Detection of Fraudulent Tweets: An Empirical Investigation Using Network Analysis and Deep Learning Technique

Jaewan Lim
Department of Business Infomration Systems and Operation Management
University of North Carolina at Charlotte
Charlotte, U.S.
jlim13@uncc.edu

Zhihui Liu
Department of Computer Science
University of North Carolina at Charlotte
Charlotte, U.S.
zliu15@uncc.edu

Lina Zhou
Department of Business Infomration Systems and Operation Management
University of North Carolina at Charlotte
Charlotte, U.S.
lzhou8@uncc.edu

Abstract— Social media has become a powerful and efficient platform for information diffusion. The increasing pervasiveness of social media use, however, has brought about the problems of fraudulent accounts that are intended to diffuse misinformation or malicious contents. Twitter recently released comprehensive archives of fraudulent tweets that are possibly connected to a propaganda effort of Internet Research Agency (IRA) on the 2016 U.S. presidential election. To understand information diffusion in fraudulent networks, we analyze structural properties of the IRA retweet network, and develop deep neural network models to detect fraudulent tweets. The structure analysis reveals key characteristics of the fraudulent network. The experiment results demonstrate the superior performance of the deep learning technique to a traditional classification method in detecting fraudulent tweets. The findings have potential implications for curbing online misinformation.

Keywords—fraudulent tweets, network analysis, deep learning, community detection

I. Introduction

Social media increasingly serves as a platform for expressing and sharing political opinions. Along with the growth of social media, there appear many unintended problems. One of them is a fraudulent account, which refers to the account posting inflammatory, extraneous, or off-topic messages in an online community [1]. This kind of accounts has been used to disseminate spiteful contents [2] and anti-social news using an aggressive language, which lures social media users into fruitless argumentation [3]. Therefore, detecting fraudulent accounts can help curb the diffusion of malicious information in social media.

We choose Twitter as a platform to investigate the problem of fraudulent accounts in this study. Twitter does not require any identity verification [4]. Consequently, it may become a breeding ground for fake accounts [5]. Twitter recently released comprehensive archives of the tweets posted around 2016 U.S. election day that are possibly connected to a propaganda effort by Internet Research Agency (IRA), an alleged Russian government-linked company located in Saint Petersburg [6]. Twitter believes that these tweets and accounts are potentially state-backed in an attempt to influence the U.S. election with an unfavorable purpose [7]. Although several studies [7, 11] have explored the dataset to detect fraudulent accounts, they are limited in two aspects. First, there has been little research into the characteristics of IRA retweet network. Fraudulent accounts are likely strongly connected among themselves as a small-world network to exert influences on others' political opinions. An analysis of the retweet network can help determine whether there exists such a strong connectivity pattern. Second, despite the promising results that deep learning techniques have produced in natural language processing applications, they have rarely been used to detect fraudulent tweets.

To address the above limitations, we develop methods for the detection of fraudulent accounts and tweet texts that combine social network analysis and deep learning technique. This study makes two-fold research contributions: 1) we identify the characteristics of IRA retweet network and communities from the network by exploring the graph properties of the network, and 2) to the best of our knowledge, this is the first study that applies a deep learning model to the detection of fraudulent tweets.

II. Related Work

Fraudulent Twitter accounts and tweet detection have been approached from the perspectives of network analysis and text analysis.

A. Network analysis

An analysis of 485,721 Twitter accounts and 14,401,157 tweets reveals that malicious accounts are likely to be connected by following one another [8]. Another study examines the connectivity and temporal behavior of honest and fraudulent accounts by analyzing the retweet networks

[9, 10]. The study confirms a strong connectivity pattern in fraudulent user networks.

B. Text analysis

Binary classifiers have been built to identify fraudulent tweets using the dataset that NBC collected from the fraudulent accounts that Twitter had removed [11]. Based on a comparison of 27 combinations of classification models and parameters such as Naïve Bayes, SVM, and C4.5, SVM with unigrams achieved the best performance with a precision of 84.9% and a recall of 84.4%. Existing machine learning techniques used in analyzing the fraudulent tweets heavily rely on feature engineering and overlook the state-of-the-art techniques for text classification.

In this study, we combine network analysis and deep learning based text analysis to gain a more complete understanding of fraudulent accounts.

III. Method

A. Dataset and preparation

The original dataset called Twitter Elections Integrity Dataset consists of tweets generated from 3,613 IRA accounts [6]. We first extract all English tweets along with their metadata from the dataset, which result in 2,997,181 tweets including 1,082,867 retweets. They are treated as fraudulent tweets in this study. The metadata contains information such as user account, retweeted account, whether it is a retweet, and hashtag, which is used to construct an IRA retweet network. The user account metadata and retweeted account are different from each other in that the user account contains IRA accounts, while retweeted account involves both IRA and non-IRA accounts. We treat fraudulent tweet detection as a binary classification problem. To support the classifier training and testing, we create a dataset of authentic tweets by collecting tweets posted on the 2016 U.S. presidential election day (November 8) [15], which consists of 338,331 tweets.

B. Network analysis

To represent the IRA network, we use *node* to denote user (account), and *edge* to denote retweet relationship between different users. Focusing on information diffusion in the fraudulent network, we first examine the structure of the retweet network by applying community detection algorithms. Specifically, the Louvain method is a heuristic method for identifying communities in large networks based on modularity optimization [16]. To further characterize the network connectivity and explore the distribution of edges among nodes, we measure node degree distribution and plot its log-log transformation.

C. Deep learning based detection model

The tweets first go through preprocessing steps including word tokenization [13] and word embedding training [12]. Word embedding allows words with similar meaning to have similar representations. The deep learning model consists of four layers: input layer, max pooling layer, denser layer, and an output layer. For the output layer, we apply binary cross entropy with Adam optimization [14] as the loss function. We set the batch size to be 128 and the number of epochs to be 50. Given that the fraudulent and authentic datasets are highly unbalanced, we build classifiers using two different settings: 1) the whole datasets, and 2) balanced datasets consisting of the authentic tweets and a subset of IRA tweets selected based on a timeframe (posted between October 8, 2016 and December 8, 2016) close to that of the authentic data, resulting in 215,176 IRA tweets.

D. Evaluation setting

We randomly split the data into training, validation and testing subsets at a ratio of 60:20:20 through stratified sampling. We choose the balanced bagging classifier as the baseline model because it is capable of handling unbalanced datasets. The metrics for classification performance include accuracy, precision, recall, and F1-score.

IV. Result

A. Properties of IRA retweet network

The IRA retweet network, as a directed graph, consists of 131,961 nodes and 442,607 edges. The modularity of the network is 0.64, and mean clustering coefficient 0.14. The community detection results in a total of 84 communities. Among them, 22 communities have the size of over 10 (nodes), top-4 have the size of over 10,000 nodes each, and the largest one has the size of 44,334 (accounting for more than 33.6% of the nodes). The results suggest that the fraudulent network has dense connections between the nodes within communities, but sparse connections across communities. There is one giant community in the network.

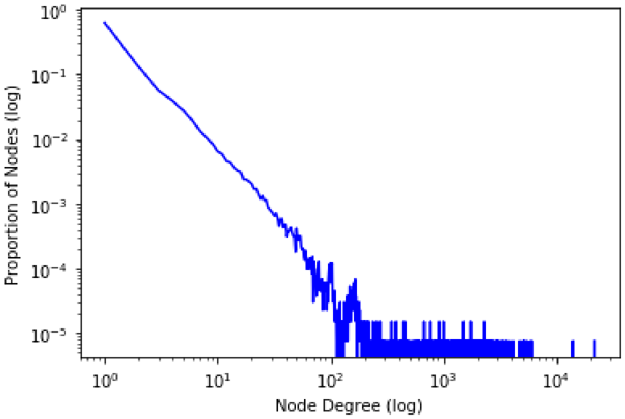

Fig. 1. Degree distribution of the IRA retweet network

Fig. 1 shows the log-log transformation plot of degree distribution in the fraudulent retweet network. The network follows a power-law distribution (α=2.18). It demonstrates the Mathew effect of accumulated advantage, that is, nodes with higher degrees in the network tend to attract more edges. The most active user made 117,168 retweet attempts

(10.8%). The user whose tweets were retweeted most accounts for 8,546 retweets (0.7%), and the most popular hashtags in those retweets include #blackquotes, #FollowFriday, #uncensored, #BrotherDarkness, etc.

B. Detecting fraudulent tweets

The loss value decreases from 0.38 to 0.25 after training our deep learning model for 50 epochs. The performances of fraudulent tweet detection models are reported in Table I. The table shows that the deep learning models outperform the baseline models in all evaluation metrics across both evaluation settings. For instance, the deep learning model achieves an accuracy of about 95.58% and an F1-score of 97.60% on the whole dataset, which are much higher than the baseline model (accuracy = 81.17%, F1-score = 88.82%). In addition, the performances on the entire dataset are superior to those on the balanced subset for both of the deep learning and baseline models. This observation suggests that a larger dataset contributes to improved performances in detecting fraudulent tweets. Compared with the deep learning model, the performance gain from employing the larger dataset is greater for the baseline model.

TABLE I. PERFORMANCE OF THE MODELS

Model	**Measure**	**Whole dataset**	**Balanced Subset**
Baseline model	Precision	96.24%	70.29%
	Recall	82.47%	66.67%
	F1-score	88.82%	68.43%
	Accuracy	81.17%	76.26%
Deep learning model	Precision	96.26%	87.51%
	Recall	98.99%	87.41%
	F1-score	97.60%	87.46%
	Accuracy	95.58%	84.80%

We also compare the hashtags (i.e., user-generated tags to represent the topics of tweets) between fraudulent and authentic tweets posted on the 2016 U.S. presidential election day. To make the comparison on a fairground, we randomly selected 6,103 tweets from the authentic dataset to match the size of tweets from the fraudulent dataset. Based on the results of frequency analyses, the top 5 hashtags in authentic tweets include #election2016, #ElectionDay, #vote, #ImWithHer and #ElectionNight; and the top 5 in fraudulent tweets #TrumpForPresident, #ThingsPeopleOnTwitterLike, #ElectionDay, #HillaryForPrison2016 and #MAGA (make America great again). The results reveal that the hashtags used in fraudulent tweets can involve political flaming and/or aggressive languages despite that they cover general topics on the presidential election as do authentic tweets.

V. CONCLUSION

This study analyzes fraudulent tweets at two levels using different analysis methods: 1) characterizing the fraudulent network at the user level by conducting a structural analysis of the fraudulent network, and 2) detecting fraudulent tweets at the tweet level by developing deep neural network models. The structure analysis reveals that the fraudulent network is mainly comprised of a small number of densely connected components. The experiment results demonstrate the superior performance of the deep learning technique to a traditional classification method in detecting fraudulent tweets. A further comparison of hashtags reveals that fraudulent tweets can involve political flaming and/or aggressive languages despite that they cover similar topics to authentic tweets This research can be continued in a number of directions such as examining temporal patterns of the fraudulent network structure and extending advanced deep learning techniques to build detection models.

VI. ACKNOWLEDGMENT

This work is supported in part by grant awards from the National Science Foundation (Award #s: CNS 1704800 and SES 1527684). Any opinions, findings or recommendations expressed here are those of the authors and are not necessarily those of the sponsor of this research.

VII. REFERENCES

[1] Galán-García, P., Puerta, J. G. D. L., Gómez, C. L., Santos, I., & Bringas, P. G. (2016). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic Journal of the IGPL*, 24(1), 42-53.

[2] Flores, Marcel, and Aleksandar Kuzmanovic. "Searching for spam: detecting fraudulent accounts via web search." International Conference on Passive and Active Network Measurement. Springer, Berlin, 2013.

[3] Fornacciari, P., Mordonini, M., Poggi, A., Sani, L., & Tomaiuolo, M. (2018). A holistic system for troll detection on Twitter. *Computers in Human Behavior*, 89, 258-268.

[4] Galán-García, Patxi, et al. "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying." Logic Journal of the IGPL 24.1 (2016): 42-53.

[5] Mihaylov, Todor, et al. "The dark side of news community forums: Opinion manipulation trolls." Internet Research 28.5 (2018): 1292-1312.

[6] Twitter. 'We're focused on serving the public conversation', 2018. [Online]. Available: https://about.twitter.com/en_us/values/elections-integrity.html [Accessed: 07- Mar- 2019].

[7] Badawy, Adam, et al. "Characterizing the 2016 Russian IRA Influence Campaign." *arXiv preprint arXiv*:1812.01997 (2018).

[8] Yang, Chao, et al. "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[9] Giatsoglou, Maria, et al. "Retweeting activity on twitter: Signs of deception." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2015.

[10] Stewart, L. G., Arif, A., & Starbird, K. (2018, February). Examining trolls and polarization with a retweet network. In *Proc. ACM WSDM, Workshop on Misinformation and Misbehavior Mining on the Web*.

[11] Griffin, Christopher, and Brady Bickel. "Unsupervised Machine Learning of Open Source Russian Twitter Data Reveals Global Scope and Operational Characteristics." arXiv preprint arXiv:1810.01466 (2018).

[12] Yang, Xiao, Craig Macdonald, and Iadh Ounis. "Using word embeddings in twitter election classification." *Information Retrieval Journal* 21.2-3 (2018): 183-207.

[13] Keras Documentation from https://keras.io/preprocessing/text/

[14] Drozdzal, Michal, et al. "The importance of skip connections in biomedical image segmentation." *Deep Learning and Data Labeling for Medical Applications.* Springer, Cham, 2016. 179-187.

[15] Ed King, 'Election Day Tweets', 2017. [Online]. Available: https://www.kaggle.com/kinguistics/election-day-tweets [Accessed: 10-Apr- 2019].

[16] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.