IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11636

Fake News Detection Using Term Frequency Tokenization

Pallavi[1], Abhishek Shettigar[2], M Karunavathi[3], Ajith[4] , Mr Ramanath Kini M G[5]

Student, Department of Electronics & Communication Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, India[1,2,3,4]

Senior Assistant Professor, Department of Electronics & Communication Engineering,

Mangalore Institute of Technology & Engineering, Moodabidri, India[5]

Abstract: The fake news on social media and various other media is wide spreading and is a matter of serious concern due to its ability to cause a lot of social and national damage with destructive impacts. A lot of research is already focused on detecting it. This paper makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning algorithm, that can classify fake news as true or false, by using tools like python scikit-learn, NLP for textual analysis. This process will result in feature extraction and vectorization; we propose using Python scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorizer and Tfidf Vectorizer. Then, we will perform feature selection methods, to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results.

Keywords: Fake news, TF-IDF,SVM,feature extraction,training classifier.

I. INTRODUCTION

This Fake News contains misleading information that could be checked. This maintains lie about a certain statistic in a country or exaggerated cost of certain services for a country, which may arise unrest for some countries like in Arabic spring. There are organizations, like the House of Commons and the Crosscheck project, trying to deal with issues as confirming authors are accountable. However, their scope is so limited because they depend on human manual detection, in a globe with millions of articles either removed or being published every minute, this cannot be accountable or feasible manually. A solution could be, by the development of a system to provide a credible automated index scoring, or rating for credibility of different publishers, and news context. This paper proposes a methodology to create a model that will detect if an article is authentic or fake based on its words, phrases, sources and titles, by applying supervised machine learning algorithms on an annotated (labeled) dataset, that are manually classified and guaranteed. Then, feature selection methods are applied to experiment and choose the best fit features to obtain the highest precision, according to confusion matrix results. We propose to create the model using different classification algorithms. The product model will test the unseen data, the results will be plotted, and accordingly, the product will be a model that detects and classifies fake articles and can be used and integrated with any system for future use.

II. METHODOLOGY

Data Preprocessing

There are some exploratory data analyses is performed on training data to prepare the data for modelling of system like null or missing values, removing social media slags, removing stop-words, correcting contraction. Also, Part of Speech (PoS) Tagging has been performed in the data to meet the accuracy of prediction model. Data has been also lemmatized to get root form of the words so that prediction algorithm gets trained on the quality data. Before model training the data was tokenize so that each word in the sentence can treat as element for model training.

Lemmatization: Lemmatization is one of the most common text pre-processing techniques used in Natural Language Processing (NLP) and machine learning in general. lemmatization involves deriving the meaning of a word from something like a dictionary. Lemmatization gives the root form of the word i.e., studying is studies. This makes sure that the root word is not just achieved by removing the suffix from a given word that is done in stemming. This makes the lemmatization algorithm slow but for the NLP technique where meaning each word is equally important by its root word. Thus, we have used lemmatization to get the root word.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11636

Tokenization:

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens. Here, tokens can be either words, characters, or sub-words as tokens are the building blocks of Natural Language, the most common way of processing the raw text happens at the token level. Hence, Tokenization is the foremost step while modelling text data. Tokenization is performed on the corpus to obtain tokens. The following tokens are then used to prepare a vocabulary. Vocabulary refers to the set of unique tokens in the corpus. Remember that vocabulary can be constructed by considering each unique token in the corpus or by considering the top K Frequently Occurring Words. TF-IDF: TF-IDF stands for "Term Frequency — Inverse Document Frequency". This is a technique to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining. TF is individual to each document and word; hence we can formulate TF as follows

$$\mathrm{TF-IDF} = \frac{\text{Count of t in d}}{\text{number of words in d}}$$

This measures the importance of document in whole set of corpus, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N.

$$\mathrm{df(t)} = \text{Occurance of t in documents}$$

IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as "is" is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

$$\mathrm{TF-IDF(T)} = \mathrm{N/df}$$

CLASSIFICATION: SUPPORT VECTOR MACHINE (SVM)-

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of Support Vector Machine (SVM) algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

MODEL TRANING

In this module the extracted features are fed into different classifiers. We have used support vector machine to learn. This computational tool uses two different classifiers so that user could get more accurate results the prediction of both the classifier are shown the output page.

FEATURE SELECTION

In this module we have performed feature selection methods from sci-kit learn python libraries. For feature selection, we have used methods like count Vectorization term frequency like TF-IDF weighting. TF-IDF (compute a weight to each word which signifies the importance of the word in the document and corpus). Count Vectorization (Vectorization is a process of converting the text data into a machine-readable form).

III. CONCLUSION

In this paper, we've used support vector machine which will predict the truthfulness of user input news, here we have presented a prediction model with feature selection used as Count Vectorization, TF-IDF which helps the model to be more accurate.

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11636

REFERENCES

1. G. Bharath, K. J. Manikanta, G. Bhanu Prakash, R. Sumathi and P. Chinnasamy, "Detecting Fake News Using Machine Learning Algorithms," 2021 International Conference on Computer Communication and Informatics (ICCCI - 2021), Jan. 27 – 29, 2021, Coimbatore, India form IEEE Xplore. DOI: 10.1109/ICCCI50826.2021.9402470.

2. Uma Sharma, Sidarth Saran and Shankar M. Patil, "Fake News Detection using Machine Learning Algorithms," 2021 International Journal of Engineering Research & Technology (IJERT), from IEEE. Aditi Vora and Narendra Shekokar "Fake News Detection Using Intelligent Techniques", 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), from IEEE.DOI: 10.1109/ICRITO51393.2021.9596438

3. Saeed Amer Alameri and Masnizah Mohd "Comparison of Fake News Detection using Machine Learning and Deep Learning Techniques", 2021 3rd International Cyber Resilience Conference (CRC) from IEEE Xplore. DOI: 10.1109/CRC50527.2021.9392458

4. Mohd Abbad, Gaurav Kumar, Md Samiullah and N. Suresh Kumar " A Predominant Advent to Fake News Detection using Machine Learning Algorithm ", 2021 International Conference on Intelligent Technologies (CONIT), from IEEE Xplore.DOI: 10.1109/CONIT51480.2021.9498436

5. Rahul R Mandical, Mamatha N, Shivakumar N, Monica R and Krishna A N, "Identification of Fake News Using Machine Learning," 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) from IEEE Xplore.DOI: 10.1109/CONECCT50063.2020.9198610

6. Jasmine Shaikh, Rupali Patil, "Fake News Detection using Machine Learning," 2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC) from IEEE Xplore. DOI: 10.1109/iSSSC50941.2020.9358890 Nihel Fatima Baarir and Abdelhamid Djeffal, "Fake News detection Using Machine Learning" 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH) from IEEE Xplore. DOI: 10.1109/IHSH51661.2021.9378748

7. Dinesh Kumar Vishwakarma and Chhavi Jain, "Recent State-of-the-art of Fake News Detection: A Review," 2020 International Conference for Emerging Technology (INCET) Belgaum, India from IEEE.Data Sciences and Analytics Group, National Security Directorate Pacific Northwest National Laboratory. IEEE access 2020.

8. Smitha N and Bharath R. "Performance Comparison of Machine Learning Classifiers for Fake News Detection", Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020), from IEEE Xplore Part Number:CFP20N67-ART ISBN: 978-1-7281-5374-2

9. Wenlin Han and Varshil Mehta, "Fake News Detection in Social Networks Using Machine Learning and Deep Learning: Performance," 2019 International Conference on Industrial Internet (ICII) from IEEE Xplore. DOI: 10.1109/ICII.2019.00070 .David Mathew Thomas, Sandeep Mathur Amity Institute of Information Technology Amity University (AUUP), "Data Analysis by Web Scraping using Python," Proceedings of the Third International Conference on Electronics Communication and Aerospace Technology [ICECA 2019] IEEE Xplore 2019.

10. Syed Ishfaq Manzoor Dr. Jimmy Singla and Nikita "Fake News Detection Using Machine Learning approaches: A systematic Review", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019) IEEE Xplore. DOI: 10.1109/ICOEI.2019.8862770

11. Karishnu Poddar , Geraldine Bessie Amali D, Umadevi K S "Comparison of Various Machine Learning Models for Accurate Detection of Fake News" , 2019 Innovations in Power and Advanced Computing Technology (i-PACT), from IEEE Xplore.DOI: 10.1109/i-PACT44901.2019.8960044