Analysis of Differences in Texts between Real and Fake News

Cristian Lopez, Jackson Maschman, Karina Sindermann, Quan Nguyen,
Sebastian Vazquez-Gasty
Affiliations

1 INTRODUCTION

Thanks to social media, our presence on the Internet has been steadily increasing. Social media connects like-minded people across the globe, fosters relationships, and even shares memorial moments with our loved ones. Nonetheless, in recent years, many companies behind these social networks faced obstacles in controlling fake news on their sites while respecting the freedom of speech. The situation worsened when companies or individuals manipulated these sites to spread misinformation or false information. Even though many companies quickly struck down these contents, the damage has been done to individuals or entities. Therefore, many companies have sought to develop preventative systems to identify fake news and prevent its spread. A common technique is having a team of reviewers that examine each post or content and decide whether the news is reliable or not [1]. However, one concerning problem plaguing them is how to efficiently analyze billions of bytes of data. They would need a team of millions of reviewers to examine all content generated in one day. Thus, our team would like to explore ways that allow companies to narrow down the list of contents needing reviews so that they can better allocate their limited human resources to only the most likely-to-be-false news. In this project, we would like to examine differences in texts, specifically word choices, grammar, and sentence structure, between real and fake news and explore their associations with whether the news is true or false.

2 OBJECTIVES

- Objective 1: Data preprocessing. Tokenizing texts in the inputs.
- Objective 2: Do association mining between word choices and fake news.
- Objective 3: Research existing models that can produce text patterns for fake news and real news and obtain their patterns. Compare these patterns to patterns generated from our association mining.
- Objective 4: Evaluate the accuracy and efficiency of association mining in objective 2 and visualization for significant rules.

3 RELATED WORK

Because fake news is a bias on information that the publisher has manipulated, fake news detection is defined as a binary classification problem [2]. The fake news detection problem is a data mining problem because it consists of two stages: (a) feature extraction and (b) model building. The feature extraction phase seeks to formalize the mathematical structure of news content and related auxiliary data. The model construction phase further develops machine learning models to more accurately distinguish between fake and legitimate news based on the feature representations. In literature, many approaches have been proposed to potentially identify fake news. For instance,

i. Visual-based: by using a classification framework, fake pictures were detected based on a variety of user preferences and features in social media that are manually designed with intuitive mode [3]. Also, visual statistical features have been implemented to identify fake news [4].
ii. Linguistics-based: main feature categories such as: lexical, grammatical, and syntactic, are used in qualitative and quantitative analysis to capture fake news [5].
iii. Post-based: fake news based on personal opinions. In coordination with linguistics-based approaches, it is intended to identify support or disavowal of certain news, and thus, the reliability of posts is evaluated by their credibility features [6]
iv. Network-based: The key idea is to track the source and the spread of misinformation based on networks of users that share common interests [7].

In this proposal, we attempt to identify fake news focused on linguistics-based features by using text mining association and comparing our results to state-of-the-art models to provide robust criteria for identifying unreliable news from any social media.

4 PROBLEM DEFINITION

The spread of false information widely through social media and other online platforms has become a major concern in today's society. This phenomenon can have serious consequences, such as shaping public opinion, influencing elections, and promoting social unrest. Therefore, it is important to be able to distinguish between real and fake news. This project aims to analyze the linguistic features that distinguish real news from fake news. The inputs for this scope are primarily social media sentences, among other attributes and the output will be how likely new sentences are to be fake. It is important to mention that for this project, the dataset is written in English, meaning that the output only will work in this language.

Considering the increasing level of fake news, it is important to address this situation and try to collaborate with the detection of this type of news to reduce social problems like undermining a democratic process or creating conspiracy theories and hate speech.

5 DATASET

We use the popular Fake News Datasets LIAR [9] and part of the FakeNewsNet [8] Dataset. Both Datasets were constructed using the fact-checking website PolitiFact. PolitiFact is an independent, nonpartisan online fact-checking website. It primarily rates the accuracy of claims or statements made by political news and politicians in the U.S.

GossipCop also fact-checked part of the news in the FakeNewsNet Dataset. GossipCop (now suggest) checks fake news from the entertainment and celebrity sectors in the U.S. published in magazines and web portals.

LIAR [9] is a publicly available Dataset for fake news detection. It contains 12836 short statements from 2007 to 2016. A POLITIFACT editor evaluated each statement. For the Truthfulness ratings, six finely nuanced labels are used (true, mostly true, half-true, mostly false, false, and pants on fire). Furthermore, the Dataset contains the statement, the subject of the statement, the context of when the statement was made, the speaker's name, job title, home state, and party affiliation, as well as their historical count of inaccurate statements. An example entry of this database is shown in Table 1.

Although the authors state that the database consists of 12836 instances, we can only access 12791 values (see Table 2).

Table 1. Example of LIAR data set.

label	4 (barely true)
statement	"Most of the (Affordable Care Act) has already in some sense been waived or otherwise suspended."
subject	"Healthcare"
speaker	"George Will"
job_title	"Columnist"
state_info	"Maryland"
party_affiliation	"columnist"
barely_true_counts	7
false_counts	6
half_true_counts	3
mostly_true_counts	5
pants_on_fire_counts	1
context	"Comments on 'Fox News Sunday'"

Table 2. Description of the LIAR data set

label	train	test	validation	Sum
barely true	1654	212	237	2103
FALSE	1995	249	263	2507
half-true	2114	265	248	2627
mostly true	1962	241	251	2454
pants-fire	839	92	116	1047
TRUE	1676	208	169	2053
Sum	10240	1267	1284	12791

The FakeNewsNet [8] Dataset for detecting fake news provides a comprehensive social media context. The Dataset includes articles and associated tweets checked by PolitiFact or GossipCop. It contains 467 thousand tweets in the PolitiFact Dataset and 1.25 million tweets in the GossipCop Dataset, labeling them as either true or fake.

Due to Twitter's privacy policy and the copyrights of news publishers, the full Dataset cannot be made public. We will use a small sample of the Dataset, which is publicly available. This sample consists of 4 files that differ by the truth value of the story (false or real) and the verifier (PolitiFact or GossipCop). The sample contains 23424 verified news titles (see Table 3). The attributes of the Datasets are id, URL of the news, title, and Twitter id.

Table 3. Description of the FakeNewsNet data set

	Real	Fake	Sum
GossipCop	16818	5335	22153
PolitiFact	798	473	1271
Sum	17616	5808	23424

6 APPROACH

For data preprocessing, we started by pulling the Dataset from their sources. After that, we removed irrelevant columns and started tokenizing the texts. For the first steps, we will focus on associations between individual words and fake news. If this path does not yield fruitful results, we will move to more complicated phrases and integrate more words to form context. We also planned to do some exploratory data analysis (EDA), like using word cloud and checking how frequently some words are, which might indicate fake news.

For the second objective, about association techniques, we would like to use the Apriori Algorithm we developed during the course or Weka to find interesting patterns. For this purpose and after the first objective, we are planning to select the attributes to optimize the association based on the previous preliminary analysis. At the same time, we will identify special features of our data by consulting with linguistics experts so that we can greedily optimize our algorithm to reduce runtime complexity. This may include applying what linguistics have identified as common phrases in fake news and focusing on building more complicated expressions from these phrases.

For the third objective, we research existing data mining models that can analyze fake news text patterns and obtain patterns generated by these models. Alternatively, we can adapt our inputs and run our inputs through these models. We will look through GitHub, past conference papers, and online Kaggle contests. The goal is to use these outputs to observe the performance and shortcomings of our association mining. We will compare the percentage of rules that are shared by both outputs. Afterward, we will look at the rules that existing models generated but our mining failed to do so.

For the fourth objective, we will evaluate the accuracy and efficiency of our association mining techniques as outlined in section 7: Evaluation. After that, we will create a visualization to showcase the support level and confidence level of notable rules. At the same time, we will also show some examples of how these rules are applied. For this task, we plan to use Tableau or PowerBI to create robust and interactive visualization.

7 EVALUATION

For associations between texts and fake news, we would like to look for rules that have at least a 90% confidence rate with at least a 5% support rate. The support rate may change depending on the types of rules since it is more common for an individual word to show up in sentences than phrases to do so. These numbers may change as we develop a further understanding of the subjects.

For the visualization in objective 4, the visualization should be interactive and display the top 10 rules in terms of the support level and confidence level. At the same time, it should show at least three examples and highlight which rules are used to determine whether the example news is fake. The visualization will allow users to look up certain words and phrases.

8 IMPLEMENTATION PLAN AND TIMELINE

We expect to have finished the preprocessing and preliminary analysis during the first week of April. Then, we plan to work simultaneously on objectives two and three. This can take at least 2-2.5 weeks. For the last objective, related to how we will present the project, we estimate to use the last week of April. Ideally, by the first week of May, everything will be completed, and we will mostly spend time reviewing current works and preparing a presentation.

REFERENCES

[1] L. Chiou and C. E. Tucker, "Fake News and Advertising on Social Media: A Study of the Anti-Vaccination Movement," SSRN Electronic Journal, 2018, doi: https://doi.org/10.2139/ssrn.3209929.

[2] M. Gentzkow, J. Shapiro, and D. Stone. "Media bias in the marketplace: Theory," Technical report, National Bureau of Economic Research, 2014.

[3] A. Gupta et al., "Faking sandy: characterizing and identifying fake images on Twitter during hurricane sandy," Proceedings of the 22nd International Conference on World Wide Web, 2013.

[4] D. Vishwakarma, and C. Jain, "Recent state-of-the-art of fake news detection: A Review," 2020 International Conference for Emerging Technology (INCET), 2020.

[5] Szczepański, M., Pawlicki, M., Kozik, R. et al. New explainability method for BERT-based model in fake news detection. Sci Rep 11, 23705, 2021.

[6] C. Castillo, M. Mendoza, and B. Poblete. "Information credibility on Twitter," Proceedings of the 22nd International Conference on World Wide Web, 2011.

[7] S. Kwon, et al., "Prominent features of rumor propagation in online social media," IEEE 13th

International Conference on Data Mining, p. 1103–1108. IEEE, 2013.

[8] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media." arXiv, Mar. 27, 2019. Accessed: Mar. 23, 2023. [Online]. Available: http://arxiv.org/abs/1809.01286

[9] W. Y. Wang, "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection." arXiv, May 01, 2017. Accessed: Mar. 23, 2023. [Online]. Available: http://arxiv.org/abs/1705.00648