FALSE DATA RECOGNITION EMPLOYED BY LOGISTIC REGRESSION AND NATURAL LANGUAGE PROCESSING TECHNIQUES

Dr.BUCIO PITTY*

ABSTRACT

Fake news, often referred to as junk news or pseudo-news, is a form of yellow journalism or propaganda created with the purpose of distributing deliberate disinformation or false news using traditional print or online social media. Fake news has become a significant problem globally in the past few years. It has become common to find popular individuals and even members of the state using misinformation to influence individuals' actions whether consciously or sub-consciously. This project report discusses the various approaches we have researched, evaluated and employed by using natural language processing, machine learning and deep learning to address the fake news problem. Using evaluation metrics such as precision/accuracy/recall/F1-scores, we would then develop classification workflows which involve machine learning and deep learning based classifiersto improve on the metrics to achieve a sustainable and resilient fake news detection text- based classifier.

Keywords: False news, discovery, Logistic Regression, Natural Language Processing.

1. INTRODUCTION

According to statistics published on statisa.com, it has been calculated that 62% of adultsbelieve that fake news is prevalent on online news websites and platforms, compared to TV/radio sources at 52% and newspapers/magazines at 52%. The demographic of this study spans across 27 countries among 19541 respondents; age group 16-74 years (Watson,A. 2021).

Current fake news detection models or tools aim to replace human judgment instead of being augmented to be a collaborative tool in making the decision. Hence it is important to consider and address the above public responses as it emphasizes on the issue of fake news and the necessity to train better and improved models to effectively achieve an accurate detection work flow of the several forms of fake information that is being published every second, every day through various online channels. The motivation behind undertaking this project is that the leading limitation with currently

deployed NLP pipelines and classification models is that insufficient or inaccurate labelled datasets and categorizations are used to train and test news data as an attempt to optimize fake news detection workflows.

The concept of this project starts off by firstly focusing on evaluating currently developed and available natural language processing pipelines (n-grams, count vectorization, Bag odf Words etc.) and classification models (Naïve Bayes, SVM, Random Forest) based on set evaluation metrics. With recall, precision, accuracy and F1-scores, we would be able to analyze and develop baseline creations to initiate a benchmark of the classification models which can be used to train and test the data on and implement improved variations of data processing pipelines on the chosen dataset, efficient machine learning or deep learning based algorithms which are able to handle huge data pre-processing or NLP pipelines.

The final deliverable has been targeted for me is to output a software module embodied as a Jupyter Notebook using Python. However, we would like to put forth here that to extend my interest in this project and explore newer scope of work; we would also be aiming to develop a Chrome extension which can be deployed to detect fake news on Chrome browsers. The fake news detection algorithm we am hoping to deploy within the Chrome extension will be based on my research and findings that I build and develop upon over the course of this final year project.

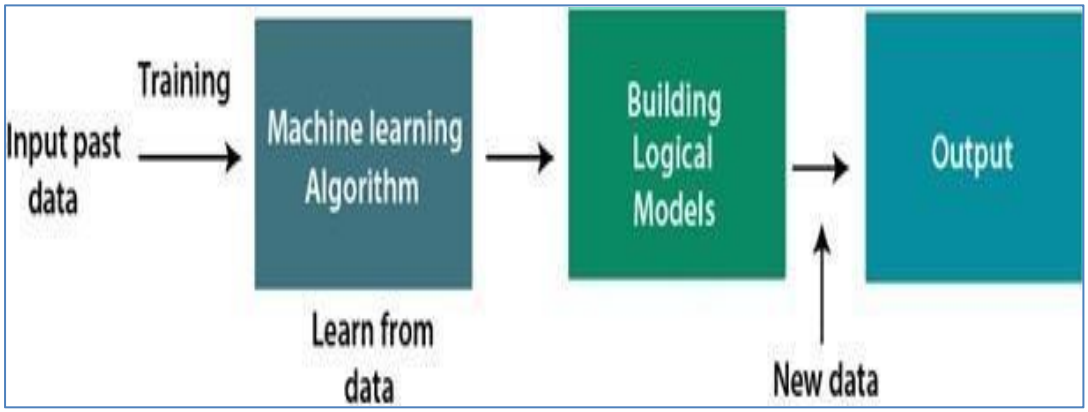

Figure 1: working of Machine Learning algorithm

2. LITERATURE REVIEW

In this section, we are going to explore my research and evaluation on the available related work in the field of fake news detection. This is such that we would be able to identify and apprise myself of knowledge gaps that we may possess, which could limit the

theoretical concepts and perspectives we could make use of to further my research and application in this project. Firstly, we have chosen to explore a similar project to the project we have chosen to undertake. Next, we will be evaluating a study assessing the significance and effectiveness of big data and the quality of the data being used for fake news detection in the current digital landscape. Finally, we have conducted critical evaluation on the NLP techniques and models that we plan to employ in this project, hence requiring reviews on the relevant methodologies, code documentation and libraries which would help support the algorithms we plan to develop.

The first paper we are going to be looking at is part of a project which was published by Rohit Kumar Kaliyar, Anurag Goswami and Pratik Narang in May 2020, as part of a published journal known as The Journal of Supercomputing (Kaliyar. 2020). The principal concept of the DeepFakE project registers the significance of the relationships that can be made between content based features found in news articles and the features which encompass the social networks that the said news articles are being propagated through to spread fake information. Kaliyar et al. also explore the different approaches to fake news detection, namely learning, knowledge and feature based approaches. Figure 1 below outlines the various approaches that they have considered as part of their project (Kaliyar. 2020).

3. MODEL DEVELOPMENT

Instrumental to developing an accurate model to address the fake news detection problem we have posed in this paper:

a. Feature Extraction
b. Classification Algorithm Analysis
c. Model Performance Evaluation

Feature extraction techniques can be generally divided into 3 core levels:

Features which relate to the characteristics of the digital circles and groups which are formed under a social agreement based on common interests, inclinations and topics. The echo chamber and filter bubble concepts I haveexplained in Section 2 plays an important role here, where data can be analyzed based how users are consuming and spreading the fake information that they encounter.

Features which relate to the demographics of entities/individuals which interact with fake news data in the digital landscape, be it news media channels, or social media platforms.

Features which relate to the lexical and linguistically features of the text content found in the bodies of news articles and headlines. These are the most powerful and weight age heavy features to fake news detection due to their importance in text classification processes.

Thanks to the literature review we have conducted we have gained a better understanding of the benefits and drawbacks that arise due to the collaboration of social and user features from the Kaliyar et al. paper from the published journal source (Kaliyar.2020). Content features provide a more direct textual and linguistically approach to feature selection and model construction. There are several methods we can use to engineer features into vectors from the raw dataset; as inputs for model construction:

1. Count Victimization
2. Word Embedding
3. Linguistically Analysis using NLP
4. TF-IDF Victimization
5. Topic Modeling

The relevant tools for the above techniques can be generally found in the sklearn library

4. METHODOLOGY

In order to develop a software module which is able to perform fake news detection, a very important step is to execute the text classification phase. There is a huge pool of machine learning models which we can use to train classifiers with the use of the feature data we have extracted in the precedent phase. Machine Learning Classifier Examples:

1. Support Vector Machine (SVM) Classifier
2. Naïve Bayes Classifier
3. Passive Aggressive Classifier
4. Bidirectional Encoder Representation from Transformers (BERT)
5. Deep Neural

Networks Deep Neural

Network Examples

1. Convolution Neural Networks (CNN)
2. Recurrent Convolution Neural Networks (RCNN)

3. Long-Short Term Model (LSTM)

We will be expanding more on these classification models and deep neural networks as part of my project structure and methodology in the following sections. These models can mostly be found from the sklearn library based on Python. As for the Deep Neural Networks, depending on the case, we would be able to access the relevant libraries frompublicly available sources.

4.1 Dataset Acquisition and Analysis

The dataset we are using for this project has been ethically considered and appropriately sourced from Kaggle under the public domain (Kajal Yadav [Kaggle]. 2020).

1. Text corpus of news articles and metadata scraped across 600 webpages
2. Approximately 10K news article data and metadata

3. 6 feature attributes: News Headline, News URL, Source, Stated On, Date, Label
4. Multi-Class Labelling: False, Pants on Fire, True, Mostly True, Barely True, Half True
5. Licensing Type: Public Domain (Creative Commons License: CC BY-SA 4.0) (Kajal Yadav [Kaggle]. 2020)

Once this step is accomplished, the project design is going to require the dataset to be split into 2 pipeline versions.

1. Non web scraped, news headline Fake-Real News Dataset
2. Web scraped, news headline + news content Fake-Real News Dataset Further details and outline will be provided in Chapter 4, Implementation.

There are going to be 2 main pre-processing pipelines using NLP and other text data cleaning techniques to prepare the raw data to be input into classification models and workflows for this paper.

1. Simple NLP
2. Extensive Data Cleanse NLP (EDC-NLP)

The Simple NLP pipeline is going to be representing a rudimentary as needed basis of pre-processing characteristic, with simple case conversions, lemmatization and stop word

removals.

To evaluate the performance and effectiveness of the implemented models effectively and sustainably among the various workflows in this project, it is important to have a proper evaluation system to compare the accuracies, recall and precision of the text classification approaches.
The standard library that can achieve this sklearn, using the sklearn metrics module. It enables us to calculate and plot confusion matrices, calculate accuracy, recall, precision and F1 scores.

Accuracy scores would quantify and measure the number of news data points that the classification model in question has predicted as fake or real news as exactly as what is represented in the test data set. Recall scores measure the ratio of the number of correctly predicted news data points to the sum of correctly predicted real news and wrongly predicted fake news. This helps us evaluate the ability of the classification model in question to correctly detect all real news data points. Precision scores would measure the ratio of correctly predicted real news data points to the sum of correctly and wrongly predicted real news data points. This helps us evaluate the ability of the classification model in question to avoid classifying a fake news data point as real news.

Using these metrics, we would be able to tabulate and compare the effectiveness and robustness of the classification models we propose to implement as part of this paper. Processes able to efficiently tackle the data point necessities for text classification while preserving the context of the news data.

The analysis of the news headline and news content features of the dataset has posed somepre-processing requirements as follows:

1. Removal of →
   a. Links, White spaces, Newlines, Tabs, Accented Characters, SpecialCharacters, Stop words, HTML tags
   b. Conversion of resultant text to lower case text
2. Reduction of →

a. Repetitive Characters, Punctuations

3. Expansion of contracted words

Spelling Correction using Auto correction Python module

4.2 The fundamental workflow for each of the proposed model approach is as follows

1. Importation of relevant libraries
2. Dataset Pre-Processing
3. Feature Extraction
4. Input Generation
5. Model Construction and Architecture Analysis
6. Model Prediction
7. Model Performance Evaluation

5. RESULTS AND DISCUSSION

5.1 Baseline Model Approach: SimpleNLP + TF-IDF + Multinomial NaïveBayes Classifier

Term Frequency refers to the frequency of a certain word which appears in a document, ortext body (Sammut C. 2011). Inverse Document Frequency refers to the frequency of the occurrence of a certain word amongst an entire collection of documents, or text corpus(Sammut C. 2011).

Using this feature extraction technique, we will be training and testing the cleansed data on the Naïve Bayes classifier as part of my baseline implementation. The predictions willbe evaluated according to Model Performance Evaluation metrics.

5.2 Alternative Approach: Simple NLP + TF-IDF + Passive Aggressive Classifier

This classifier ultimately works as a family of algorithms based on the Perception model (T. Zhai. 2022), and is usually used in social media companies such as Twitter (T. Zhai.2022),where there is information and data being constantly in fluxed into the frame work and an algorithm which is able to handle the huge data in flow is required.

The Passive Aggressive classifier foundationally works such that if the prediction of theclassifier is correct, the machine learning model remains unchanged and if the

prediction is otherwise, then the classifier alters the model to achieve a more accurate prediction.

5.3 BERT Classification Approach (Using EDC-NLP Pipeline)

The BERT (Bi-Directional Encoder Representation from Transformers) approach is a slightly more unique approach. This partially supervised learning model can comprehend and analyze the context and relationships between words in a text string, or sentence structures amongst a huge text corpus.

We propose to use the BERT (Base) model, which has 12 Encoding layers in its stack. Using the BERT (Base), we will automatically tokenize the text corpus of the news headlines
Being input and pass the input through the encoder layers to perform the classificationof the fake/real news data.

5.4 Deep Learning Approach using Neural Networks

Deep neural networks are networks with a complex architecture, where obscure layers perform extremely complicated tasks. Using deep neural networks in the field of fake newsdetection has proven to be a heavily researched and experimental task. For this project, we propose to implement 2 familiar, different types of neural networks, which are expanded on as shown below.

5.5 Deep Neural Network Approach – EDC-NLP + CNN Deep Learning Model

For this approach, we are going to adopt a convolution neural network (CNN) approach on the dataset which has been cleaned using the EDC-NLP process. Even though, convolution neural networks are usually used for image and video recognition, my decision to use the CNN approach to detect fake news would be as such:

1. After the word embedding process, CNN networks are optimal in extracting important features and conducting supervised learning on those vector data inputs
2. CNNs are able to provide a dense network along with the pooling layer

architecture hence increasing the efficiency of feature recognition and hence prediction amongst large datasets.

There is going to be a sequential compilation of spatial dropouts, convolution 1-Dimensional layers and Dense layers which will employ the use of a binary cross entropy loss function as it is a binary classification dataset and focus on the accuracy metrics toisolate the effectiveness of the module in detecting fake news.

5.6 Deep Neural Network Approach – EDC-NLP + RCNN + LSTM Deep Learning Model

Using deep learning neural networks such as these enable us to construct varied model architectures to optimize the text classification workflow process in project designs. For the second deep learning approach, we propose to implement a Recurrent Convolution Neural Network (RCNN) collaborated with a Long Short Term Memory (LSTM) network.

My decision to choose the above 2 configurations to build and compile a Deep Learningnetwork would be as such:

3. RCNN models are adept at learning contextual information, which is important in our case of distinguishing the text contained within several news articledata points

LSTM networks are effective with modeling sequential data points and hence good at handling long scope dependencies

However, where we are going to develop on these networks are the variations of the layers and layer implementation structures across the network model architecture, with the goalof achieving the best possible accuracy in detecting fake news from a dataset.

5.7 Baseline PRO Model Approach: EDC-NLP + TF-IDF + Multinomial NaïveBayes Classifier

This is a supplemental approach to the initial Baseline approach we have included

in myproject design.

The difference between this approach and the Baseline approach is that this approach usesthe EDC-NLP pipeline to pre-process its data, and the dataset is derived from the web scraped Fake-Real news dataset.

The rationale behind this approach is to be able to identify the relationship between using a web-scraped, comprehensive and well formatted, data cleaned dataset on the accuracy of Fake News Detection as compared to using a less competitive dataset.

5.8 Alternative PRO Approach: EDC-NLP + TF-IDF + Passive Aggressive Classifier

This is a supplemental approach to the initial Alternative approach .The difference between this approach and the Alternative approach is that this approach uses the EDC-NLP pipeline to pre-process its data, and the dataset is derived from the web scraped Fake-Real news dataset.

The rationale behind this approach is to be able to identify the relationship between usinga web-scraped, comprehensive and well formatted, data cleaned dataset on the accuracy of Fake News Detection as compared to using a less competitive dataset.

5.9 Fake News Detection Chrome Extension

Due to project time constraints owing to the training of the Deep Learning models for fake news detection, we decided to not go ahead with the Chrome Extension Deployment, as initially reported. This decision was made by me so that we were able to allocate sufficient headroom and duration for the completion of the project proper first.

This aspect of the project is beyond the scope, which we have undertaken as a personal interest to implement a robust, deployable software module for fake news detection. We propose to implement a Chrome Extension, which will be deployed using Python andFlask.

The 3-month outlook for this prototype requires me to take a series of incremental

and model corrective steps to develop an improved, highly accurate fake news detection model.

As mentioned in the Project Design section, the plan is to develop a BERT classificationmodel and Deep Learning model and to effectively tune and refine the classification confidence of the models to achieve higher detection accuracy, so that we are aligned with the aims of this project. The subsequent models are currently a work in progress, andfurther details, methodologies and the outline of the implemented models will be includedin the Final Report submission of this project. The high-level scope below shows some of the important milestones that is to be takenfor the further propagation of this prototype:

1. Improving text classification models
   a. Hyper parameter Tuning of models – model performance refinement
   b. Models & Output Coalescing - improve performance based oncombining multiple models and their outputs
   c. Improved dataset cleaning pipelines and text data cleansing
2. Packaging improved model structures into a more comprehensive software module for fake news .Beyond the scope: Implementation of a chrome extension module based on the most accurate fake news detection model I can implement (not included in Final Submission)

6. CONCLUSION AND FUTURE WORK

To conclude the Misinformation Acknowledgement using fake news detection employed bynatural language processing paper. Let us see if these questions have been addressed, and hence have the aims of this project been accomplished. Is it necessary to build an increasingly accurate NLP pipeline and model system to detect fake news when considering the capabilities of currently available options?

Are deep learning neural network models better in comparison to classification models for the chosen dataset and to achieve better accuracies when it comes to evaluating the improved models? The results we have presented in the previous chapter falls in line with this question. It can be well concluded that Deep Learning neural network models are

better in terms of text classification as compared to conventional machine learning techniques. However, it is also important to have the right input data to train the models. Hence in conclusion, Deep Learning models such as convolution neural networks produce high accuracies in fake news detection applications, and ought to be supplemented with optimized model architectures, and well formatted input data with the appropriate NLP pipelines. The RCNN and LSTM learning models are much more suitable for time series data, as we are using the LSTM layers which deal with sequential data. Hence for such models, it would be appropriate if the news or text corpus dataset had information such as date of publication etc.

References

1. Abhijnan chakraborty, bhargavi paranjape, sourya kakarla, and niloy ganguly. 2016. Stop clickbait: detecting and preventing clickbaits in online news media. pages 9–16

2. Alexandre bovet & hernán a. makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. nature communications, 10(7).

3. Davis, r., proctor, c. 2017. Fake news, real consequences: recruiting neuralnetworks for the fight against fake news.

4. S, Banerjee r, choi y. 2012 syntactic stylometry for deception detection. in:proceedings of the 50th annual meeting of the association for computational linguistics: short papers, vol 2. association for computational linguistics, pp 171–175 https://www.statista.com/statistics/1112026/fake-news-prevalence-attitudes-worldwide/

5. I. Traore et al. 2017 “detection of online fake news using n-gram analysis and machine learning techniques.” international conference on intelligent, secure, and dependable systems in distributed and cloud environments (pp. 127–138)

6. K Ludwig, m creation. 2020 “dissemination and uptake of fake-quotes in laypolitical discourse on facebook and twitter” j. pragmat, 157, 101–118.

7. Kajal yadav. 2020 fake-real news. kaggle dataset source.

8. Kaliyar, r.k., goswami, a., & narang, p. 2020. deepfake: improving fake news detection using tensor decomposition-based deep neural network. the journal of supercomputing, 1-23.

9. M. Chang et al j. devlin. bert. 2017: pre-training of deep bidirectional transformers for language understanding

10. M. Maimaiti, a. wumaier, k. abiderexiti, and t. yibulayin. 2017 “bidirectional long short-term memory network with a conditional random field layer for uyghur part-of-speech tagging,” information, vol. 8, no. 4, article no. 157.

11. Natekin a, knoll a. 2013 gradient boosting machines, a tutorial. front neurorobotics 7:21

12. P. M. Sosa, “twitter sentiment analysis using combined lstm-cnn models,” 2018 [online].

13. Papanastasiou f, katsimpras g, paliouras g. 2019 tensor factorization with label information for fake news detection.

14. Pavleska t, školkay a, zankova b, et al. 2018 performance analysis of fact-checking organizations and initiatives in europe: a critical overview of online platforms fighting fake news. in: eids6 (european integration and democracy series), pp. 1–29.

15. Pérez-rosas v and mihalcea r. 2015 experiments in open domain deception detection. in: proceedings of the conference on empirical methods in natural language processing, pp. 1120–1125.

16. Rabanser s, shchur o, gnnemann s. 2017 introduction to tensor decompositions and their applications in machine learning.

17. Rashki hannah and choi eunsol et al. 2017 truth of varying shades: analyzing language in fake news and political fact-checking. in proceedings of the 2017 conference on empirical methods in natural language processing, pages 2931–2937, copenhagen, denmark. association for computational linguistics.

18. Sammut c, webb, gi (eds). 2011 encyclopedia of machine learning. springer, boston, ma.

19. Shu k, wang s, liu h. 2019 beyond news contents: the role of social context for fake news detection. in: proceedings of the twelfth acm international conference on web search and data mining. acm, pp 312–320.

20. Swire b, ecker uk, lewandowsky s. 2017 the role of familiarity in correcting inaccurate information. journal of experimental psychology. Learning, memory, and cognition 43(12): 1948. crossref. pubmed.

21. T. Zhai and h. wang. 2022 "online passive-aggressive multilabel classificationalgorithms," in ieee transactions on neural networks and learning systems

22. Tom young, devamanyu hazarika, soujanya poria, and erik cambria. 2017. Recent trends in deep learning based natural language processing.

23. Torabi asr, f., & taboada, m. 2019. big data and quality data for fake news and misinformation detection. big data & society.

24. Wang wy. 2017 ‘liar, liar pants on fire’: a new benchmark dataset for fake news detection. in: proceedings of the 55th annual meeting of the association for computationallinguistics, vol. 2, vancouver, canada, pp. 422–426.

25. Dr. Subhani Shaik, Dr. K. Vijayalakshmi, Vaishnavi Anthireddy, Indhu Kethireddy, laxmi Kavya Marikanti,”Mitigating covid-19 transmission in schools using digital contact tracing”, Dickensian journal, volume 22, issue 6, June, 2022.

26. Monisha Singh, Dr. Sunil Bhutada, Dr. K Vijayalaxmi, Dr.Subhani Shaik,” Online social networks fake news on covid-19 in the South India Region”, Neuo Quantology, Vol.20, Issue 10, August 2022.