FAKE NEWS DETECTION USING LOGISTIC REGRESSION

Fathima Nada[1], Bariya Firdous Khan[2], Aroofa Maryam[3], Nooruz-Zuha[4], Zameer Ahmed

[1,2,3,4]*Anjuman Institute of Technology and Management , Bhatkal*
[5]*Under the guidance of (Professor of Computer Science and Engineering department AITM, Bhatkal)*

---***---

Abstract - *Proliferation of misleading information in everyday access media outlets such as social media feeds, news blogs, and online newspapers have made it challenging to identify trustworthy news sources, thus increasing the need for computational tools able to provide insights into the reliability of online content. In this paper, we focus on the automatic identification of fake content in the news articles. First, we introduce a dataset for the task of fake news detection. We describe the pre-processing, feature extraction, classification and prediction process in detail. We've used Logistic Regression language processing techniques to classify fake news. The pre-processing functions perform some operations like tokenizing, stemming and exploratory data analysis like response variable distribution and data quality check (i.e. null or missing values). Simple bag-of-words, n-grams, TF-IDF is used as feature extraction techniques. Logistic regression model is used as classifier for fake news detection with probability of truth.*

Key words: **Fake news detection, Logistic regression, TF-IDF vectorization.**

1. INTRODUCTION

Fake news detection has recently attracted a growing interest from the general public and researchers as the circulation of misinformation online increases, particularly in media outlets such as social media feeds, news blogs, and online newspapers. A recent report by the Jumpshot Tech Blog showed that Facebook referrals accounted for 50% of the total traffic to fake news sites and 20% total traffic to reputable websites. Since as many as 62% of U.S. adults consume news on social media (Jeffrey and Elisa, 2016), being able to identify fake content in online sources is a pressing need.

Social media and the internet are suffering from fake accounts, fake posts and fake news. The intention is often to mislead readers and or manipulate them in purchasing or believing something that isn't real. So a system like this would be a contribution in solving a problem to some extent.

As human beings, when we read a sentence or a paragraph, we can interpret the words with the whole document and understand the context. In this project, we teach a system how to read and understand the differences between real news and the fake news using concepts like natural language processing, NLP and machine learning and prediction classifiers like the Logistic regression which will predict the truthfulness or fakeness of an article.

2. LITERATURE REVIEWS

In general, Fake news could be categorized into three groups. The first group is fake news, which is news that is completely fake and is made up by the writers of the articles. The second group is fake satire news, which is fake news whose main purpose is to provide humour to the readers. The third group is poorly written news articles, which have some degree of real news, but they are not entirely accurate. In short, it is news that uses, for example, quotes from political figures to report a fully fake story. Usually, this kind of news is designed to promote certain agenda or biased opinion [1].

In the article published by Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu [2], they explored the fake news problem by reviewing existing literature in two phases: characterization and detection. In the characterization phase, they introduced the basic concepts and principles of fake news in both traditional media and social media. In the detection phase, they reviewed existing fake news detection approaches from a data mining perspective, including feature extraction and model construction.

Hadeer Ahmed, Issa Traore, and Sherif Saad [3] proposed in their paper, a fake news detection model that uses n-gram analysis and machine learning techniques. They investigated and compared two different features extraction techniques and six different machine classification techniques. Experimental evaluation yields the best performance using Term Frequency-Inverted Document Frequency (TF-IDF) as feature extraction technique, and Linear Support Vector Machine (LSVM) as a classifier, with an accuracy of 92%.

Perez-Rosas, Veronica & Kleinberg, Bennett and Lefevre Alexandra and Rada Mihalcea [4] in their publication "Automatic detection of fake news" focus on the automatic identification of fake contents in online news. For this they introduced two different datasets, one obtained through crowd sourcing and covering six news domains (sports, business, entertainment, politics, technology and education) and another one obtained from the web covering celebrities. They developed some classification models using linear sum classifier and five-fold cross verification with accuracy, precision and recall and FI measures averaged over the five iterations that rely on the combination of lexical, syntactic and semantic information as well as features representing text readability properties which are comparable to human ability to spot fakes.

E.M Okoro, B.A Abara, A.O. Umagba, A.A. Ajonye and Z. S. Isa [5] in their publication _A Hybrid approach to fake news detection on social media using a combination of both human-based and machine-based approaches. Since traditional and machine based approaches have some limitations and can't single handedly solve the problems like human literacy and cognitive limitations and the inadequacy of machine based approach. To solve all these problems, they proposed a Machine Human (MH) model for fake news detection in social media. This model combines the human literacy news detection tool and machine linguistic and network-based approaches. This way, the two parallel approaches of detection are at work, each helping to provide a balance for the other. The existing system and research work reveal that most classification algorithms perform well to detect or predict the fakeness of a news article. Though the logistic regression serves well for this purpose, our system is based on this information and thus we focus to work with classification algorithms like the logistic regression.

3. METHODOLOGY

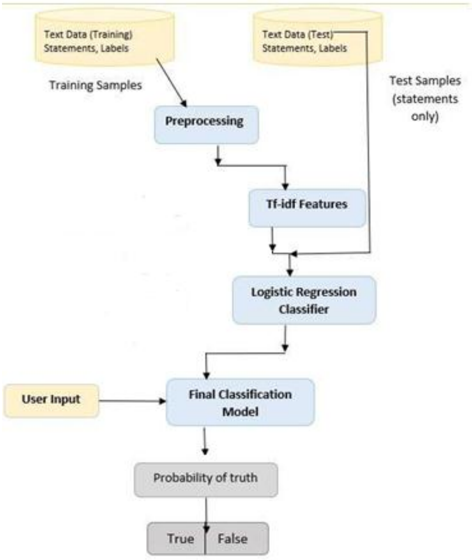

Fig 3.1: Flow chart of the proposed system

3.1 Data pre-processing

This module contains all the pre processing functions needed to process all the input documents and texts. First we read the train, test and validation data files then perform some pre processing like **tokenizing**, **stemming** etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

Stemming: In linguistic morphology and information retrieval, stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

Tokenizing: Tokenization is the process of replacing sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security. Tokenization, which seeks to minimize the amount of data a business needs to keep on hand, has become a popular way for small and mid-sized businesses to bolster the security of credit card and e-commerce transactions while minimizing the cost and complexity of compliance with industry standards and government regulations.

3.2 Feature Selection

In this module we have performed feature extraction and selection methods from sci-kit learn python libraries. For feature selection, we have used methods like simple bag-of-words and n-grams and then term frequency like tf-tdf weighting.

Count features:

The CountVectorizer provides a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary. You can use it as follows:

1. Create an instance of the *CountVectorizer* class.
2. Call the *fit()* function in order to learn a vocabulary from one or more documents.
3. Call the *transform()* function on one or more documents as needed to encode each as a vector.

An encoded vector is returned with a length of the entire vocabulary and an integer count for the number of times each word appeared in the document. Because these vectors will contain a lot of zeros, we call them sparse. Python provides an efficient way of handling sparse vectors in the scipy.sparse package. The vectors returned from a call to transform() will be sparse vectors, and you can transform them back to numpy arrays to look and better understand what is going on by calling the toarray() function.

3.3 Classifier

In this module we build all the classifiers for predicting the fake news detection. The extracted features are fed into different classifiers. We have used Logistic Regression classifier from sklearn. Each of the extracted features were used in the classifier. Once fitting the model, we compared the f1 score and checked the confusion matrix. After fitting all the classifiers, two best performing models were selected as candidate models for fake news classification. Finally selected model was used for fake news detection with the probability of truth. In Addition to this, we have also extracted the top 50 features from our term-frequency tfidf Vectorizer to see what words are most and important in each of the classes. We have also used Precision-Recall and learning curves to see how training and test set performs when we increase the amount of data in our classifiers.

Logistic regression Classifier:

It is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X.

4. CONCLUSION

In this paper, we've used Logistic Regression classifier which will serve the model and work with the user input. Here, we've presented a detection model for fake news using TF-IDF analysis through the lenses of different feature extraction techniques. We have investigated different feature extraction and machine learning techniques. The proposed model achieves accuracy of approximately 72% when using TF-IDF features and logistic regression classifier.

5. ACKNOWLEDGEMENT

We consider it as a privilege to articulate a few words of gratitude and respect to all those deserving individuals who guided us in this project. First and foremost, we would like to extend our profound gratitude and sincere thanks to our guide **Prof. Zameer Ahmed**, Department of computer science and Engineering, AITM Bhatkal who constantly supported and encouraged us during every step of dissertation. We really feel highly indebted to them for constantly guiding us to continue our work and giving us short term goals.

We are thankful to our project co-ordinator **prof. Bhagwat S G** and our HOD **prof. Anil Kadle** Department of Computer Science and Engineering, AITM, Bhatkal for their immense support.

We take this opportunity to thank **Dr.M.A Bhavikatti,** Principal, AITM Bhatkal for the encouragement and useful suggestions to pursue this work.

6. REFERENCES

[1] Schow, A.: The 4 Types of 'Fake News'. Observer (2017). http://observer.com/2017/01/ fake-news-russia-hacking-clinton-loss/

[2] Fake News Detection on Social Media: A Data Mining Perspective
Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu
Computer Science & Engineering, Arizona State University, Tempe, AZ, USA
Charles River Analytics, Cambridge, MA, USA
Computer Science & Engineering, Michigan State University, East Lansing, MI, USA

[3] Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques
Hadeer Ahmed, Issa Traore, and Sherif Saad
ECE Department, University of Victoria, Victoria, BC, Canada
School of Computer Science, University of Windsor, Windsor, ON, Canada

[4] Verónica Pérez-Rosas, Kleinberg Bennett, Alexandra Lefevre, and Rada
Mihalcea, ―Automatic detection of fake news,‖ *Proceedings of the 27th*
International Conference on Computational Linguistics, pp. 3391–3401,
Santa Fe, New Mexico, USA, 2018.

[5] E. M. Okoro, B. A. Abara, A. O. Umagba, A. A. Ajonye, and Z. S. Isa,
―A Hybrid Approach to Fake news detection on social media,‖ vol. 37,
no. 2, pp. 454-462, 2018.