Is the News Deceptive? Fake News Detection using Topic Authenticity

Aviad Elyashar, Jorge Bendahan, and Rami Puzis
Telekom Innovation Laboratories and Department of Software and Information Systems Engineering
Ben-Gurion University of the Negev, Beer-Sheva, Israel
Email: aviade@post.bgu.ac.il, jorgeaug@post.bgu.ac.il, puzis@bgu.ac.il

***Abstract*—In this paper, we propose an approach for the detection of fake news in online social media (OSM). The approach is based on the authenticity of online discussions published by *fake news promoters* and *legitimate accounts*. Authenticity is quantified using a machine learning (ML) classifier that distinguishes between *fake news promoters* and *legitimate accounts*. In addition, we introduce novel link prediction features that were shown to be useful for classification. A description of the processes used to divide the dataset into categories representing topics or online discussions and measuring the authenticity of online discussions is provided. We also discuss new data collection methods for OSM, describe the process used to retrieve accounts and their posts in order to train traditional ML classifiers, and present guidelines for manually labeling accounts. The proposed approach is demonstrated using a Twitter pro-ISIS fanboy dataset provided by Kaggle. Our results show that the method can determine a topic's authenticity from *fake news promoters*, and *legitimate accounts*. Thus, the suggested approach is effective for discriminating between topics that were strongly promoted by *fake news promoters* and those that attracted authentic public interest.**

Keywords–Fake News Detection; Link Prediction; Data Collection; Topic Authenticity.

I. INTRODUCTION

Traditionally, television and newspapers were the kinds of media devices used to inform people about the news and other topics of interest. However, in recent decades new vehicles for news delivery have been introduced, such as computers and mobile devices. Moreover, the popularity of viewing news on these devices has grown due to the easy access of online news using smart devices, and content generators, that provide users with a steady stream of personalized news, derived from a wide variety of news sources. As a result, online news is rapidly replacing traditional media devices [1].

Although online news provides numerous benefits, this domain is also problematic. For example, the nature of online news publication has changed, to the point that traditional fact checking and vetting performed to prevent potential deception are sometimes absent or incomplete due to the flood of material from content generators [2]. The flood of unchecked news has contributed to the growing problem of fake news publication, which has been defined as particular news articles, which are intentionally deceptive and their publication and propagation [2].

There are dangers associated with the publication of deceptive news. In many cases, these news are published for spreading rumors, influence, and intentionally mislead people [3]. For example, nasty rumors about organizations, which are published by malicious users, can result in serious reputation damage [4].

In many cases, a *fake news promoter* takes over a specific online discussion and may have a strong influence on the other participants writing on the topic. In this study, we propose a method for detecting fake news in online social media (OSM) based on a machine learning (ML) classifier capable of distinguishing between *fake news promoters* and *legitimate accounts* participating in the same online discussion. The classifier is based on behavioral features (e.g., total number of retweets), network analysis features (e.g., co-citation closeness centrality), and link prediction features (e.g., max total friends in common posts graph from *fake news promoters*), which are used to compare OSM accounts participating in a particular online discussion to both confirmed *fake news promoters* and *legitimate accounts*. The classifier attempts to quantify the authenticity level of accounts, where *fake news promoters* and *legitimate accounts* are placed on opposite ends of the authenticity scale. We demonstrate that the distribution of accounts' authenticity is different in topics that are prone to OSM manipulation and in those topics that attract authentic public interest. In order to evaluate our method, we used the Twitter Propaganda dataset provided by Kaggle [5].

The contributions of this paper are:

- identifying link prediction-based features that are found useful for account classification. To the best of our knowledge, we are the first to identify link prediction features that address the account type classification;
- developing a novel method for data collection for cases in which the samples come from only one class. The method is based on topic detection and is useful for retrieving unlabeled samples with the same context;
- providing guidelines for manually labeling accounts with respect to fake news; and
- demonstrating the account authenticity distribution within OSM discussions;

The rest of this paper is organized as follows: In Section II, we review approaches for the detection of fake news, and abusers who are capable of spreading fake news within OSM, and review the concept of topic modeling. We describe the proposed method, including: a new method for data collection (Section III-A), the general guidelines for labeling accounts manually with respect to fake news (Section III-B), a novel classifier, which was found useful for the classification of *fake news promoters* and *legitimate accounts* (Section III-C) and the proposed topic authenticity estimation approach (Section III-D). Section III-E discusses ethical considerations, and we conclude the paper in Section IV with a summary and our plans for future work.

 ISBN: 978-1-61208-595-1

II. Related Work

In this section, we provide the necessary background information regarding the major issues focused on this study: fake news detection methods, methods for identifying abusers, and topic modeling in OSM.

A. Fake News Detection

The approaches for fake news detection can be divided into two categories as depicted by [2]: linguistic and network analysis. The linguistic approach is based on extraction of the content of deceptive messages, and analysis to associate language patterns related to deception. One of the simplest models based on this approach is the bag-of-words. The methods based on this model rely on shallow lexico-syntactic cues. Most of them are based on dictionary-based word counting using Linguistic Inquiry and Word Count (LIWC) [6], such as [7]. Others take advantage of ML techniques using simple lexico-syntactic patterns, such as n-grams and part-of-speech (POS) tags [8], or location-based words [9]. Recently, [10] developed hybrid convolutional neural network model which integrates metadata with text. They showed that hybrid model improves a text-only deep learning model. The shortcoming of this approach is that it relies solely on language: it does not differentiate between various meanings of words that look the same, and word counting does not categorize combinations of words or phrases that might imply different meanings [11].

Syntax analysis is a more sophisticated approach that attempts to detect fake news. Feng et al. [12] extended Ott et al.'s n-gram feature set [13] by incorporating deep syntax features derived from probability context free grammar (PCFG) parse trees. It was found to be useful for deception detection with about 90% accuracy. Later, others developed third party tools in order to automate this process, such as the Stanford Parser [14], and AutoSlog-TS syntax analyzer [15].

Semantic analysis is an other means of deception detection. [16] proposed a method that uses profile compatibility in order to differentiate between genuine and fake product reviews. They extended their previous n-gram plus syntax model [12] by mixing profile compatibility features. The researchers proposed the use of additional signals of truthfulness by characterizing the degree of compatibility between the personal experience described in a test review and a product profile derived from a collection of reference reviews about the same product. They showed that such additional signals of truthfulness significantly improve their model's performance. The shortcomings of this approach include the fact that it is restricted to the domain of application, and the limited ability to select alignment between features and descriptors that are dependent on the content of a profile.

Discourse analysis is another method that can help in the detection of fake news. Rubin et al. [17] used rhetorical structure theory (RST) that served as the analytic framework for the identification of systematic differences between deceptive and truthful stories in terms of their coherence and structure. They used a vector space model (VSM) that estimates each story's position in a multi-dimensional RST space with respect to its distance from truth and deceptive centers as a measure of the story level of deception and truthfulness. Recently, automatic tools for rhetorical classification have become available, but they have not yet been employed in the context of veracity assessment.

One more common method for the detection of deceptive cues is the use of ML classifiers, such as support vector machines (SVM) [18], and Naive Bayesian models [15].

As opposed to the linguistic approach there is the network analysis approach that provides aggregate deception measures, based on network, along with behavior features, such as message metadata or structured knowledge network queries. This approach is used in many applications, which involve real-time content, such as micro-blogging applications (e.g., Twitter) [2]. For example, [19] attempted to differentiate between human, bot, and cyborg users in terms of tweeting behavior.

B. Identification of Abusers

Several studies involving the identification of abusers have been conducted. [20] proposed a method for clustering accounts based on the similarity of the posted URL; they classified each cluster as either malicious or not by extracting behavioral and content features of each cluster. A method for the identification of crowdturfers on Twitter was presented by [21]. They extracted features that were related to account properties, activity patterns, and linguistic properties. [22] proposed a method for detection of artificially promoted objects, such as posts, pages, and hyper-links, as part of crowdturfing tasks. In this study, we extracted variety of features in order to classify an object as artificially promoted or not. [23] and [24] used supervised ML techniques for bot detection. [23] based their detection on sentiment analysis, social network analysis, posted content, and account property features. [24] presented BotOrNot, a bot identification platform that can be used through a Web user interface. They detected bots based on all the features like [23], including behavior features.

Recently, [25] identified hoaxes within Facebook based on the users who interacted with them rather than their content. [26] found evidence that socialbots play a key role in the spread of fake news. [27] proposed a method for estimating the authenticity of online discussions based on several similarity functions of OSM accounts participating in the online discussion. They found that the similarity function with the best performance across all the datasets was bag-of-words. This study is different from the current study in the goal (estimating the authenticity of online discussions within the domain of 'Virtual TV' versus fake news detection), the datasets, the method for account labeling, and the evaluation method (KNN with similarity function versus ML classifiers).

C. Topic Modeling

[28] introduced a technique called Latent Dirichlet Allocation (LDA) for identifying topic proportions in documents. k topics are defined for the entire corpus and each document in the corpus contains these topics in different proportions. LDA has been applied in a large number of areas, including text summarization, document search, and clustering. In LDA, topics are defined as probability distributions over a fixed set of terms within a corpus [28]. It means that a topic related to specific issue will include all of the words in the corpus, but words co-occurring together across multiple documents in the corpus revolving around this issue will have the highest probability in that topic's distribution, whereas words that appear less frequently will have a lower probability. In this study, we used LDA for identifying prominent online discussions or topics.

 ISBN: 978-1-61208-595-1

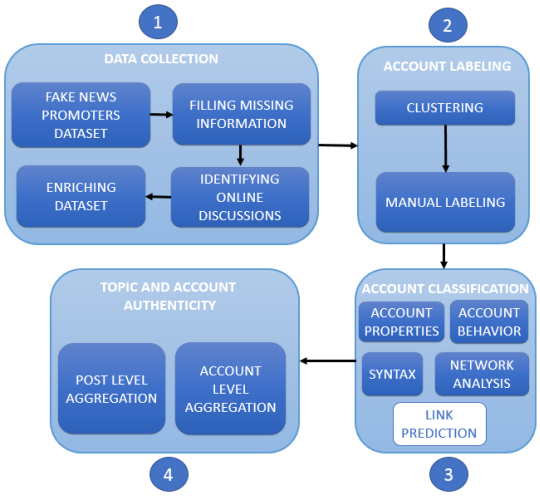

Figure 1. Estimation of account and topic authenticity.

III. PROPOSED METHOD

We propose an approach for detecting fake news based on estimating the prevalence of *fake news promoters* among the accounts that contributed to the given online discussion. In this section, we provide a comprehensive description of the proposed method, from data collection to authenticity of accounts and topics (depicted in Figure 1).

A. Data Collection

We used the Twitter Propaganda dataset which includes solely *fake news promoters*: 17,410 tweets published by 112 pro-ISIS fanboys from around the world from the November 2015 Paris terrorist attacks until May 2016. The dataset includes information about the account, such as account's full name, username, description, location, number of statuses, and number of followers. The information regarding the tweet included content, publication date, and time-stamp.

The first step in the data collection process includes using provider services in order to fill in the missing information regarding the given accounts. The more information we are able to obtain, the greater the number of helpful features that can be extracted. In this case, we used the Twitter REST API public service [29] to obtain the missing information about the *fake promoters*' accounts (e.g., number of friends). However, in the case of the Twitter Propaganda dataset, all of these accounts were suspended by Twitter administrators. It is important to note that at this point we only have samples of *fake news promoters*; we also need samples of *legitimate accounts* in order to use traditional binary supervised learning techniques (described in Section III-C). The next step is to use topic detection algorithms, such as LDA or latent semantic analysis (LSA) in order to identify online discussions. Each online discussion or topic is composed of several terms. We took the top ten terms in each topic by probability and retrieved 100 recent tweets that included these terms using the Twitter REST API. We decided on the top ten terms, because we believe that this number is sufficient to cover a topic. Additional terms are not necessarily provide posts, which are directly related to a given topic. Moreover, increasing the number from ten to a greater number would increase the amount of time spent crawling. Eventually, we collected 27,654 tweets that were published by 360 unlabeled accounts.

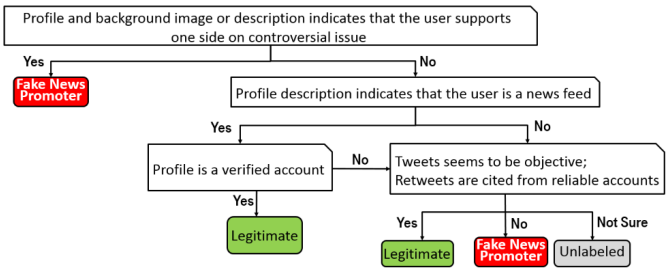

Figure 2. The manual labeling guidelines.

B. Account Labeling

In order to train a ML classifier or directly estimate the authenticity of the rest of the accounts in the dataset, we need samples of *legitimate accounts*. The overall approach is based on selection of the 'right' accounts for labeling, as well as strict unambiguous labeling guidelines. There are too many unlabeled accounts to manually label all of them. The simplest idea is to randomly choose unlabeled accounts and label them. However, by doing this we may inadvertently choose accounts from one type and not from the other. For this reason, we clustered the accounts, i.e., grouping a set of accounts in such a way that similar accounts will be in the same group. The clustering was carried out using features described in Section III-C. Selecting an equal number of samples from each cluster will preserve the highest variability among accounts.

Next, we manually inspect the unlabeled accounts and assign labels to them. We developed guidelines for the manual classification of an OSM account as a *fake news promoter* or *legitimate account*. The manual labeling process is presented in Figure 2 and described below.

(1) First, we look at self-descriptors, such as the profile, background image, and the description section of the account. A profile image is one of the most important personal attributes on OSM [30]. In many cases, it expresses the user's main motto or idea. In cases in which the profile image expresses support for one side of a controversial issue, we mark it as a potential *fake news promoter*. For example, an account in which the profile and background image contain extreme statements, such as 'Free Palestine' with clenched fists soaked in blood, would be marked as a potential *fake news promoter* due to its subjective opinions regarding the Israeli-Palestinian conflict. Likewise, extreme statements in the description like 'Evil Assad' with a profile image of the flag of Syrian Arab Republic (opposition) or Kurdish forces in Syria would be marked as a potential *fake news promoter* due to its subjective opinions regarding the Syrian civil war.

(2) Then, in cases in which an OSM account declares itself as a news feed or any other type of content aggregator, we check to determine whether it is a verified account. Tweets that are published by an authority are likely more reliable than tweets from an account with less credibility [31]. In cases, such as those mentioned above, we mark the account as a *legitimate account*. In those cases in which the news feed account remains unverified we look closer at the content published. If the tweets present an objective perception, or retweets from other reliable news feeds, we mark it as *legitimate*. If the tweets seem to be subjective or surrounding one side of a conflict, we mark it as a *fake news promoter*. If no clear decision can be

 ISBN: 978-1-61208-595-1

made regarding the source and nature of the news, the account remains *unlabeled*.

(3) Finally, we inspect the account's published content. If the majority of the posts published by the given account contain authentic content or the account retweets from a reliable account, we mark it as a *legitimate account*. In cases in which the posts are subjective and bias toward one side, we mark it as a *fake news promoter*. If no clear decision can be made regarding the source and nature of the posts, the account remains *unlabeled*.

We used the Committee of Experts approach in order to reach an agreement on the account labels. Three annotators (students) participated in the manual labeling process. They independently reviewed the same groups of unlabeled accounts and analyzed their Twitter profiles and posts. The annotators then assigned a label of either *legitimate account* or *fake news promoter* to each account. In case of full agreement among them, the label was set. It is important to reach full agreement in order to avoid biased labeling in case that the annotators belong to a specific cultural background or political community while the unlabeled accounts belong to the opposite community that protests against the annotators' community.

C. Account Classification

In this section, we describe the features used to classify Twitter accounts and present the results of our evaluation. In this study, we used features that were reported to perform well in the past [21], including a mixture of static account properties, behavioral features, and content / syntax related features. Moreover, we present the link-prediction features, which were found to be useful for classification.

Features based on account properties: screen name length. Other features (e.g., account age, friend-follower ratio, and others) were calculated, but later removed. Due to the suspension of the pro-ISIS fanboys, we could not complete these features.

Features based on account behavior: number of retweets, average retweets, and number of received retweets.

Features based on syntactic characteristics: average hashtags, average links, average user mentions, and average post length.

Features based on network analysis: we created two graphs: common posts, and co-citation.

- Common posts. OSM accounts that publish the same content might be part of a crowdturfing campaign [32]. *Common-posts* graphs emphasize which OSM accounts spread the same content across the OSM.
- Co-citation. There is significant evidence that one of the main malicious tasks is to spread hyperlinks across the OSM [33]. A *co-citation* graph shows which OSM accounts share the same hyperlinks, thereby discovering potential malicious activities.

The network analysis features were calculated as a Cartesian product between 1) the two graphs described above, 2) algorithms for calculating centrality in graphs, such as closeness centrality, clustering, and degree centrality, and 3) aggregation functions, such as mean, standard deviation, kurtosis, and skewness.

Features based on link prediction: these features are novel and they were identified during the current research and are not part of previous studies. These new features are used to assess the likelihood of each account to be a *fake news promoter*. First, we choose small number of known *fake news promoters* randomly. Afterwards, we create the common posts and co-citation graphs as described previously. Later, for each account, which is a node in the given graph, we calculate how much it has in common with other *fake news promoters*. For example, feature named 'Link prediction - max - total friends - common posts - fake news promoter' depicts the maximal number of friends a given account has in common posts graph with all the random *fake news promoters*. Actually, these features are a Cartesian product between 1) the two previously described graphs (common posts and co-citation), 2) the link prediction measures: Jaccard's coefficient, common neighbors, preferential attachment [34], Adamic-Adar index [35], total friends, transitive friends, opposite direction friends [36], and Bayesian promising [37], and 3) the aggregation functions: minimum, maximum, mean, median, skewness, and kurtosis.

In order to evaluate the predictive power of the extracted features, we applied information gain feature selection [38]. The most significant features are described in Table I. Among the top ten most significant features, six features are related to link prediction. These results suggest that it is important to consider both the topic affinity of an author and the behavioral properties of the account during classification.

TABLE I. TOP FEATURES ORDERED BY INFORMATION GAIN.

Rank	Feature	InfoGain
1	Average post length	0.2731
2	Average retweets	0.2352
3	Average user mentions	0.2231
4	Link prediction - max - total friends - common posts - fake news promoter	0.1894
5	Link prediction - median - total friends - common posts - fake news promoter	0.1894
6	Link prediction - min - total friends - common posts - fake news promoter	0.1894
7	Link prediction - mean - total friends - common posts - fake news promoter	0.1894
8	Average links	0.1062
9	Link prediction - kurtosis - total friends - common posts - fake news promoter	0.0672
10	Link prediction - skewness - total friends - common posts - fake news promoter	0.0672

We trained several ML classifiers (XGBoost showed the best results) in order to determine the differences between *fake news promoters* and *legitimate accounts*. Each classifier was trained with multiple sets of features having the highest information gain score. The performance of the classifiers was evaluated in terms of the area under ROC curve (AUC), accuracy, precision, and recall during internal 10-fold cross-validation. The results of the best classifier for each algorithm are summarized in Table II. We note that the best classifier was trained using XGBoost on all the features with an AUC of 0.935, accuracy of 0.89, and precision and recall of 0.913, and 0.923 respectively.

TABLE II. THE PERFORMANCE OF THE BEST CLASSIFIERS.

Algorithm	Num of features	AUC	Accuracy	Precision	Recall
XGBoost	All	0.935	0.89	0.913	0.923
Random Forest	All	0.93	0.85	0.87	0.911
Random Forest	20	0.919	0.87	0.88	0.941
XGBoost	20	0.89	0.83	0.85	0.912
AdaBoost	All	0.86	0.832	0.875	0.875
Decision Tree	All	0.83	0.84	0.89	0.87

 ISBN: 978-1-61208-595-1

D. Authenticity of Accounts and Topics

We estimate the authenticity of accounts using the confidence level provided by the best trained classifier. We define authenticity of the account x as the confidence of x being a *legitimate* account. The last step of the proposed approach is aggregating the authenticity of individual OSM accounts into authenticity of topics. We consider the following two aggregations:

a) Post level aggregation: In this case, every post is associated with the authenticity of its account. The authenticities are accumulated in terms of topic probabilities. The LDA-based topic detection determines the probability $T_{p,i}$ that post p belongs to topic i. Next, for each OSM account x and each topic i, we compute the average probability that x's posts belong to topic i.

$$topic\text{-}auth\text{-}1(i) = \sum_{p \in P} T_{p,i} \cdot acc\text{-}auth(A(p)) \qquad (1)$$

b) Author level aggregation: First, a set of authors of posts for a specific topic is determined. Then, the authenticities of the author accounts are aggregated. We define the set of accounts involved in a specific topic i as

$$D(i) = \{A(p) : T_{p,i} = MAX_j \{T_{p,j}\}\}.$$

Here, every post is associated with a single topic – the one it belongs to with the highest probability. An account is associated with a topic if at least one of its posts is associated with that topic. The account level authenticity of the topic i is then determined as follows:

$$topic\text{-}auth\text{-}2(i) = \sum_{x \in D(i)} acc\text{-}auth(x). \qquad (2)$$

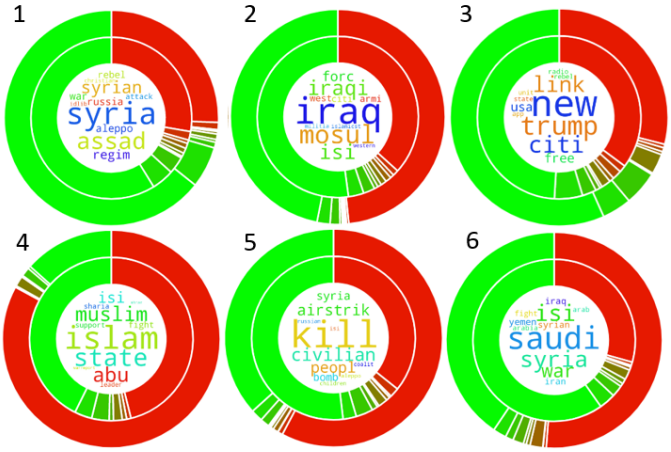

Figure 3. Authenticity distribution in six topics from Twitter Propaganda

In order to visually represent the authenticity of each topic, we used donut charts as depicted in Figure 3. The number of topics in the collected dataset was optimized empirically to produce coherent topics. In total, we found twenty-three coherent topics. For the sake of brevity, we show the authenticity distribution of six topics. In the middle of each donut chart, we include the word cloud representing the topic. Each word cloud includes the terms with the highest probabilities. For example, topic 1 includes the terms: 'Syria', 'Assad', 'regime', 'Aleppo', 'Russia', 'war', 'attack', etc. The inner cycle of the donut chart enclosing each word cloud represents the account authenticity distribution. Similarly, the outer cycle depicts the post level authenticity distribution. Green and red color represent authenticity scores that equal to 1, and 0 respectively. High and low authenticity scores resemble *a legitimate account*, and *fake news promoter* respectively.

We can see, that in some cases (e.g., topics 4, 5, and 6) the fraction of posts is disproportional to the fraction of accounts having the same authenticity level. This means that a few *fake news promoters* took over the online discussion in these topics and may have had strong influence on the rest of the accounts who wrote on this topic. Moreover, we succeeded in detecting several meaningful topics: topic 1 focused on reports of the Syrian civil war in Aleppo, along with the intervention of Russia in support of the Assad regime; topic 2 focused on the war between ISIS and the Iraqi forces in Mosul; topic 3 centered on President Trump and U.S.; topic 4 focused on ISIS and Islamic issues; topic 5 focused on the air strike in Syria during Syrian civil war with emphasis on the killing of civilians; and topic 6 centered on ISIS operations in several locations in the Middle East, such as Saudi Arabia, Yemen, Iraq, and Syria. Based on these results, we believe that the online discussions surrounding the reports of the Syrian civil war in Aleppo regarding Russia forces and the Assad regime, as well as the reports regarding President Trump are genuine. In addition, we can see a higher level of *fake news promoter* participation in topics centering on ISIS and the air strikes in Syria, which raise doubts about the reliability of this news.

E. Ethical Considerations

Collecting information from OSM has raised ethical concerns in recent years [39]. In order to minimize the potential risks that may arise from such activities, this study follows recommendations presented by [40], which deal with ethical challenges regarding OSM and Internet communities.

For this study, we used the Twitter REST API public service for two purposes: first, for obtaining the missing information about the pro-ISIS fanboys (e.g., number of followers), and second, in order to enrich the Twitter Propaganda dataset of unlabeled posts and accounts who posted the same context as the pro-ISIS fanboys. The Twitter REST API collects the information of accounts that agree to share their information publicly. Moreover, the research protocol was approved by the Ben-Gurion University of the Negev Human Research Ethics Committee.

IV. Conclusion

In this paper, we proposed a method for detecting fake news based on distinguishing between *fake news promoters* and *legitimate accounts* participating in the same online discussion. Using the proposed method, we demonstrated the distribution of accounts' authenticity for each topic. As a result, we could identify topics that are prone to OSM manipulation, as well as topics that attract authentic public interest. We believe that the proposed method can be useful for users to detect fake news and misinformation within OSM. Moreover, we introduced an approach for collecting data when there is only information from one class. We believe that this method can be valuable to others when there is a need of collecting samples from the other class in the same context.

 ISBN: 978-1-61208-595-1

Finally, the discovery that link prediction features are capable of improving author type classification is very important and may be an indication of the key role these features play in the domain of fake news identification. In the future, we plan to evaluate the presented approach on additional datasets spanning fake news in multiple domains, such as politics, product reviews, etc. We think that it would be interesting to estimate whether the link prediction features are useful only in the domain of ISIS fake news or they are also useful for classification in other different domains.

ACKNOWLEDGMENT

The authors would like to thank Robin Levy-Stevenson for proofreading this article.

REFERENCES

[1] T. Lavie, M. Sela, I. Oppenheim, O. Inbar, and J. Meyer, "User attitudes towards news content personalization," International journal of human-computer studies, vol. 68, no. 8, 2010, pp. 483–495.

[2] N. J. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: methods for finding fake news," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, 2015, pp. 1–4.

[3] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception detection for news: three types of fakes," Proceedings of the Association for Information Science and Technology, vol. 52, no. 1, 2015, pp. 1–4.

[4] J. Kostka, Y. A. Oswald, and R. Wattenhofer, "Word of mouth: Rumor dissemination in social networks," in International Colloquium on Structural Information and Communication Complexity. Springer, 2008, pp. 185–196.

[5] "How isis uses twitter," https://www.kaggle.com/fifthtribe/how-isis-uses-twitter/discussion/21034, accessed: 2017-08-20.

[6] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep.

[7] J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth, "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication," Discourse Processes, vol. 45, no. 1, 2007, pp. 1–23.

[8] D. M. Markowitz and J. T. Hancock, "Linguistic traces of a scientific fraud: The case of diederik stapel," 2014.

[9] J. T. Hancock, M. T. Woodworth, and S. Porter, "Hungry like the wolf: A word-pattern analysis of the language of psychopaths," Legal and Criminological Psychology, vol. 18, no. 1, 2013, pp. 102–114.

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

[11] D. F. Larcker and A. A. Zakolyukina, "Detecting deceptive discussions in conference calls," Journal of Accounting Research, vol. 50, no. 2, 2012, pp. 495–540.

[12] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Volume 2, 2012, pp. 171–175.

[13] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam." in HLT-NAACL, 2013, pp. 497–501.

[14] M.-C. De Marneffe, B. MacCartney, C. D. Manning et al., "Generating typed dependency parses from phrase structure parses," in Proceedings of LREC, vol. 6. Genoa, 2006, pp. 449–454.

[15] S. Oraby, L. Reed, R. Compton, E. Riloff, M. Walker, and S. Whittaker, "And thats a fact: Distinguishing factual and emotional argumentation in online dialogue," in Proceedings of the 2nd Workshop on Argumentation Mining, 2015, pp. 116–126.

[16] V. W. Feng and G. Hirst, "Detecting deceptive opinions with profile compatibility." in IJCNLP, 2013, pp. 338–346.

[17] V. L. Rubin and T. Lukoianova, "Truth and deception at the rhetorical structure level," Journal of the Association for Information Science and Technology, vol. 66, no. 5, 2015, pp. 905–917.

[18] H. Zhang, Z. Fan, J.-h. Zheng, and Q. Liu, "An improving deception detection method in computer-mediated communication." JNW, vol. 7, no. 11, 2012, pp. 1811–1816.

[19] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: human, bot, or cyborg?" in Proceedings of the 26th annual computer security applications conference. ACM, 2010, pp. 21–30.

[20] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, 2012, pp. 15–15.

[21] K. Lee, P. Tamilarasan, and J. Caverlee, "Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media." in ICWSM, 2013.

[22] J. Song, S. Lee, and J. Kim, "Crowdtarget: Target-based detection of crowdturfing in online social networks," in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. ACM, 2015, pp. 793–804.

[23] J. P. Dickerson, V. Kagan, and V. Subrahmanian, "Using sentiment to detect bots on twitter: Are humans more opinionated than bots?" in ASONAM 2014. IEEE, 2014, pp. 620–627.

[24] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in Proceedings of the 25th International Conference Companion on WWW, 2016, pp. 273–274.

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

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

[27] A. Elyashar, J. Bendahan, R. Puzis, and M.-A. Sanmateu, "Measurement of online discussion authenticity within online social media," in Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2017.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, 2003, pp. 993–1022.

[29] "Twitter rest api," https://dev.twitter.com/rest/public, accessed: 2017-08-22.

[30] C. Zhao and G. Jiang, "Cultural differences on visual self-presentation through social networking site profile images," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2011, pp. 1129–1132.

[31] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in Data engineering (icde). IEEE, 2012, pp. 1273–1276.

[32] K. Lee, S. Webb, and H. Ge, "Characterizing and automatically detecting crowdturfing in fiverr and twitter," Social Network Analysis and Mining, vol. 5, no. 1, 2015, pp. 1–16.

[33] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Serf and turf: crowdturfing for fun and profit," in Proceedings of the 21st international conference on WWW. ACM, 2012, pp. 679–688.

[34] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," science, vol. 286, no. 5439, 1999, pp. 509–512.

[35] E. Adar and L. A. Adamic, "Tracking information epidemics in blogspace," in Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence. IEEE Computer Society, pp. 207–214.

[36] D. Kagan, M. Fire, and Y. Elovici, "Unsupervised anomalous vertices detection utilizing link prediction algorithms," arXiv preprint arXiv:1610.07525, 2016.

[37] R. T. Stern, L. Samama, R. Puzis, T. Beja, Z. Bnaya, and A. Felner, "Tonic: Target oriented network intelligence collection for the social web." in AAAI, 2013.

[38] C. Stachniss, G. Grisetti, and W. Burgard, "Information gain-based exploration using rao-blackwellized particle filters." in Robotics: Science and Systems, vol. 2, 2005, pp. 65–72.

[39] A. Elyashar, M. Fire, D. Kagan, and Y. Elovici, "Guided socialbots: Infiltrating the social networks of specific organizations employees," AI Communications, vol. 29, no. 1, 2016, pp. 87–106.

[40] Y. Elovici, M. Fire, A. Herzberg, and H. Shulman, "Ethical considerations when employing fake identities in online social networks for research," Science and engineering ethics, vol. 20, no. 4, 2014, pp. 1027–1043.

 ISBN: 978-1-61208-595-1