, PA, July 2013
19. Wang, F., Wang, H., Xu, K.: Diffusive logistic model towards predicting information diffusion in online social networks. In: Proceedings of IEEE ICDCS Workshop on Peer-to-Peer Computing and Online Social Networking (HOTPOST), Macao, China, June 2012

20. Dai, G., Ma, R., Wang, H., Wang, F., Xu, K.: Partial differential equations with robin boundary condition in online social networks. Discret. Contin. Dyn. Syst. - Ser. B (DCDS-B) **20**(6), 1609–1624 (2015)
21. Rubin, V.L., Chen, Y., Conroy, N.J.: Deception detection for news: three types of fakes. In: Proceedings of the Association for Information Science and Technology, February 2015
22. Wang, W.Y.: Liar, liar pants on fire: a new benchmark dataset for fake news detection. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), July 2017
23. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of International Conference on Learning Representations (2013)