Labelling and Classification of Fake News on Social Media using Citizen Science

Ankit Gupta(CS1922), Mohamed Anas(CS1915), Subrat Panda(CS1913)
M.Tech(CS), Database Management Systems, Second Semester,
Indian Statistical Institute, Kolkata

July 20, 2020

Abstract

The proliferation of Internet and Mobile Technologies are causing many services to go online. The use of social media platform is pervasive nowadays. One of the services which have got the attention of the researcher is news. Availability of any news is more than easy than before due to the use of the online medium of providing services. The news can now reach faster to us than before. More often, news reaches via social media faster than news articles on authentic resources. The news while being shared online changes its authenticity by less informed users or by some miscreants wanting to cause chaos in public and gaining commercial or political advantage. The detection of fake news in social media is a big challenge for researchers to design an automated process to perform the task. This citizen science project work on using knowledgeable citizens to label the news for fake or real. The labelled news by citizens can then be used to challenge news article from the social media for their authenticity. This project deals with the development of web interface to label and classify news article shared on social media over the Internet using the help from informed citizens.

1 Introduction

The information can be misleading with deceptive words. The rise in usage of social media platforms such as Facebook, Twitter, Whatsapp etc. are causing the spread of information at an exponential rate by rapid sharing of information online. The number of users on such social media is growing at a very fast rate due to development in Internet and Mobile Technologies.

Fake news intentionally presents misinformation or deception spreading through both traditional print news media or online social media [9]. Fake news existed for a long time, since the "Great moon hoax" publisher in 1835 [6]. In the last decade, due to the booming of various social media, the appearance of fake news has been increased for various commercial and political purposes. The deceptive words in any news article consume online users' rational judgement, which can bring chaos in the general public. For example, in recent times of COVID pandemic, the fake news related to medicine/cure of COVID-19 caused

chaos in general public and it took very much effort to bring down such fake articles. The fake news also significantly affects the electoral results [9]. It has become very important to improve the trustworthiness of information shared in online social media by identifying the fake news.

The reckless sharing of fake news on the social media platform has allowed the content to reach wide audiences which need proper monitoring. Nowadays, researchers are contributing to artificial intelligence(AI) algorithms to identify fake news [1]. However, the results from algorithms are yet to match up with the rate of sharing of fake news. Most of the algorithms perform web scrapping from the shared articles on social media platforms. Nevertheless, it's quite difficult to label all the news article for fake or real. Until the labelled data of news is available, most of the AI algorithm will get quite complex to perform accurate monitoring of fake news.

The labelling of a news article to be real or fake manually will be a very tedious task. The automated AI algorithm as of now might get very complex to produce accurate results which might consume very computational resources. The concept of "Citizen Science" can be used for labelling of news articles. The citizen science projects involve an informed and knowledgeable public who participates in scientific research. Nowadays, due to the availability of online facilities, there is a rise of citizen science projects which takes advantage of online platforms, therefore, covering online public all over the world. In citizen science, the public need not have lot of knowledge regarding the important problem they are solving. The problem is reduced or converted to a simple form such that people can solve them using basic qualitative and quantitative skills.

This project involves informed online public to classify the news article for its truthfulness by participating in the online platform developed as a part of this project. To avoid damage to accuracy the participants need to provide the source information of authenticity of news while classifying the news. The labelled news articles is then used for training an online machine-learning(ML) algorithm to classify further similar news whenever requested for checking of the authenticity of any similar news article. The news articles are collected by web scrapping from social media platforms such as Twitter, Facebook etc. and from the News provider APIs.

The code-base of this project is available on the Github at https://github.com/dbms-ctzs/sage, nevertheless, this project(sage) is live on https://sage-citizen.herokuapp.com with some reduced functionality. The later section of this report discusses web-scraping of the news article, database design and machine-learning techniques used in this project.

2 Web-scraping for News Articles

The spread of fake news profoundly caused by two of the important sources a) few news publishing companies and b) social media platforms. Very few specific news publishing companies produce fake news articles, mostly these articles are targeted for humour purposes, however, these humorous news articles can be assessed earnestly by some audiences and can affect the readers' psychology adversely which may initiate rumours. The social networking sites such as Twitter, Facebook etc. are primary platforms for sharing and commenting on any news articles. These platforms significantly boost the spread of a news article because of

re-posting, sharing etc. This effect is disastrous when the news is fake or false which perhaps leads to chaos.

Web scraping involves extraction of required information from online sources using tools such as web crawlers, application-programming-interfaces(APIs), RSS feeds etc. The APIs for collecting data is made available by the online platforms to the public for accessing a large amount of required data which are difficult to acquire manually [8].

For this project, fake articles and headlines were obtained from Twitter-API [5] and Newscatcher API [3]. Twitter API allows searching tweets with filtering on a wide range of topic covering science, politics, country, date-published, number-of-retweets, and also allows searching of articles by using key-words like COVID19. Whereas, Newscatcher-API allows searching for the relevant news topics which are published by newspaper publishers companies on their online platform. The news articles collected from different resources available online are then stored in our database. The code for the web-scraping of the news article is available at https://github.com/dbms-ctzs/sage/tree/master/webscrapping.

3 Database Design and Evaluation

The database is important for all online services which handle users authentication. This project uses the SQLite database for all the activities involving a database. SQLite stores the entire database (definitions, tables, indices, and the data itself) as a single cross-platform file on a host machine. It implements this simple design by locking the entire database file during writing. SQLite read operations can be multitasked, though writes can only be performed sequentially [7].

The database design for our project involves the following relation tables. **UserInfo**, **NewsInfo**, **FeedbackInfo**. The Entity-Relationship(ER) diagram for our project is demonstrated in Figure 1. The purpose and details of each tables are discussed below.

1. **UserInfo** : This tables handles the details of users registered on our website. Primarily used for authentication purpose.
    - **userID(Primary Key)** - A unique number generated for each user.
    - **name** - Name of the user.
    - **e-mail** - Email address of the user.
    - **password** - Password of the user.
2. **NewsInfo**: Stores news articles gathered using web-scraping.
    - **newsID(Primary Key)** - A unique number generated for each news article/tweet that is scraped.
    - **content** - News content of the article/tweet.
    - **timestamp** - Date time details of the article/tweet.
3. **FeedbackInfo Schema**
    - **userID(Foreign Key)** - ID of user who gave feedback for the news with newsID.

- **newsID(Foreign Key)** - ID of news for which feedback was given by the corresponding userID.
- **timestamp** - Date time details of the feedback.
- **feedback** - Feedback on whether the news is fake or real.

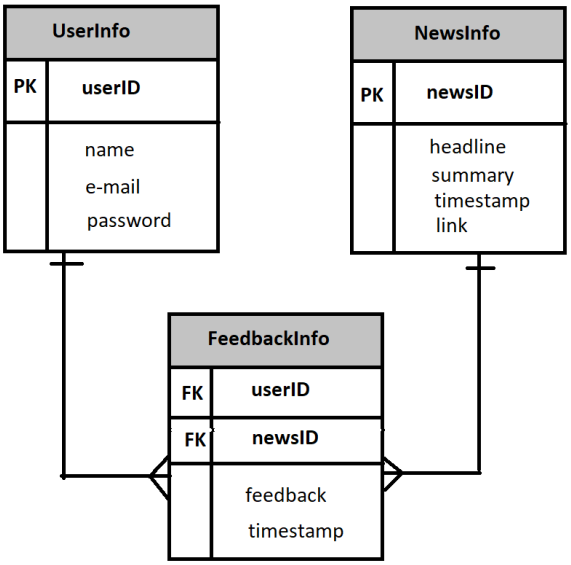

Figure 1: ER diagram of the database.

There is a many to many relationship between **UserInfo** and **NewsInfo** which is captured by the **FeedbackInfo** table.The **FeedbackInfo** table has been designed taking care that a user can select whether a news is fake or real only once, so both userID and newsID can act as a unique identifier. If in case the application allows people to change their status on whether a news is fake or real then the feedback given should also be used as an identifier. The contribution of the user, in terms of how many authentic feedbacks etc., are calculated using the FeedbackInfo table. Other relevant information can be obtained using joins. For example, the inner join between UserInfo and FeedbackInfo gives the news articles commented by a particular user. These three tables form the backbone of the application. Other tables may be added as per the need for any advanced feature.

4 Machine Learning for Fake News Detection

The news article gathers from various online sources using web-scraping as discussed previously is exhibited to online users which distinguished under "citizen science" aspects of the project. The online public helps to label the news articles to be fake or real from the database of our project. The ML algorithm is then used on the labelled news article dataset for classifying similar news contents which may arrive in future. The purpose of this classification is to improve the time required to detect fake news.

In this project, we have used the NLP (Natural Language Processing) techniques for detecting the 'fake news', that is, misleading news stories that come from the non-reputable

sources. The goal can be achieved only by building a model based on a count vectorizer (using word tallies) or a Term Frequency Inverse Document Frequency (TFIDF) matrix i.e. word tallies relative to how often they're used in other articles in your dataset. However, these models do not consider important qualities such as ordering in words or context, perhaps, two news articles having similar words can have a completely different meaning [4].

In this project, we have applied the **Multinomial Naive Bayes** classifier for the required classification of news articles. Multinomial Naive Bayes classifier is a specific instance of a Naive Bayes classifier which uses a multinomial distribution for each of the features and it is suitable for classification with discrete features (e.g., word counts for text classification). In order to create a model to classify an article into fake or real based on its words and phrase, we have employed a Multinomial Naive Bayes classifier on a dataset with labels of fake or real [2].

About Dataset, Model and Performance:

The labelled test dataset **"news.csv"** of news articles is taken from the Kaggle which contains 53429 rows and 3 columns(features). Features of this dataset are title, text and label.

The ML model is build based on the **count vectorizer** and a **TF-IDF matrix** i.e word tallies relatives to how often they are used in other articles. Since this problem is a kind of text classification, implementing a Multinomial Naive Bayes classifier will be suitable which is considered standard for text-based processing.

Confusion Matrix is taken as a performance matrix here. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa).

5 Modules

The web application is developed using open-source tools such as Django and Flask for back-end and Bootstrap for front-end.The project is primarily divided into three main modules: 1) Login and User Registration, 2) Dashboard and Profile and 3) ML for Fake News Detection.

1. **Login and Registration:** There are two modes for login, user mode and administrator mode having separate privileges. Administrator mode has more privileges than users, an administrator can confer some extra privileges to users too. The process of registration is different for administrators as they have to create their account by running command "`python manage.py createsuperuser`" in the shell. Normal users can use registration and login option available on the front-end web page. The facility to login with **Github** and **Google+** is also provided.

2. **Dashboard and Profile:** After user authentication, the dashboard fetches news article from the database where users can help us to label news article using provided "fake" and "real" buttons. The authentic source of the news article can be shared with the provided text box area for commenting.

3. **ML for Fake News Detection:** Users can check the authenticity of the news articles by providing the link of that news article. Now, a ML algorithm is started to work and gives result with high accuracy. "`sage/ctzs/app.py`"

6 Conclusion and Future Work

The explosive growth of social media fostering people to consume online news contents significantly higher than traditional printed media. However, social media also has been used inaptly to spread fake news, which has undesired impacts on individual users and broader aspects of society. In this project, we explored the problem of fake news by reviewing existing works of literature in two phases: characterization and classification. In the characterization phase, we introduced the use of citizen science for detecting fake news articles. In the classification phase, we reviewed existing fake news classification approaches from a data mining perspective including features extraction and model construction. Although, the project has successfully achieved implementation of every individual modules separately, nevertheless, the integration between modules will be one of the future goals of the project.

References

[1] Jain, A., Kasbe, A.: Fake News Detection. In: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). pp. 1–5 (Feb 2018). https://doi.org/10.1109/SCEECS.2018.8546944

[2] Kai Shu, Amy Sliva, S.W.J.T., Liu, H.: Fake News Detection on Social Media:A Data Mining Perspective. pp. 11–12. https://www.kdd.org/exploration_files/19-1-Article2.pdf

[3] Newscatcher: Newcatcher api, https://newscatcherapi.com

[4] Pantech Solutions: Fake news detection using machine learning, https://www.pantechsolutions.net/fake-news-detection-using-machine-learning

[5] Twitter: Twitter api docs, https://developer.twitter.com/en/docs

[6] Wikipedia: Great Moon Hoax (May 2020), https://en.wikipedia.org/w/index.php?title=Great_Moon_Hoax&oldid=958664972, page Version ID: 958664972

[7] Wikipedia: SQLite (Jul 2020), https://en.wikipedia.org/w/index.php?title=SQLite&oldid=965673692, page Version ID: 965673692

[8] Wikipedia: Web scraping (Jul 2020), https://en.wikipedia.org/w/index.php?title=Web_scraping&oldid=967764860, page Version ID: 967764860

[9] Zhang, J., Dong, B., Yu, P.S.: FakeDetector: Effective Fake News Detection with Deep Diffusive Neural Network. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE). pp. 1826–1829. IEEE, Dallas, TX, USA (Apr 2020). https://doi.org/10.1109/ICDE48307.2020.00180, https://ieeexplore.ieee.org/document/9101457/