

FAKE NEWS ARTICLES CLASSIFYING USING NATURAL LANGUAGE PROCESSING TO IDENTIFY IN-ARTICLE ATTRIBUTION AS A SUPERVISED LEARNING ESTIMATOR

Dr.CHANDRA MOHAN

¹Associate Professor, CSE,Chalapathi Institute of Technology,Guntur, India

²UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

³UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

⁴UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

⁵UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

ABSTRACT: Intentionally deceptive content presented under the guise of legitimate journalism is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most so-called 'fake news' is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as making excessive use of unsubstantiated hyperbole and non-attributed quoted content. In this paper, the results of a fake news identification study that documents the performance of a fake news classifier are presented. The NLTK, Natural Language, and SciPy Toolkits were used to develop a novel fake news detector that uses quoted attribution in machine learning system as a key feature to estimate the likelihood that a news article is fake. This process is called influence mining and this novel technique is presented as a method that can be used to enable fake news and even propaganda detection. In this paper, the research process, technical analysis, technical linguistics work, and classifier performance and results are presented. The paper concludes with a discussion of how the current system will evolve into an influence mining system.

1.INTRODUCTION

Intentionally deceptive content presented under the guise of legitimate journalism (or 'fake news,' as it is commonly known) is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns.

Most fake news is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a fake news classifier are presented and discussed in this paper.

As an increasing amount of our lives is spent interacting online through social media platforms, more and more people tend to hunt out and consume news from social media instead of traditional news organizations.[1] The explanations for this alteration in consumption behaviors are inherent within the nature of those social media platforms: (i) it's often more timely and fewer expensive to consume news on social media compared with traditional journalism, like newspapers or television; and (ii) it's easier to further share, discuss, and discuss the news with friends or other readers on social media. For instance, 62 percent of U.S. adults get news on social media in 2016, while in 2012; only 49 percent reported seeing news on social media [1]. It had been also found that social media now outperforms television because the major news source. Despite the benefits provided by social media, the standard of stories on social media is less than traditional news

organizations. However, because it's inexpensive to supply news online and far faster and easier to propagate through social media, large volumes of faux news, i.e., those news articles with intentionally false information, are produced online for a spread of purposes, like financial and political gain. It had been estimated that over 1 million tweets are associated with fake news "Pizzagate" by the top of the presidential election. Given the prevalence of this new phenomenon, "Fake news" was even named the word of the year by the Macquarie dictionary in 2016 [2]. The extensive spread of faux news can have a significant negative impact on individuals and society. First, fake news can shatter the authenticity equilibrium of the news ecosystem for instance; it's evident that the most popular fake news was even more outspread on Facebook than the most accepted genuine mainstream news during the U.S. 2016 presidential election. Second, fake news intentionally persuades consumers to simply accept biased or false beliefs. Fake news is typically manipulated by propagandists to convey political messages or influence for instance, some report shows that Russia has created fake accounts and social bots to spread false stories. Third, fake news changes the way people interpret and answer real news, for instance, some fake news was just created to trigger people's distrust and make them confused; impeding their abilities to differentiate what's true from what's not. To assist mitigate the negative effects caused by fake news (both to profit the general public and therefore the news ecosystem). It's crucial that we build up methods to automatically detect fake news broadcast on social media [3]. Internet and social media have made the access to the news information much easier and comfortable [2]. Often Internet users can pursue the events of their concern in online form, and

increased number of the mobile devices makes this process even easier. But with great possibilities come great challenges. Mass media have an enormous influence on the society, and because it often happens, there's someone who wants to require advantage of this fact. Sometimes to realize some goals mass-media may manipulate the knowledge in several ways. This result in producing of the news articles that isn't completely true or maybe completely false. There even exist many websites that produce fake news almost exclusively.

They intentionally publish hoaxes, half-truths, propaganda and disinformation asserting to be real news – often using social media to drive web traffic and magnify their effect. The most goals of faux news websites are to affect the general public opinion on certain matters (mostly political). Samples of such websites could also be found in Ukraine, United States of America, Germany, China and much of other countries [4]. Thus, fake news may be a global issue also as a worldwide challenge. Many scientists believe that fake news issue could also be addressed by means of machine learning and AI [5]. There's a reason for that: recently AI algorithms have begun to work far better on many classification problems (image recognition, voice detection then on) because hardware is cheaper and larger datasets are available. There are several influential articles about automatic deception detection. In [6] the authors provide a general overview of the available techniques for the matter. In [7] the authors describe their method for fake news detection supported the feedback for the precise news within the micro blogs. In [8] the authors actually develop two systems for deception detection supported support vector machines and Naive Bayes classifier (this method is employed within the system described during this paper as well) respectively. They collect the info by means

of asking people to directly provide true or false information on several topics – abortion, execution and friendship. The accuracy of the detection achieved by the system is around 70%. This text describes an easy fake news detection method supported one among the synthetic intelligence algorithms – naïve Bayes classifier, Random Forest and Logistic Regression. The goal of the research is to look at how these particular methods work for this particular problem given a manually labelled news dataset and to support (or not) the thought of using AI for fake news detection. The difference between these article and articles on the similar topics is that during this paper Logistic Regression was specifically used for fake news detection; also, the developed system was tested on a comparatively new data set, which gave a chance to gauge its performance on a recent data.

2. LITERATURE SURVEY

Mykhailo Granik et. al. in their paper [3] shows a simple approach for fake news detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set of Facebook news posts. They were collected from three large Facebook pages each from the right and from the left, as well as three large mainstream political news pages (Politico, CNN, ABC News). They achieved classification accuracy of approximately 74%. Classification accuracy for fake news is slightly worse. This may be caused by the skewness of the dataset: only 4.9% of it is fake news.

Himank Gupta et. al. [10] gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of tweets in 1 sec. Firstly, they have collected 400,000 tweets from

HSpam14 dataset. Then they further characterize the 150,000 spam tweets and 250,000 non-spam tweets. They also derived some lightweight features along with the Top-30 words that are providing highest information gain from Bag-of-Words model. 4. They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately 18%.

Marco L. Della Vedova et. al. [11] first proposed a novel ML fake news detection method which, by combining news content and social context features, outperforms existing methods in the literature, increasing its accuracy up to 78.8%. Second, they implemented their method within a Facebook Messenger Chabot and validate it with a real-world application, obtaining a fake news detection accuracy of 81.7%. Their goal was to classify a news item as reliable or fake; they first described the datasets they used for their test, then presented the content-based approach they implemented and the method they proposed to combine it with a social-based approach available in the literature. The resulting dataset is composed of 15,500 posts,

Twitter by learning to predict accuracy assessments in two credibility-focused Twitter datasets: CREDBANK, a crowd-sourced dataset of accuracy assessments for events in Twitter, and PHEME, a dataset of potential rumors in Twitter and journalistic assessments of their accuracies. They apply this method to Twitter content sourced from BuzzFeed's fake news dataset. A feature analysis identifies features that are most predictive for crowd-sourced and journalistic accuracy assessments, results of which are consistent with prior work. They rely on identifying highly retweeted threads of conversation and use the features of these threads to classify stories, limiting this work's applicability only to the set of popular tweets. Since the majority of tweets are rarely retweeted, this method therefore is

only usable on a minority of Twitter conversation threads.

In his paper, Shivam B. Parikh et. al. [13] aims to present an insight of characterization of news story in the modern diaspora combined with the differential content types of news story and its impact on readers. Subsequently, we dive into existing fake news detection approaches that are heavily based on text-based analysis, and also describe popular fake news datasets. We conclude the paper by identifying 4 key open research challenges that can guide future research. It is a theoretical Approach which gives Illustrations of fake news detection by analyzing the psychological factors

3. EXISTING SYSTEM

The methodology used here after the research is using the Naive Bayes algorithm which is one of the algorithms of supervised learning. In this paper, the libraries of Python such as Sci-kit are used. Python is one of the open source IDE or software to download and install. Python has many and huge extensions and libraries which helps to develop with one of the emerged technology such as Machine learning. All the different types of machine learning algorithms are available in Sci-kit libraries which are available for Python. Django is one of the deployment model which can be used to develop the web pages using HTML, CSS, Java script which will provide the client side.

4. PROPOSED SYSTEM

In our proposed system of "classifying fake news fake news article"

We are using the Random forest algorithm, Decision tree algorithm, Support vector machine, bidirectional rnn, gated recurrent units which are the supervised machine learning algorithms. Streamlit is used for the Front end. csv file for back end.

5.SYSTEM ARCHITECTURE

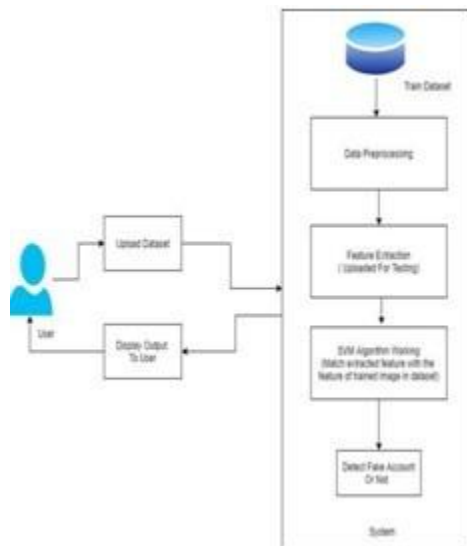


Fig: .1 System Architecture

6. IMPLEMENTATION

The 4 main factors to determine fake news are:

Data Collection:

Collect sufficient data samples and legitimate software samples.

Data Pre-processing:

Data Augmented techniques will be used for better performance.

Pre-Processing:

We will provide the Dataset of both real and fake news articles.

Data will be in many different formats. We need to categories the data. Then, cleaning of the data involves various stages:

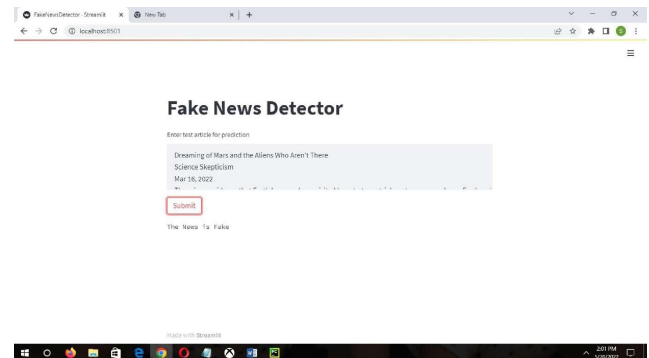
Train and Test Modelling:

After all the preprocessing is done we will train our model Using train and test data applying the algorithms.

Predicting output:

Now we give our test data to the model and the model predicts the output whether the news article is fake or real.

7.SCREEN SHOT



8. CONCLUSION

This research is the results of a study that produced a limited fake news detection system. The work presented here in is novel in this topic domain in that it demonstrates the results of a full-spectrum research project that started with qualitative observations and resulted in a working quantitative model. The work presented in this paper is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes.

REFERENCES

- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective" arXiv: 1708.01967v3 [cs.SI], 3 Sep 2017.
- Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu, "Fake News Detection on Social Media: A Data Mining Perspective" arXiv:1708.01967v3 [cs.SI], 3 Sep 2017.

- M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), Kiev, 2017, pp. 900-903.
- Fake news websites. (n.d.) Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Fake_news_website. Accessed Feb. 6, 2017.
- Cade Metz. (2016, Dec. 16). The bittersweet sweepstakes to build an AI that destroys fake news.
- Conroy, N., Rubin, V. and Chen, Y. (2015). "Automatic deception detection: Methods for finding fake news" at Proceedings of the Association for Information Science and Technology, 52(1), pp.1-4.
- Markines, B., Cattuto, C., & Menczer, F. (2009, April). "Social spam detection". In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web (pp. 41-48).
- Rada Mihalcea, Carlo Strapparava, The lie detector: explorations in the automatic recognition of deceptive language, Proceedings of the ACL-IJCNLP
- Kushal Agarwalla, Shubham Nandan, Varun Anil Nair, D. Deva Hema, "Fake News Detection using Machine Learning and Natural Language Processing," International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7, Issue-6, March 2019
- H. Gupta, M. S. Jamal, S. Madisetty and M. S. Desarkar, "A framework for real-time spam detection in Twitter," 2018 10th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, 2018, pp. 380-383
- M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro and L. de Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," 2018 22nd Conference of Open Innovations Understanding the random forest by Association (FRUCT), Jyväskylä, 2018, pp. 272-279.
- C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," 2017 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, 2017, pp. 208-215.
- S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Miami, FL, 2018, pp. 436-441
- Scikit-Learn- Machine Learning In Python Dataset- Fake News detection William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648, 2017.
- Shankar M. Patil, Dr. Praveen Kumar, "Data mining model for effective data analysis of higher education students using MapReduce" IJERMT, April 2017 (Volume-6, Issue-4).
- Aayush Ranjan, "Fake News Detection Using Machine Learning", Department Of Computer Science & Engineering Delhi Technological University, July 2018.
- Patil S.M., Malik A.K. (2019) Correlation Based Real-Time Data Analysis of Graduate Students Behaviour. In: Santosh K., Hegadi R. (eds) Recent Trends in Image Processing and Pattern Recognition. RTIP2R 2018. Communications in Computer and Information Science, vol 1037. Springer, Singapore.
- Badreesh Shetty, "Natural Language Processing (NLP) for machine learning" at towardsdatascience, Medium.
- NLTK 3.5b1 documentation, Nltk generate n gram
- Ultimate guide to deal with Text Data (using Python) – for Data Scientists and Engineers by Shubham Jain, February 27, 2018
- Understanding the random forest by Anirudh Palaparthi, Jan 28, at analytics vidya.
- Anirudh Palaparthi, Jan 28, at analytics

vidya.

[24] Shailesh-Dhama, "Detecting-Fake-News-with-Python", Github, 2019

[25] Aayush Ranjan, "Fake News Detection Using Machine Learning", Department Of Computer Science & Engineering Delhi Technological University, July 2018.