Fake News Classification By Using Machine Learning

[1]V.Poojitha Govind [2]M.Subashini

[1]V.Poojitha Govind, Department of CSE, Narayana Engineering College, Gudur

[2]M.Subashini Asst. Professor, Department of CSE, Narayana Engineering College, Gudur

ABSTRACT :

Social media is a popular medium for the dissemination of real-time news all over the world. Easy and quick information proliferation is one of the reasons for its popularity.An extensive number of users with different age groups, gender, and societal beliefs are engaged in social media websites. Despite these favorable aspects, a significant disadvantage comes in the form of fake news, as people usually read and share information without caring about its genuineness. Therefore, it is imperative to research methods for the authentication of news. To address this issue, this article proposes a two-phase benchmark model named WELFake based on word embedding (WE) over linguistic features for fake news detection using machine learning classification.

The first phase preprocesses the data set and validates the veracity of news content by using linguistic features. The second phase merges the linguistic feature sets with WE and applies voting classification. To validate its approach, this article also carefully designs a novel WELFake data set with approximately 72 000 articles, which incorporates different data sets to generate an unbiased classification output.

KEYWORDS : MachineLearning;NaturalLanguage;ClassificationTechniques.

1. INTRODUCTION

1.1Background Study:

Nowadays people around the world are getting much involved on online social networks regardless of age, community, or sex. Communicating using social networks is simple, fast, and attractive to share and transfer information. Currently, social network sites like Facebook trailed by Twitter are the market pioneers, facilitating over 1.3 billion clients with a dynamic monthly variation of 300 million users in average. Their collaborations generate Terabytes of information every second . Online social networks are attractive because of the simple and convenient way to access and circulate information with other people. However, the fast scattering of data at a high rate with minimal effort enables the

widespread of false information, such as fake news, which are harmful to society and people.

Fake news are low-quality information with purposefully false data, propagated by individuals or bots that deliberately manipulate message for tattle or political plans.

Schudson and Zelizer claimed that the term "fake news" originated in previous centuries together with the mass media itself. Nevertheless, this term attracted increased attention after the U.S. presidential elections of 2016, when the propagation of fake news on social media pulled the attention of a larger number of online users than traditional newsreaders. In the last five months before the elections, approximately 7.5 million tweets contained a link to exceptionally one-sided or false news websites. An interesting and worrying aspect is that false and unsubstantiated news from doubtful sources attracts more audiences than credible information. Relevant work on this topic concluded that fake news spread quicker, penetrate further, and have a deeper impact than true news. There are numerous cases where people accept and spread news without checking their correctness certified by sources.

2. RELATED WORK:

2.1 Over view :

This survey is an analysis of distinctly assorted systems or techniques that are being used previously for Fake News detection. The primary objective of this paper is to observe and determine most efficient and non-biased techniques for stated problem statement. Also, following survey explores every methodology implemented among mentioned Literatures (see References). The prominent causes and prevalence of fake news are perplexing issues. There are numerous approaches that can and had been embraced by individuals as well as organizations [5].

How ever in our survey, it is observed that Prominence regarding this approaches are given to (1) Fact Checking, (2) Rumor detection, (3) Stance Detection, and(4)Sentiment Analysis. In [14], A sentiment analysis is done to detect Fake News with the help of Neural Networks. More-less this procedure is followed in other surveyed literatures irrespective of approaches, tools and resources utilised. Hence it is observed that Machine Learning is a common domain for text analysis. Therefore it seems, a fake news detector is an informally titled data science implementing model which is capable of detecting and classifying fake and true news from provided data. Neural networks are incapable of estimating text driven data and hence requires word embedding [11]. TF-IDF, Fast Text, Bag of Word (BOW) and Word2Vec are frequently appeared across this survey. Addition to this [14] introduces flair library for NLP.

The Problem of news detection is classification oriented specifically binary classification, so machine learning algorithms such as logistic Regression, Supported Vector Machine (SVM), and Naïve bayes are utilized more often. However in following survey it can be noted these algorithms

are not very lenient on varying data and hence do not seems to provide required accuracy. New methods such as Deep Learning and Natural language processing are explored to provide solution.

METHODOLOGY:

In this project, we propose a machine learning-based approach for fake news classification, using a combination of natural language processing (NLP) techniques and classification algorithms.

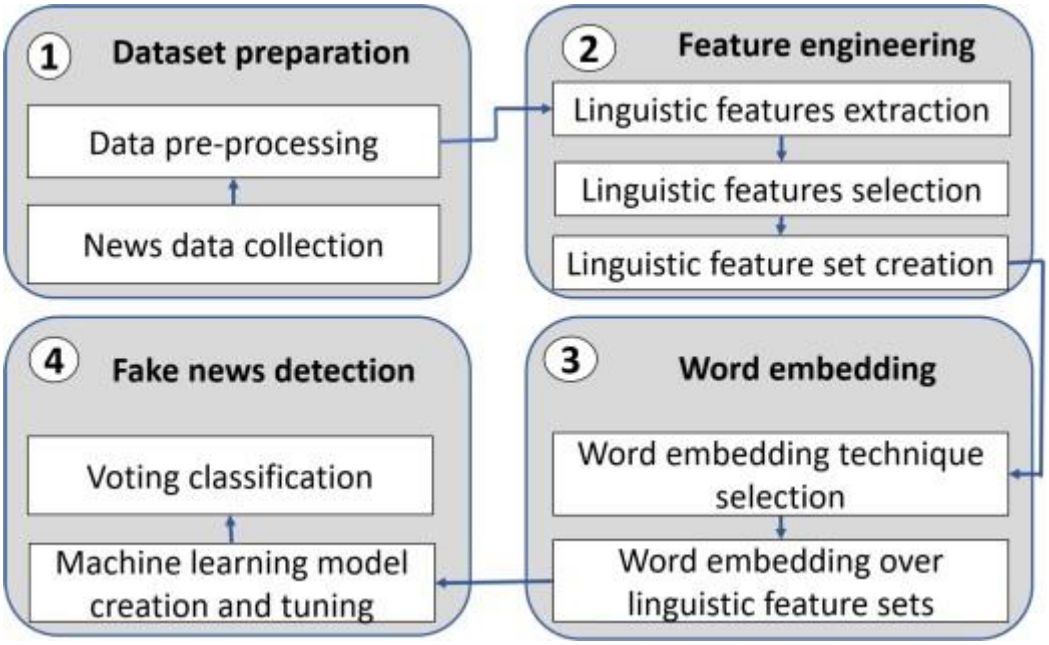
We first preprocess the raw news articles and extract relevant features using NLP techniques such as tokenization, stemming, and part-of-speech tagging.

3. SYSTEM DESIGN:

3.1 System Architecture:

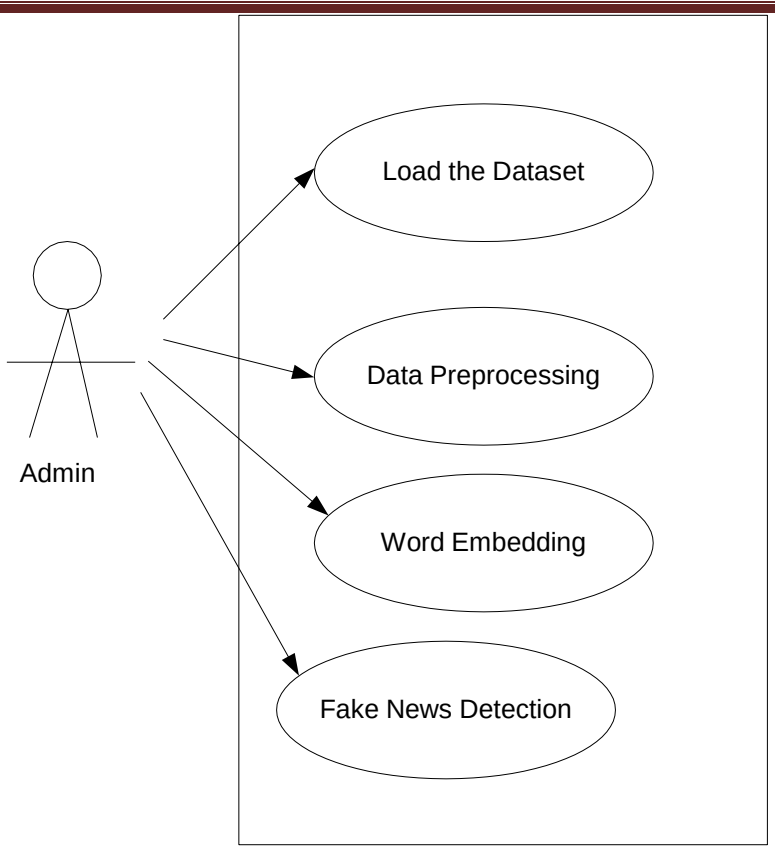
System architecture is the conceptual design that defines the structure and behavior of a system. An architecture description is a formal description of a system, organized in a way that supports reasoning about the structural properties of the system. It defines the system components or building blocks and provides a plan from which products can be procured, and systems developed, that will work together to implement the overall system.

The System architecture is shown below:



3.2 Use case Diagram of the system :

A use case diagram is a type of behavioral diagram created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases.

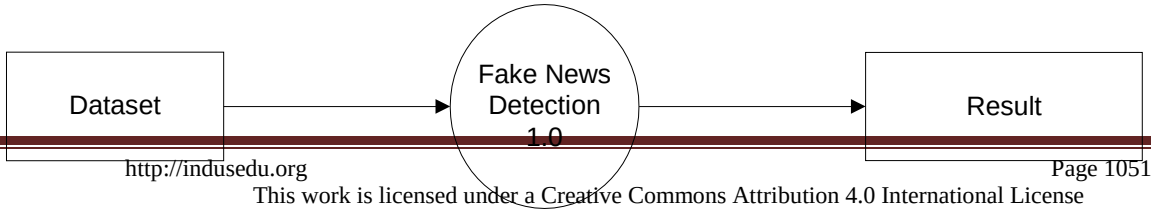


3.3 Data Flow Diagram of the system:

An information stream outline (DFD) is a graphical representation of the "stream" of information through a data framework. DFDs can likewise be utilized for the perception of information handling (organized outline). On a DFD, information things stream from an outside information source or an inner information store to an interior information store or an outer information sink, through an inward process.

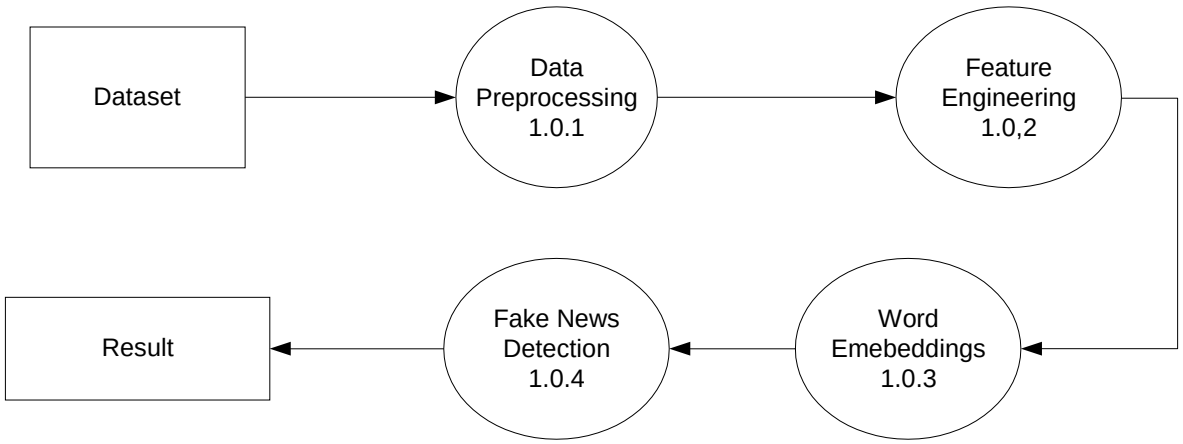
3.3.1 Level 0 Data stream chart :

A connection level or level 0 information stream chart demonstrates the collaboration between the framework and outside specialists which go about as information sources and information sinks. On the connection chart (otherwise called the Level 0 DFD) the framework's associations with the outside world are displayed simply as far as information streams over the framework limit. The connection chart demonstrates the whole framework as a solitary process, and gives no pieces of information as to its inward association.



3.3.2 Level 1 Data flow diagram:

The Level 1 DFD shows how the system is divided into sub-systems (processes), each of which deals with one or more of the data flows to or from an external agent, and which together provide all of the functionality of the system as a whole. It also identifies internal data stores that must be present in order for the system to do its job, and shows the flow of data between the various parts of the system.



4. Data Preprocessing:

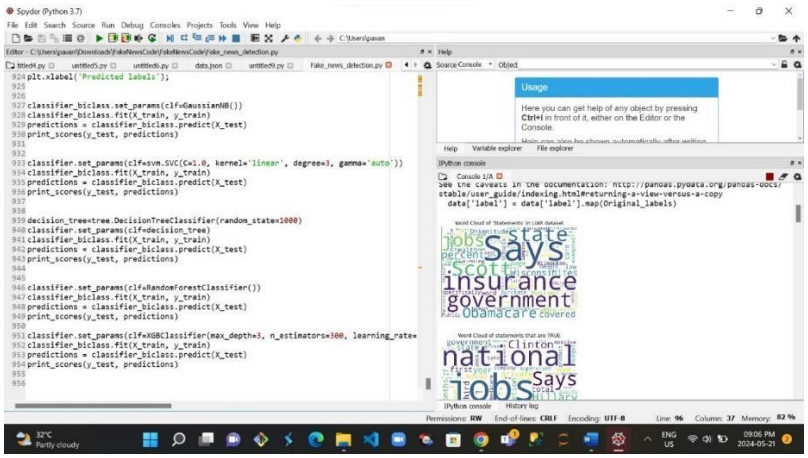
The main goal of this part is to use NLP techniques to preprocess the input data and prepare for the next step to extract the proper features.

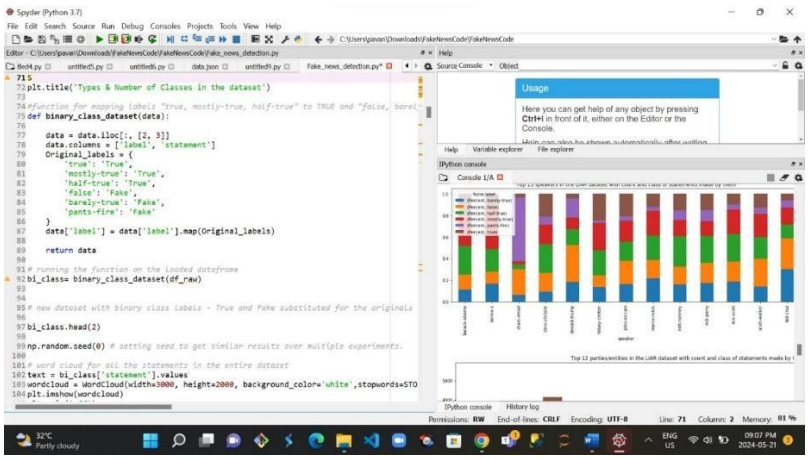
The data we use contains news titles and texts. Each of the titles is about 12.45 words long, while each of the texts is about 405.28 words long. In our project, we only use the titles for the fake news detection because the texts are too large for us to train efficiently. Also, the text contains too many details and information for a piece of news, which may distract the models during training.

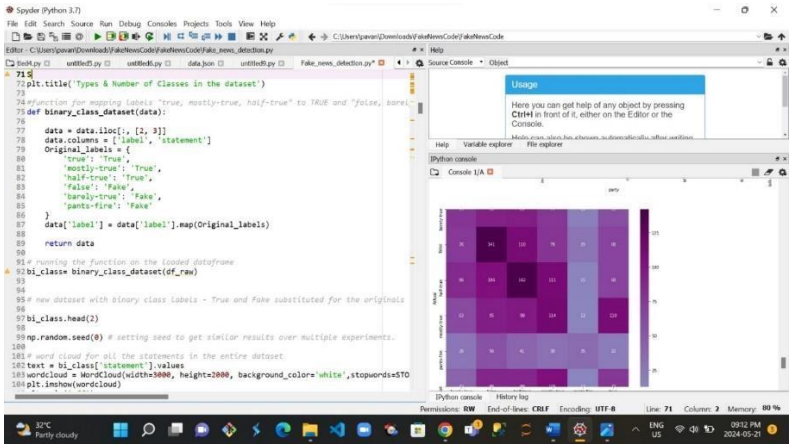
4.1 Word Embedding:

This part is important because we need to convert the dataset into a form that models can handle. We use different types of word embedding for different models we built. For LSTM, Bidirectional LSTM, and CNN, we first create a Tokenizer to tokenize The words and create sequences of tokenized words. Next, we zero-padded each sequence to make the length of it 42. Then, we utilized the Embedding layer that initialized with random weights to let it learn an embedding for all of the words in the training dataset. The Embedding layer [10] will convert the sequence into a distributed representation, which is a sequence of dense, real-valued vectors. For BERT, we utilized BERT tokenizer to tokenize the news titles in the dataset first.

6. RESULTS AND ANALYSIS:







7. CONCLUSION:

We presented a new model called WELFake for text fake news detection. For this purpose, we prepared a larger data set called WELFake with over 72 000 news articles combining four open-source data sets (i.e., Kaggle, McIntire, Reuters, and BuzzFeed) to reduce their individual limitation and bias. Afterward, we analyzed over 80 linguistic features from state-ofthe-art works and selected

20 significant ones to minimize the computational complexity and increase the standard classifiers' accuracy. We applied two WE-based methods (i.e., TF-IDF, CV) over these linguistic features using six ML models (i.e., KNN, SVM, NB, DT, Bagging, and AdaBoost) and found out that CV produces better overall accuracy than TF-IDF with an SVM model. We, therefore, used CV over LFS and classified the 20 features based on four categories: writing pattern, readability index, psycho-linguistics, and quantity. As the number of predictors that participate in the voting classifier needs to be odd, we prepared three LFS by distributing the twenty selected features in a balanced manner across these categories. Afterward, we embedded CV with these LFS and applied all six ML models. We determined the most accurate ML model and took its predicted results from each WE-enabled LFS data set for voting classification. We finally applied the result of this voting classifier to the next level voting classification with the best model results of TF-IDF and CV over LFS and obtained the final classification. Experimental results show that the WELFake model produces a high 96.73% accuracy on the WELFake data set. To further analyze its advantage we compared it with two state-of-theart works and found out that it improves the overall accuracy by 1.31% compared to BERT and 4.25% compared to CNN models. The proposed WELFake model also improved the accuracy by up to 10% on the McIntire and BuzzFeed data sets [37]. We also analyzed the performance of different ML models in terms of accuracy, precision, recall, and F1-score, and found out that SVM produced the most accurate results. Finally, our frequency-based model focused on analyzing writing patterns outperformed predictive-based related works implemented using the Word2vec WE method by up to 1.73%.

We plan to extend our work in the future with other factors like knowledge graphs and user credibility for further verification of the output generated by the WELFake model.

8. REFERENCE:

[1] Ajeet Ram Pathak, Aditee Mahajan, Keshav Singh, Aishwarya Patil, Anusha Nair: "Analysis of Techniques for Rumor Detection in Social Media", International Conference on Computational IntelligenceandDataScience(ICCIDS),2019,volume167,pp.2282296,DOI:10.1016/j.procs.2020.03.281.

[2]Aswinithota,PriyankaTilak,SimratAhluwalia,NibratLohia:"FakeNewsDetection:ADeepLearnin gApproach",S MUDataScienceReview,2018,volume-1,issue-3,Article-10.

[3] Caio V. Meneses Silva, Raphael Silvsa Fontes, Methanias Colaco Junior: "Intelligent Fake News Detection: A Systematic Mapping", Journal of Applied Security, 2020,nDOI:10.1080/19361610.2020.1761224.

[4] Dong-Ho Lee, Yu-Ri Kim, Hyeong-Jun Kim, Seung-Myun Park, Yu-Jun Yang: "Fake NewsDetectionusingDeepLearning",JournalofInformationProcessingSystems(JIPS),2019,volume-15,issue5,pp.1119-1130, DOI: 10.3745/JIPS.04.0142.

[5] Fatmeh Torabi Asr, Maite Taboada:"Big Data and quality data for Fake News and misinformation detection", Big Data & Society, 2019, Article-14, DOI:10.1177/2053951719843310.

[6] Harita Reddy, Namratha Raj, Manali Gala, Annappa Basava: "Text-mining-based Fake News Detection Using Ensemble Methods", International Journal of Automation and Computing,DOI:10.1007/s11633-019-1216- 5.

[7] James Bradbury, Stephen Merity, CaimingXiong, Richard Socher: "Quasi-Recurrent Neural Networks", International Conference of Learning Representations(ICLR),2017.

[8] MeichangGuo, ZhiweiXu, Limin Liu, MengjieGuo, and Yujun Zhang: "An Adaptive Deep Transfer Learning Model for Rumor Detection without Sufficient Identified Rumors", Mathematical Problems in Engineering, 2020, article ID-7562567, DOI:10.1155/2020/7562567. [

9] PranavBharti,MohakBakshi,R.AnnieUthra:"FakeNewsDetectionUsingLogistic

[10] Regression, Sentiment Analysis and Web scrapping", International Journal of Advanced Science and Technology (IJAST),2020,volume-29,issue-9,pp.115–1167.

[11] Pritika Bahad, Preeti Saxena, Raj Kamal: "Fake News Detection using Bi-Directional LSTM-Recurrent Neural Network", International Conference on Recent Trends in Advanced Computing (ICRTAC), 2019, volume-165,pp.74-82,DOI:10.1016/j.procs.2020.01.072.

[12] Ray Oshikawa, Jing Qian, William Yang Wang: "A Survey on Natural Language Processing for Fake News Detection", Language Resources and Evaluation Conference (LREC), 2020,volume-12.

[13] Rohit KumarKaliyar, Anurag Goswami, Pratik Narang: "Deep Fake: Improving fake news detection using tensor decomposition - based deep neural network", The Journal ofSupercomputing,2020.

[14] SachinKumar, Rohan Asthana, Shashwat Upadhyay, Nidhi Upreeti, Mohammad Akbar: "Fake News Detection using Deep Learning models: A Novel Approach", Transactions on Emerging Telecommunication Technologies, 2019, volume-31, issue-2, DOI:10.1002/ett.3767.

[15] Sebastian Kula, Michal Choras, Rafal Kozik, Pawel Ksieniewicz, Michal Woznaik: "Sentiment Analysis for Fake News Detection by Means of Neural Networks", International Conference on Computational Science (ICCS), 2020, volume-12140, pp.653-666, DOI:10.1007/978-3-030-50423-6_49.

[16] Tejaswini Yesugade, Shrikant Kokate, Sarjana Patil, Ritik Varma, Sejal Pawar: "Fake News and Deep Fake Detection", Journal of Critical Reviews, 2020, volume-7, Issue-19.

[17] Xinyi Zhou, Reza Zafarani: "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities", ACM Computing Surveys, Article-109,DOI:10.1145/3395046.