Identifying Misinformation

FARHANA ALAM, Boise State University, USA
MOSTOFA NAJMUS SAKIB, Boise State University, USA
QUDRAT E ALAHY RATUL, Boise State University, USA
ABISHAI JOY, Boise State University, USA

Additional Key Words and Phrases: datasets, Politifact, Feature Engineering, random forest, logistic regression, chi-square, SVM, Buzzfeed

ACM Reference Format:
Farhana Alam, Mostofa Najmus Sakib, Qudrat E Alahy Ratul, and Abishai Joy. 2018. Identifying Misinformation. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Fake news spreading through media outlets poses a real threat to the trustworthiness of information and detecting fake news has attracted increasing attention in recent years. The topic of fake news has drawn attention both from the public and the academic communities. One of the most discussed phenomena after the 2016 US presidential election was the spread of fake news and its possible influence. Fake news is typically written intentionally to mislead readers, which determines that fake news detection merely based on news content is tremendously challenging. In this paper, we are proposing fake news detection based on the news and users' social network features.

1.1 Problem Statement

Consider a set of news 'N' and a set of user 'U', where $u_i, u_j \in$ 'U' and u_j are the followers of u_i. Given a weighted bipartite graph with $G(V, E)$, where $V_L \in N$ and $V_R \in U$. Our goal is to predict whether the given news shared by any user from social media is real or fake and to utilize the results of the analysis, to create a general awareness on the importance of news literacy for assessing the ground truth.

2 RELATED WORKS

The Web2.0 technology has accelerated the consumption of fake news in social media. In one instance, the middle schoolers at Philadelphia believed that the earth is flat because of the idea they picked up from basketball star Kyrie Irving, who said that on a podcast[20].The instructors further reported that it is difficult to change a child's misconceptions once they are exposed to misinformation [20].With the alarming rise of fake news, there is a compelling need for understanding the dynamics of misinformation in social media.

Shu et al. [16] focus on considering a tri-relationship among user, publisher, and news content. User engagements represent the news proliferation process over time, which provides useful auxiliary information to infer the veracity of news articles. Shu et al. [15] have proposed a detection method using explicit and implicit features from data, which has the potential to differentiate fake news.

Jonathan Albrig [1] analyzed the role of Facebook in the election of 2016, the article also discussed the influence of hyper biased news over the search engine and all over the internet. The study claims that not only large social media advertisements are an emotional spread of the misinformation, but the algorithm amplifies it. The study also found that only 60% of the misinformation source is a social media platform, the rest 40% of the misinformation comes from the direct website. The study vastly discussed the Micro Propaganda Machine, which is an influence network, that can tailor users opinions, emotional reaction, and create a viral sharing.

In the paper [8],the authors Srijan Kumar and Neil Shah presented a comprehensive study of the actors, rationale, and impact of successful spreading of false information, as well as, characteristics of it and detection algorithm. They found the creation and spread ing of false information are generally controlled by a single entity with synthetically created fake accounts following each other. The spreading rate of fake is very high, especially at the initial state, even before it is debunked. This paper's survey found that generally, the text is longer, more exaggerated, repetitive, incoherent, and more opinionated in false news, created and spread in a short period of time from the same group of relatively new accounts of fewer reviews with overlapping local networks. Also, the"echo-chamber effect" (self-selective polarizing effect on the content of predetermined beliefs), or improved technologies like content personalization may cause the same kind of fake news to come up several times and make people believe it. However, several algorithms Feature-based, Graph-based, Propagation-modeling based have been created for operative detection of false information.

Horne and Adali enlisted that [5], (f1) fake news articles tend to be shorter in terms of content but use repetitive language, smaller words, less punctuation, and fewer quotes. (f2) fake news articles require a lower educational level to read, use fewer analytic words, use more personal pronouns and adverbs, but fewer nouns. (f3) fake titles are longer, contain shorter words, use more capitalized words, fewer stop words, and fewer nouns overall but more proper nouns. (f4) titles are a strong differentiating factor between fake and real news. Anu et al. [11] reproduced the study by Horne and Adali

Authors' addresses: Farhana Alam, Boise State University, USA; Mostofa Najmus Sakib, Boise State University, USA; Qudrat E Alahy Ratul, Boise State University, USA; Abishai Joy, Boise State University, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
0730-0301/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

and confirm most of the original paper's findings regarding title and body feature differences between fake and real news. However, differently from Horne and Adali, they found that fake titles contain more stop words, and, both titles and bodies express more negative emotions and sentiments than the real ones. Real news articles are more descriptive than fake news ones. Also, they highlight some differences between two different news domains: political and gossip. For instance, among stylistic, psychological, and complexity features in the news title, psychological features are the most important group of features for gossip news, while for political news it is the stylistic features. This shows how gossip news titles tend to be more persuasive than other news domains.

Anu et al. [12] found that News headlines are more informative than news body content, suggesting that we can avoid "read" the news excerpt and focus on other modalities to better detect misleading news. By comparing the news headline and excerpt content, the authors observe that headline characteristic are more consistent than excerpt ones across datasets. They also argued publisher bias is a strong predictor of news validity. In fact, by analyzing information collected from mediabiasfactcheck.com ("the most comprehensive media bias resource on the Internet"), they showed that hyper-partisan (extremely biased towards a political party) news sources are more likely to spread misleading stories than other sources.

Images in news articles also play a role in misleading news detection [2; 4; 6; 18]. Fake images are used in news articles to provoke emotional responses from readers. Images are the most eye-catching type of content in the news; a reader can be convinced of a claim by just looking at the title of the news and the image. Jin et al. [7] showed that including visual and statistical features extracted from news images improves the results for microblogs news verification.

The paper [9], focuses on news and media literacy and defines, explicates, and operationalizes news literacy combining it with the knowledge of news production, distribution, and consumption with five conceptualized domains. For The rationale and purpose of the definition of news literacy, authors firstly put emphasis on importance on knowledge and skill and secondly, discussed how news literacy should be defined and distinguished from related literacies. Authors came to a definition like, " news literacy as knowledge of the personal and social processes by which news is produced, distributed, and consumed, and skills that allow users some control over these processes. ". Authors proposed five knowledge and skills domains – context, creation, content, circulation, and consumption, which make up the core of news literacy though empirically there are some overlaps between each of these domains.

3 DATA DESCRIPTION

The data mainly has two datasets that contain both the news content and social context information. The fake news ground truth was collected from two platforms: BuzzFeed and PolitiFact [14]. The first file (user-user network) from these datasets have a primary identification for the user influencer and their respective followers. While the second file (user news relationships) contains detailed information on news and users, who shared it. There were 120 real news articles and 120 fake, news articles from Politifact media. Similarly, Buzzfeed had 91 real news and 91 fake news articles For identifying the nodes, we used networkx library and the details of the datasets can be seen in table 1. A glimpse of real (Green) and fake (Red) news is in figure 1. The figure 2 shows the percentage of missing value against each attribute in the news.

Table 1. Details of the dataset from Politifact and Buzzfeed medias

	Politifact	Buzzfeed
Real News + Fake News	120+120	91+91
Number of nodes (user-user file)	23865	15257
Number of edges (user-user file)	574744	634750
User-News Relationships	32789	22779

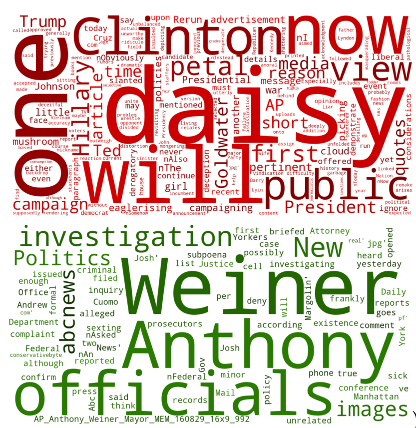

Fig. 1. Glimpse of Fake and Real News in Dataset

The content of the news from Politifact and Buzzfeed media contains the following attributes:

- Source: It indicates the author or publisher of the news article.
- Headline: It refers to the short text that aims to catch the attention of readers and relates well to the major of the news topic.
- Body text: It elaborates the details of the news story. Usually, there is a major claim which shaped the angle of the publisher and is specifically highlighted and elaborated upon.
- Image video:It is an important part of the body content of a news article, which provides visual cues to frame the story.
- Title: Title of the news article.
- Publish dates: The publish date of the news article.
- Canonical link: A canonical link element is an HTML element that helps webmasters prevent duplicate content issues.

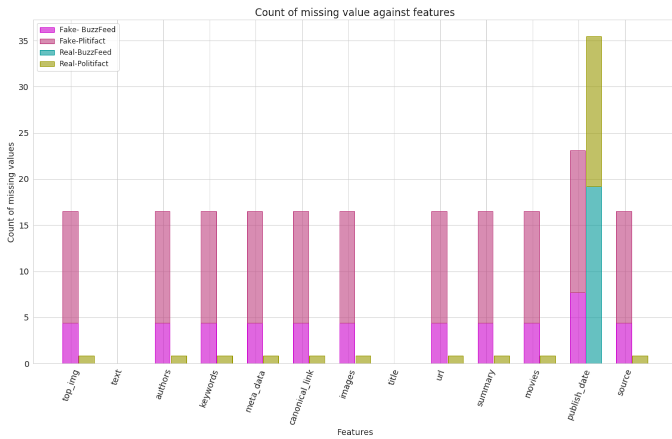

Fig. 2. Distribution of Missing data in Dataset

4 METHODOLOGY AND FEATURE ENGINEERING

In our dataset, we have the source, URL, canonical link, image, and top image link for the news data. For source, we have the news source's web link, whereas URL points out to the specific news along with the source. For all the above-described source and URL related available data, we wanted to understand their relations with the news being real or fake. We divided all of them into three parts: a subdomain, domain, and suffix/ top-level domain. Once the features were computed, we examined multiple rows for the image and URL as there were multiple links were available as a list. Initially, we experimented with label encoding and one-hot encoding to create distinguishable features for fitting in the classifier. These features were later fed in the classifier but they had much lower accuracy. So we removed all these features. Other types of features we used were the author/organization name and the length of the author/organization name. Both the features were calculated and later used for classification purposes but had lower accuracy.

Another type of feature category we focused on was the emotional features. For the emotional features, we used the NRC emotion lexicon [10]. This lexicon has 2 sentiments:positive and negative, along with other emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. We have used the latest version of the lexicon (0.92) which has more than 14000 words. Each of the words is categorized as different emoticons. A function was created which yields an emoticon score and results in the specific emoticon for a word from the above-mentioned category. We have used all those categories as separate features, but none of them performed well in the classifier.

Although we calculated several types of features, most of them generated very low accuracy, forcing us to eliminate the majority of them. Our understanding is that on a larger data-set, all the features that developed low accuracy might help gain better accuracy.

After this failed attempt, the computation of text and network features proved to be successful. As explained by Shu et al., [13], network features are helpful in detecting fake news.

We will be considering the below set of features for our classification problem:

- Text Analysis and News Features : The paper by Horne and Adali [5] describes the importance of text-based features for identifying misinformation. Our text features include, – Word Counts: We computed the number of words in text and title.
  - Word Counts: We computed the number of words in text and title.
  - Punctuations Counts: We computed the number of punctuation in text and title.
  - Uppercases Lowercases Counts: We computed the number of lower and upper case words in text and title.
  - Shares: Number of shares per news. Fake news spread significantly faster, farther and more deeper than the real news [19]. The figure 3 describes distribution of shares among fake and real news.
  - Number of Authors: Number of authors per news. It is important to understand credibility of the fake news. The findings by Sitaula et al., [17] suggest that an author's history of association with fake news, and the number of authors of a news article, can play a significant role in detecting fake news.

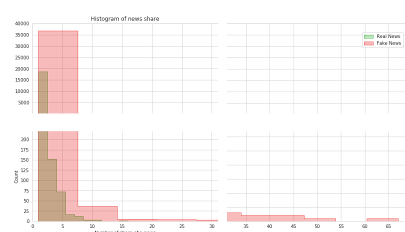

Fig. 3. Distribution of Shares among fake and real news

  - @ Mention Counts: We computed the number of words in text and title.

  - Count of common/extreme words : We computed the number of common and extreme words (e.g., shares, Breaking, Do not) in text.
  - Title Length: We computed the length of title using spacy and NLP.
  - Sentiment Analyzer - Polarity: The paper by Bhutani et al.,[3] explains the importance of sentiment analysis on fake news detection problem. We used cleaned news text, replaced emojis and then applied VADER sentiment analyzer. The sentiment analysis was performed to segregate the texts in three separate categories i.e. positive negative and neutral.
  - Topic Modeling: We used LDA (Latent Dirichlet Allocation) for our analysis and computed 20 topics. LDA helps to build topics per document. Each topics also consists of 20 words with a available weight/importance for each word. All the news body's were later tokenized and computed with the corresponding topics. So for each news we had 20 scores for 20 topics and those were used for the classification.
- Network Features: As explained by Shu et al.,[13], network analysis helps in the detection of fake news. We have datasets that show user-user relations. We used them to identify the importance of the node. Our network features include,
  - Degree Centrality: In our network, connections indicate the number of followers and followees. So, we computed degree centrality to get these features.
  - Closeness Centrality: Closeness centrality gives higher ranks to those nodes which can quickly reach other nodes. So, if we compute closeness centrality of user nodes, those news-nodes connected with higher closeness centrality user-nodes may get a chance of quicker share or spread.
  - Betweenness Centrality: Betweenness centrality quantifies how important nodes are in connecting other nodes. A node with higher betweenness centrality would have more control over the network. Therefore, if an influencer has higher betweenness centrality, then more information passes through him in the network, to his followers. We have data for the influencer-follower network.
  - Eigen vector Centrality: This measure ranks nodes according to their connection with other important nodes. This can also be used to identify the importance of news from user-news data.
  - Pagerank Centrality: This measure is good as it takes care of dangling nodes. It is similar to eigenvector centrality, PageRank ranks nodes according to the probability that a random surfer has to reach that node.

5 EXPERIMENTS AND RESULTS

After feature engineering, we tested our features using the binary classification of whether the news is fake or not on various machine learning algorithms, namely Logistic Regression, Support Vector Machine (SVM), Extra-trees, XGBoost, Naive Bayes, and Random Forest. Since our dataset was balanced, we considered Accuracy, Precision, Recall, and F1-Score and performed 10-fold stratified cross-validation. Out of the 6 classifiers, Random Forest and Extra trees classifier outperformed with an Accuracy of 83.63 and 85.77 respectively as shown in the table 2.

We have also extracted the feature importance for predicting fake news using extra trees classifier. As per fig 4, fig 5 and fig 6, the network feature and text feature are statistically significant. From fig 5 and fig 6, it is evident that degree centrality, count of lower case letters, and the number of authors are relatively important.

Table 2. Details of the dataset from Politifact and Buzzfeed medias

Classifier	Accuracy	Precision	Recall	F1-Score
Logistic Regression	79.63	79.81	78.90	81.08
SVM	78.92	79.09	77.80	81.08
Random Forest	83.63	83.74	83.98	83.87
Extra trees Classifier	85.77	85.70	86.71	84.84
XGBoost	81.51	81.85	80.63	83.39
Naive Bayes	70.62	72.14	69.56	75.36

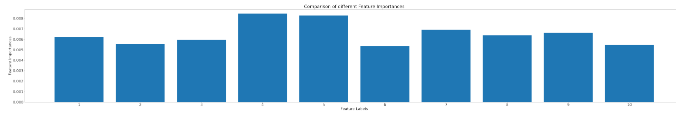

Fig. 4. Feature importance using Extra Trees

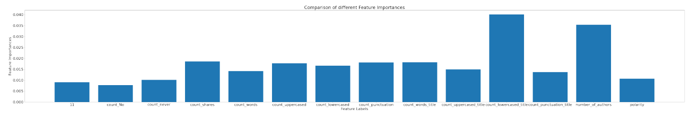

Fig. 5. Feature importance using Extra Trees

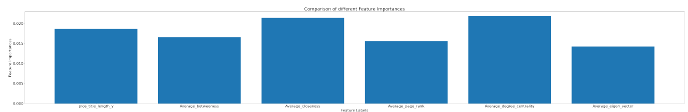

Fig. 6. Feature importance using Extra Trees

6 CONCLUSION

With the increasing popularity of social media, more people consume news from social media instead of traditional news media. However, social media has also been used to spread fake news, which negatively impacts society. In this project, we analyzed the fake news problem by the news content and users' network features. Many existing studies focus on detecting fake news by either content or source analysis, we have used both contents of the news and network properties to detect fake news. Further, this research can be improved by incorporating Spatio-temporal information and analyzing the news's images.

REFERENCES

[1] J Albright. The# election2016 micro-propaganda machine. medium. 2016.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[3] Bhavika Bhutani, Neha Rastogi, Priyanshu Sehgal, and Archana Purwar. Fake news detection using sentiment analysis. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–5. IEEE, 2019.

[4] Manish Gupta, Peixiang Zhao, and Jiawei Han. Evaluating event credibility on twitter. In *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012*, Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012, pages 153–164, United States, 2012. Society for Industrial and Applied Mathematics Publications. Copyright: Copyright 2020 Elsevier B.V., All rights reserved.; 12th SIAM International Conference on Data Mining, SDM 2012 ; Conference date: 26-04-2012 Through 28-04-2012.

[5] Benjamin D. Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, 2017.

[6] Z. Jin, J. Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. *Proceedings of the 25th ACM international conference on Multimedia*, 2017.

[7] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 19(3):598–608, 2017.

[8] Srijan Kumar and Neil Shah. False information on web and social media: A survey, 2018.

[9] Seth Ashley Emily K. Vraga Melissa Tully (corresponding author), Adam Maksl and Stephanie Craft. False information on web and social media: A survey.

[10] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. 29(3):436–465, 2013.

[11] Anu Shrestha and Francesca Spezzano. Textual characteristics of news title and body to detect fake news: A reproducibility study, 2020.

[12] Anu Shrestha, Francesca Spezzano, and Indhumathi Gurunathan. Multi-modal analysis of misleading political news. In Max van Duijn, Mike Preuss, Viktoria Spaiser, Frank Takes, and Suzan Verberne, editors, *Disinformation in Open Online Media*, pages 261–276, Cham, 2020. Springer International Publishing.

[13] Kai Shu, H Russell Bernard, and Huan Liu. Studying fake news via network analysis: detection and mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, pages 43–65. Springer, 2019.

[14] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 8, 2018.

[15] Kai Shu, Suhang Wang, and Huan Liu. Understanding user profiles on social media for fake news detection. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 430–435. IEEE, 2018.

[16] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 312–320, 2019.

[17] Niraj Sitaula, Chilukuri K Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. Credibility-based fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 163–182. Springer, 2020.

[18] D. Tian. on image feature extraction and representation techniques. 2013.

[19] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[20] A Wolfman-Arent. The ongoing battle between science teachers and fake news.