DETECTING THE SPREAD OF ONLINE FAKE NEWS USING NATURAL LANGUAGE PROCESSING AND BOOSTING TECHNIQUE

MSc Research Project
Data Analytics

Nandhini Haridas
Student ID: X17165989

School of Computing
National College of Ireland

Supervisor: Theo Mendonca

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	**Nandhini Haridas**
Student ID:	**X17165989**
Programme:	Data Analytics
Year:	2018
Module:	**MSc Research Project**
Supervisor:	Theo Mendonca
Submission Due Date:	12-12-2019
Project Title:	**DETECTING THE SPREAD OF ONLINE FAKE NEWS USING NATURAL LANGUAGE PROCESSING AND BOOSTING TECHNIQUE**
Word Count:	XXX
Page Count:	15

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	28th January 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

DETECTING THE SPREAD OF ONLINE FAKE NEWS USING NATURAL LANGUAGE PROCESSING AND BOOSTING TECHNIQUE

Nandhini Haridas
X17165989

Abstract

Social media is flooded with fake news these days. This impacts the quality of the social media. By this way online social media is purely affected by the enormous amount of fake news spread across the online platform. This will affect the reputation of the company that publishes the news. Unfortunately, very few research has been conducted to understand the intensity of this issue and to help overcome fake news that is spread in the social media. Among those few research were deep learning techniques that were used to help identify the fake news with better accuracy. But, the one main drawback of this technique is its high latency in getting the prediction as the data used will be very large and enormous. This paper is proposed in a way to eliminate this high latency issue and get a better, faster and sharper accuracy level. This research is proposed in using machine learning algorithms like Logistic regression, LightGBM and XGBoost to get a better, faster accurate results as they are good in low computational complexity. Also this research uses dimensionality reduction technique like PCA. By this way the time and space complexity will be low as the main motive of the proposed project is to predict the fake news in social media with low latency and better accuracy.

1 Introduction

In today's growth in technology, fake news are all around us. Social media is crowded with fake news Shabani and Sokhn (2018). This affects the readers to differentiate the original and the fake news. The fake news can mislead to many conspiracy theories of business man, celebrities, government officials, schools, colleges, employers etc. By this way the social platform that published the news would be losing their reputation. Due to this diverse impact on many VIP's, many researchers are working on to identify the fake news that are published in social media and are trying to help them with eliminating them Shabani and Sokhn (2018). Most of the fake news are usually on the political parties, celebrities and business person. They invest a lot on researchers to find out the fake news which is spoiling their reputation. Due to the fake news, people tend to either believe them or consider them to be as a conspiracy theory.

What is fake news?

In most understanding term, fake news are nothing but false news that are published in the online platforms like Facebook, Twitter, LinkedIn, Google news, etc Mahir et al. (2019). It has been recorded since the early 1890's during the moon landing and this is considered as one of the oldest fake news recorded Shabani and Sokhn (2018). The best example of the major political fake news that was spread during the recent days would be the US election for presidential that was held on 2016 Shabani and Sokhn (2018). With these kind of false news, the reputation of the targeted person is affected and it takes a huge step in order to make it back to normal.

Previous researches were done on this topic to understand the severity and the damage that the person who is affected by the fake news is going through. The study was conducted to understand the severity of the fake news and to get the maximum accuracy percentage using a hybrid machine learning models like Logistics regression, SVM, Random forest, Gradient boosting, and Neural networks and they were able to get 87 percentage of accuracy Shabani and Sokhn (2018). In another research Mahir et al. (2019), conducted to understand the fake news generated in twitter using five different machine learning models like SVM, Naïve Bayes, Logistic regression, and Recurrent neural networks. The research was conducted with the comparison of these machine learning models to understand the best performed model out of which, SVM and Naïve Bayes performed better on comparison with 89 percentage and 84 percentage respectively. The problem that was figured with both the research papers were that, even though the accuracy was high the latency was equally high in both the cases. This led to the research idea after understanding the exact need that a machine model should give high accuracy as well as low latency. As the research is conducted for the real-world scenario, a model that has low latency with high accuracy will be beneficial for using with large dataset. This led to the research question for taking this project.

"To what heights can a fake news on social platform be detected using NLP models like bag of words and machine learning algorithms like LightGBM, XGBoost and Logistic regression used to identify fake news with low latency?".

This research question is evolved in understanding the necessity of getting an output with low time complexity with high accuracy. This motivated in building this project on an aim of getting an accurate machine learning model with low latency. The machine learning models like Logistic regression, LightGBM, XGBoost are the best when it comes to performing with best accuracy. To achieve the goal with low latency and to reduce the space and time complexity, PCA dimensionality reduction technique is used to reduce the time complexity.

2 Related Work

2.1 FAKE NEWS ON SOCIAL MEDIA

Tyagi et al. (2019) analysed the spread of misinformation in Online Social Media by considering Twitter as one medium. According to the paper submitted by her, the information spread is more with the fake content than the real content. What comes out first seemed to be believed by the most than what is True. Another most vital reason for the spread of misinformation was the lack of central moderation on the flow of information because of the fact that all the information are crowd sourced. The proposed solution for this problem was to classify the information based on the veracity. The Tweets will be collected based on the Keywords and the Hashtags used and will be used as the datasets to build a machine learning model. The Information was then fed to the model which will be classified by Naïve Bayes and Decision Tree algorithms for classification using the Python's NLTK module. The Model is then evaluated for Performance using the Confusion Matrix. The confusion matrix helps us to understand the errors made by the classifiers and its types. The same study can be applied to different social media, For example, Facebook.

Also, one more problem in collecting Datasets in the form of news articles and Tweets is that most of the Data are unnecessary for learning. Splitting the news articles into related and unrelated ones are really important.Jang et al. (2019) have proposed a word2vec CNN (Convolutional neural network) to classify the news articles. To construct CNN, two-word embedded algorithms of word2vec, Continuous Bag-of-word (CBOW) and skip-gram were used.He proposed a embedding technique called word2vec. The Performance of the CNN model with three-word embedding CBOW, skip gram and Random vector was analysed for accuracy, recall, precision and F1 score as performance metrics. The CNN with CBOW exhibited values corresponding to 0.9147 and 0.9167 for F1 and accuracy respectively, and the CNN with Random vector had values of 0.8475 and 0.8409. The results proved that the CBOW algorithm performed better on news articles and skip gram algorithm performed better on Tweets. The news articles follow a uniform pattern and so the CNN models can be used to extract features and perform faster accurate classification when the data is processed with the CNN based classification.

Spam messages or Comments posted in social media is believed widely by people around. Although spam emails are detected by SMTP servers, the same cannot be applied in detecting spam posts in social media. Sohrabi and Karimi (2018) have analysed a feature selection approach in detecting spam in Social media. Facebook was considered a medium in collecting the Datasets. PSO (Particle swarm optimization) was the feature selection algorithm that was proposed. Number of selected features were calculated after which a solution to determine the selected features were used. After this, the Data was prepared for clustering.The Evaluation was carried out on three methods of clustering detection, SVC and Decision tree. The clustering method using DB index proved in the proper detection of legitimate messages and SVC showed a proper detection of spam messages. So, the combination of clustering method using DB index and SVC were used in the proposed system. The unsupervised machine learning method of clustering method combined with supervised machine learning method of SVC showed more precision in detecting spam messages. The results proved that the proposed method had a very good detection rate.

Social media is a platform with a perfect mixture of communication between a business and an individual. To fulfil this, Aswani et al. (2018) proposes a hybrid model that detects fake profile by using Bio inspired computing and analysis from social media. Levy flight firefly Algorithm (LFA) with chaotic maps are used to detect fraud in Twitter business by using K-means integration. Later in the paper to identify spammers and non-spammers Fuzzy C-Means clustering approach is used to detect overlapping profile between the two clusters. The proposed K-means levy flights algorithm (LFA) gives a best accurate result with an accuracy of 97.98 percentage after a statistical t-test by combining different significant factors.

Social media content has become mode dangerous with all the fake news created by social media users. The fake news and real news are to be detected to eradicate the disadvantages of the social media. Novel approach by Ozbay and Alatas (2019) for fake new detection and metaheuristic algorithms like Grey Wolf Optimization (GWO) and Salp Swarm Optimization (SSO) has been used to detect the fake news in social media. The real time data is compared with several supervised algorithms. However, Grey Wolf Optimisation has promising results with high accuracy. As the paper suggests GWO optimises the best value and F-score with social media. Advantage of false news detection is they construct models that consist of mind specified words.

From a technician perspective the fake news detection draws some attention. The fake detection can be determined as straight forward, binary classification problem for fake and non-fake tweets. The research paper by Helmstetter and Paulheim (2018) used weakly supervised approach which comprises multiple data set at once and train the classifier based on this data set. The classification can be used for fake and non-fake tweets by dividing them into classification target. Despite this unclear data the data can be supervised and can produce a good output and F1 score up to 0.9. The data sets used are small that are collaborated with the model. The approach yields a very good result achieving f1 score of 0.77

The study uses twitter-based mining approach with machine learning models like time-aware semantic analysis of users. The data is classified in to two different classifiers politics and non-politics. The paper by Abu-Salih et al. (2018) uses Logistic regression, Support Vector Machine and Decision tree classification as a part of the project. These tweets are divided into domain detection user level to find the negative tweets on user level. Domain detection Tweet level comprises of negative tweets on personal tweet level. OSN represents the context of user content the major challenge is understanding the domain using OSN.

2.2 EMAIL AND SMS SPAM DETECTION

In a study, a research-based company out of 201 billion emails that were sent daily 18.5percentage seemed irrelevant to the recipient and 22.8percentage were sent unnecessarily. These spam mails result in the cause of loss of money to the company. Alurkar et al. (2017) proposed a data science-based system architecture that could possibly classify the emails between spam and ham. The structure is built in steps with the first step being collecting the Data in the form of emails and formatting them after which EDA (Explore and Analyze Data) is executed. After the Data is prepared the data Analysis was processed. The Decision theory was used to classify the emails as s[pam/ham. Since the output of the study was binary (Spam/Ham), the focus was to categorize the Data accordingly. The paper proposed the Dataset to be taken independently. In this paper, A statistical model was not implemented completely rather it proposes Progressive web app that could platform independent. A classifier like this would not only detect the emails being spam or ham but also prevents them and save and in return save lot of time and money for the organization.

Fake news has an immense impact on society to overcome this disadvantage the research paper by Aphiwongsophon and Chongstitvatana (2018) used machine learning techniques like Naïve Bayes, Neural Networks and support vector mechanism. Data is cleaned using normalization before using it in machine learning model. As a result, fake news can be identified using machine learning model. However, it may not be easy to detect the fake news in real time but given the fact that it is not hard to detect the fake news. Other models can also be used to broaden the effect of detecting the effect in real time model.

This paper by Gupta et al. (2018) relates spam detection by SMS using machine learning classifiers. As this spam SAM could lead to loss of data and privacy which could be harmful. To avoid this the researcher proposed to classify different classification techniques using different data sets. The techniques like accuracies, precision, recall and CAP Curve techniques are used. In this paper 8 different classifiers are compared to get a high variant and pression on the datasets. In these classifiers Convolution Neural network Classifier produced a highest result and accuracy of 99.19 percentage . However, CNN is used in this classification to achieve this result.

2.3 FAKE NEWS DETECTION

Klyuev (2018), in his paper has said that spread of fake news in Social media has a greater impact on a larger audience as most of the people from across the globe are not into reading newspapers or watching news on television. He has analysed and has come up with machine learning model that could automatically detect and tag fake publications to help people. The procedure followed for Machine learning was to gather datasets from Twitter and to specify set of features to be extracted from them. The Testing methods proposed were Naïve bayes, Unsupervised clustering and decision trees. The Model created using naïve Bayes proved to be robust as it was able to flag the Fake news

Gupta et al. (2018) have analysed the spread of fake news in Twitter and tried to build a Machine learning and deep learning algorithm to attack this problem. A Five different types of classification models that includes Support Vector machine, naïve Bayes, Logistic regression, Long-short term memory, recurrent neural network was implemented. For Implementation, Python 3.6.5 was used as a programming language. As the Dataset were mostly "Text", other techniques like Count Vectors, TF-IDF and word embedding were used. After Cross validation, an accuracy of 62.47percentage for Logistic Regression, 84.56percentage for Naïve Bayes and 89.3percentage for SVM was achieved. For Deep learning algorithms, Long-Short-term Memory and recurrent Neural networks were used and it achieved an accuracy of 74percentage and 76percentage respectively. The precision seemed to be higher for Naïve Bayes and SVM models as it acquired a F1 score of 0.94. So, the final conclusion was that the SVM model performed the best for characterization technique even with the largest dataset of nearly 20360 observations.

Fang et al. (2019) tried to analyse the Fake news spread in social media by taking a public Dataset and building model with CNN (Convolutional Neural Network) and self multi-head attention mechanism. They also have compared the accuracy of a fake news classifier with previous results. The Result was rather obvious showing great accuracy in the fake news which also created a good effect in their study of detecting fake news. The Proposed model acquired an accuracy of 95.5percentage under 5-fold cross-validation in the chosen public Dataset. The Dataset had nearly 24,000 news articles in which 12,228 fake news and 9,762 non-fake news were used in the experiment finally. A confusion matrix was used to classify the precision. The method of SMHA-CNN -1200's precision reached 95.5percentage in the detection of the news articles. Other methods like LSTM and Att-LSTM were also used for checking the precision. It seemed that LSTM was inefficient with a model of 1200 as input length.

"Fake News Detection Enhancement with data Imputation". The Datasets that are collected often contains noises like Missing values that needs to be imputed. Kotteti et al. (2018) have proposed a Data Imputation preprocessing strategy for Detecting fake news using machine learning Techniques. The Dataset chosen had possible Missing values in it which was treated with scikit-learn's Imputer with Mean strategy which replaces the missing values with mean. Categorical Imputer was used for handling categorical missing values.The Traditional methods that were used as a proxy to build models were SVC, Decision Trees, Multi-Layer Perceptron and Gradient Boosting Gradient Boosting. Feature Extraction proposed by Kotteti et al. (2018) uses Term Frequency and inverse document frequency (TF-IDF) to identify the useful features from the news contents. For Evaluation, F1 score, precision, accuracy and recall were used to select the best model. Further, the computational complexity of the models was monitored in the training Dataset by noting down the prediction time of different classifiers. Around 4000 records were deleted from the Dataset and MLP classifier proved to stand out among the other classifiers with a F1 score of 0.37 and an accuracy of 0.416percentage. The remaining missing values were filled with empty text which prevented Data loss and still the MLP classifier outperformed other classifiers with a F1 score of 0.454 and an accuracy of 0.458 in Training Dataset.

Mobile phones are source of communication, information seeking and publishing using social media as a key instrument. The freedom in online media also promotes overwhelming users producing fake news. The research in this paper by Klyuev (2018) used text features utilising Natural language processing methods and detect spam bots using machine learning model from social networks. This combined mechanism can produce high level accuracy in filtering the fake news. Machine learning techniques like Support Vector Machines (SVM), Gradient Boosting, Bounded Decision Tree, Random Forest techniques are used to overlap the fake news by analysing the tweets in social media. The users can use open source resources to analyse the quality of the tweet.

This research paper by Katsaros et al. (2019) uses various machine learning techniques for fake news detection and classification to avoid vertigo of fake news. The researcher examined L1 regularized logistic regression, Support vector classification, Gaussian naïve Bayes, Multinomial native Bayes, Decision tree, Random forests and Neural networks to determine fake news vs legit news. The efficiency and training speed were tested between these algorithms. As a result, the paper concludes that a space with hundred dimensions is capable to capture high accuracy of detection between fake and legit news. As far as the results go, Neural networks is best suitable for downside of higher training time.

3 Methodology

KDD- This project uses KDD methodology. Knowledge Discovery in Database also known as KDD is used in finding the in depth information and the hidden details in the data. This methodology helps in understanding the high-level information of the data and their hidden pattern. As this project uses high volume of data, this methodology would help in identifying the every minute information and utilises the at most information out of the data.

3.1 Business Understanding:

To take forward a project, the very first step is to understand the business needs and propose a model that will solve the business problem. After finalising the topic, the research was carried out to understand what was lacking in this particular domain. It is understood that, there were no much research carried out in detecting fake news in social media which gives an accurate result with low latency. The researches that was carried out did not provide a faster and simpler solution even though they gave good accuracy.

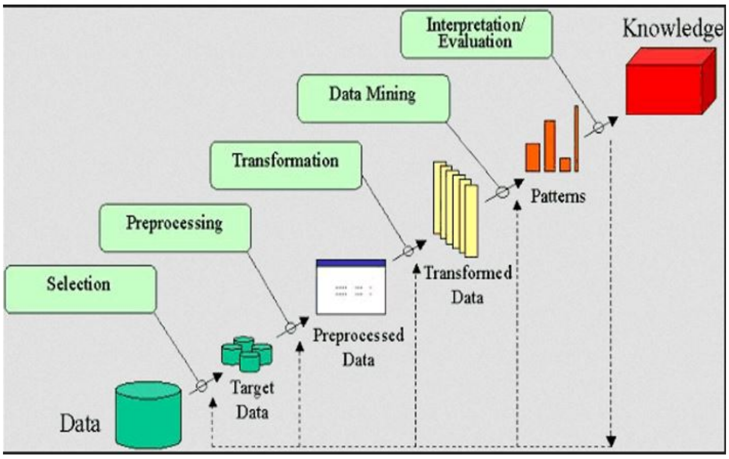


Figure 1: KDD

This was taken into consideration and a project was carried out to get a better, faster and simple machine learning model that will not only provide accurate results but also a machine learning model that will perform faster.

3.2 Data Understanding:

Only when the data is understood the project can be carried forward. Initially after the data is taken from the website the first step was to check the total number of rows. The data that is used in this project has 22,652 rows and 5 columns. Out of this 5 columns, only two columns were used in this project as it serves best for the prediction.

3.3 Data Preparation:

Once the data is studied the next step is to prepare the data for analysis. In the taken dataset, there are unwanted columns that needs to be removed as they do not contribute much for the prediction. Also, it is seen that the dataset is imbalanced meaning the fake news in the dataset is low compared to the true news. This class variation needs to be either over sampled or under sampled to bring it to right proportion. In this project random under sampling is carried out to train the model in a better way and remove the imbalance in the data. Along with this, the text data is also normalised to lower cases and then complete stop words were removed from the text data. Finally text data was lemmatised so as to make it more useful for the analysis.

3.4 Modelling:

This is the main step where the machine learning algorithm is interpreted with document term matrix. Uni-gram term matrix is created to interpret the results using machine learning algorithms. High dimensionality was found when Uni-gram term matrix was created. For reducing the dimensionality, PCA dimensionality reduction technique is used. PCA also known as Principal Component Analysis, is the method that is used in this project to carry out the dimensionality reduction. This method is used to remove the unwanted columns that do not contribute much to the analysis that will eliminate the noise and reduce the quality of the model.

Once the dimensionality reduction technique is carried out, machine learning algorithms are used to interpret the results. In this project, machine learning models like Logistic regression, XGBoost, LightGBM and Decision tree are used to do the analysis and to understand whether they give a model that has high accuracy level with low latency.

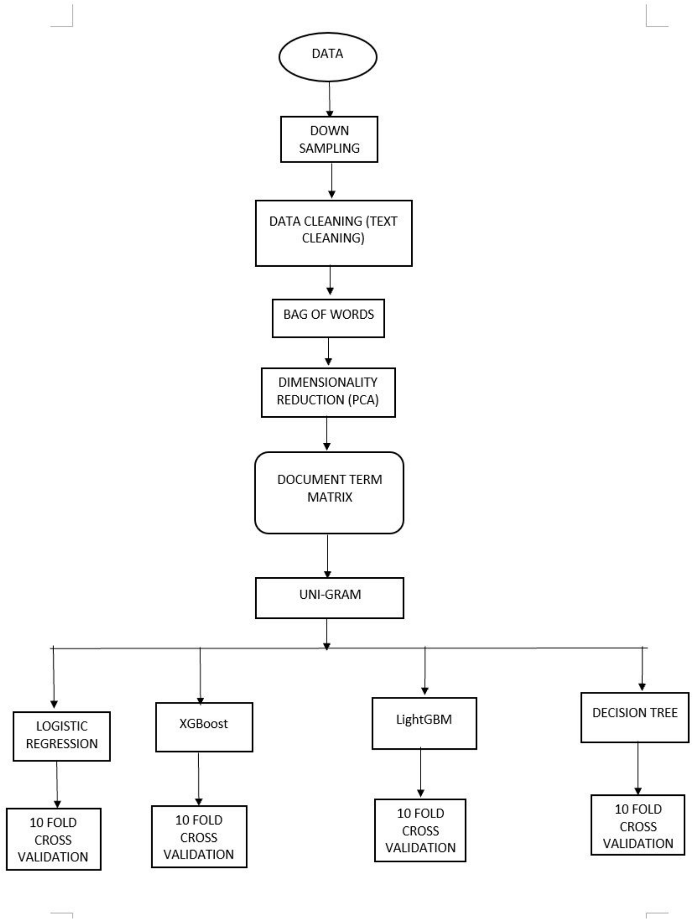


Figure 2: PROJECT FLOW

3.5 Evaluation:

Once the model is created it has to be evaluated using various evaluation metrics. In this project the following evaluation metrics are used to understand the performance of the machine learning model. They are10-fold cross validation, AUC, accuracy, and F1-score are used.

4 Design Specification

The project was initially done in desktop, but since the data is too large the project was carried out using Google colab. This is a free cloud platform which supports GPU. This platform is widely used for improvising python programming and several deep learning applications as it has inbuilt several packages. There are many free cloud platform, but the Google Colab is used in this project due to its various inbuilt features like TensorFlow, PyTorch, Keras etc along with free GPU service. By this, the process would be fast and simple. The colab session works under a virtual machine platform and gives an option of using 13Gb of ram using CPU, GPU or TPU. This project is supported using GPU with 13Gb ram and 120 Gb of drive space.

5 Implementation

5.1 Data Selection:

The dataset used for this project is taken from https://www.kaggle.com/c/fake-news/data. The total number of rows in this dataset are 20250 and it has 5 columns out of which only 2 columns are contributing well for this project. The other 3 columns were not used in this project. The dataset also has huge class imbalance meaning, the fake news found in the dataset were only 7600 and the genuine news were 17720. The data is selected after verifying about the privacy and ethical issues. This data is publicly available and do not contain any information that is against the data privacy.

5.2 Data sampling:

Once the data is sorted, then it is studied to check the class imbalances. It is understood that there is huge difference between the fakes news and genuine news. To overcome this, data sampling needs to be done. There are two types of data sampling process; Over sampling and under sampling. Here, under sampling is done as the genuine news class is in majority. Both sampling has their own advantages on machine learning models, but in this project it was better to do under sampling as the minority class was fake news. Random under sapling approach has been carried forward in this project. This process brings down the majority class as equal to the minority class making it easier for the machine learning model to do the analysis. By this was the class imbalance will be eliminated.

5.3 Data Pre-processing:

In every machine learning projects, it is very important to process the data before fitting into the model as the data would have number is missing values, errors etc. The following steps were done for data pre-processing:

5.4 Cleaning Text data:

In this project cleaning the data was carried with different levels. First, the data was checked for null values and unwanted columns and they were removed as there were columns that do not add value to the project. The next process was to remove the stop words from the data. The reason for removing stop words is because it causes dimensionality to the model. Hence removing the stop words will help reduce the dimensionality for the model. Then the data was lemmatised using WordNetLemmatiser package. Lemmatising is a process of removing the words that has similar meaning e.g.: Store, Storing, Stored. Once the lemmatise is done only the word “Store” will be kept removing the other two words. By this way the when the document matrix is created it will not take them as three different words thus reducing the time and complexity. Finally the data is normalised by converting it into lower cases. This is the important step as this will reduce the duplication of the data.

5.5 Document Term Matrix:

To implement machine learning model, the data needs to be in numerical format. The conversion of text to numbers is termed as document matrix. To do this process bag of words model is used. Bag of words is a Natural language processing (NLP) technique to convert the words to numbers for making the machine learning model to best fit in the process. By this way we can segregate the model in a way to get the desired output. What is Natural Language Processing (NLP)? Natural Language Processing is technique that is the combination of Artificial Intelligence with machine learning models. The interference of the human language to allow the computer To get the output is the basic idea of Natural Language Processing. This is widely used for converting the large volume of text data into useful information for analysing [https://medium.com/greyatom/a-dive-into-natural-language-processing-103ae9b0a588].

BAG OF WORDS: Bag of words is a technique to segregate the whole text into separate columns and then convert them into numerical format by count of vectorisation. This method is very simple and the easiest way to format the text data in such a way to get the best out of the machine learning model. This method helps in implementing the machine learning model. For eg. Lets assume this sentence:

"It is a warm sunny day"
"It is a hot sunny day"
"It is a bad day"
"It is a good day"

The sentence above will be split as follow: 'It', 'is', 'a', 'warm', 'sunny', 'day', 'hot', 'bad', 'good'. Each words will be separated as a document and then they will form a vector as "o and 1". Each token will be taken as a gram and will be further separated as Unigram, Bi-gram, Tri-gram etc. https://medium.com/greyatom/an-introduction-to-bag-of-words-in-nlp-ac967d43b428 This project is further separated as Uni-gram.

UNI-GRAM This is a type of model which uses single term of sequence from the given sequence of data.

5.6 DIMENSIONALITY REDUCTION:

In projects based on using data, there would be random variables that will create noise and impact the output. In this project as the data used is text, it has huge random variables and unwanted columns that is not useful. To eliminate this and to get the proper accuracy level it is important to remove the dimensionality. Machine learning models usually face the curse of dimensionality which will impact their output. Due to this the model faces noise, over-fitting and high time complexity. This can be eliminated by various dimensionality reduction techniques. This project uses Principal Component Analysis (PCA) technique to reduce the dimensionality. **Principal Component Analysis**:

PCA is a technique used to reduce the dimensionality between the variables. The main idea behind PCA is to show the output of the data from a most informative way by using the best relevant variables suited for the project Ganaa et al. (2019). The advantage of using PCA is that it will not remove the essence of the original data rather keep the required information needed for supporting the machine learning models.

In this project there were more random variables found that was causing noise which in turn caused space and time complexity. The dataset contained few columns that were not needed for the project and caused noise and time complexity. To remove this complexity, PCA technique has been used in this project to remove unwanted columns and random variables. PCA has been showing promising results for dimensionality reduction.

5.7 CLASSIFICATION USING MACHINE LEARNING ALGORITHMS:

- **LOGISTIC REGRESSION:**

Logistic regression is considered to be one of the best model in machine learning when it comes to predictive analytic. They are used to predict the outcome for a discrete set of data. This model works best for the classification problem when dealing with natural language processing Anggraina et al. (2019). There are three types of logistic regression namely binary logistic regression, multi-linear logistic regression and nominal logistic regression. Logistic regression works on a probabilistic approach with 0 and 1. By this way it makes the model more simple to be used for real time problem.

Binary logistic regression: This type of logistic regression deals with only having two types basically 0's and 1's. Multi-linear logistic regression: This type of logistic regression have 2 or more variables. Ordinal logistic regression: This type of logistic regression will have variables that are ordered values.

- **XGBoost:**

XGBoost is a type of machine learning model that helps the model to boost its performance. XGBoost also known as eXtreme Gradient Boosting works on a principle of tree formation which helps the model to increase its speed. The name by itself explains that this will help the model achieve the highest accuracy with boosted tree algorithmReis et al. (2019).

This model serves best in giving the highest accuracy level. This works on a way by boosting the model until there is no improvement is required to train them. This also helps in making the corrections that the new models created when being added; meaning XGBoost is basically working on a way that it trains the new models to detect the errors created by its previous models. By this way the model boosts its performance and give better prediction. This way the model will perform with low latency.

- **LightGBM:**

LightGBM is another type of boosting technique that is based on a tree learning algorithms. Meaning, the model grows in the leaf-wise rather level-wise like other models. By this way it will have low loss. They are very efficient is memory storage, high efficiency and train the models fast, high accuracy, also supports parallel computation and it works on GPU learning, it can also handle large dataset. LightGBM do not require to encode the categorical data. It can work directly without encoding them. It also supports weighted query group data type Yang et al. (2019).

- **Decision Tree:**

The term decision tree by itself describes that the model works on a tree like graph model. This way the internal node acts as a test, the branch of the nodes represents the result of the test and the leaf node denotes the class label. The classification logic is the entire root to leaf representation. In machine learning models the tree based models are said to be best as they provide results on a leaf-wise structure rather level-wise. This means, the model will automatically detect the errors in each step and making it overcome the error before it is passed onto the next level. Decision tree in this project is used to compare the results with the other machine learning models and verify the results to check which model gives better performance Thorne et al. (2017).

- **K-FOLD CROSS VALIDATION:**

K-Fold cross validation is a statistical technique used to enhance the efficiency of the machine learning model. The term k refers to the number of groups the data is folded to get an efficient value. This is a very simple method that can be used to get a desired output. This method is usually used while predicting the model and not while training the model. The process is usually carried out by choosing a random set of unique group in the given dataset and then keeping the rest as training set. The model is built using the unique set of samples and the output is interpreted. The output is summarised by taking the overall mean value Lam et al. (2009).

6 Evaluation

After running the models the results are interpreted using specific evaluation metrics in order to understand the performance of the models. The following evaluation metrics are used in this project to for interpreting the results:

6.1 ACCURACY

This metric tells us how much percentage is a machine learning model accurate. It is calculated by the formula: Accuracy = TP+TN/TP+FP+FN+TN Where, TP- True positive TN-True negative FP-False Positive FN-False Negative

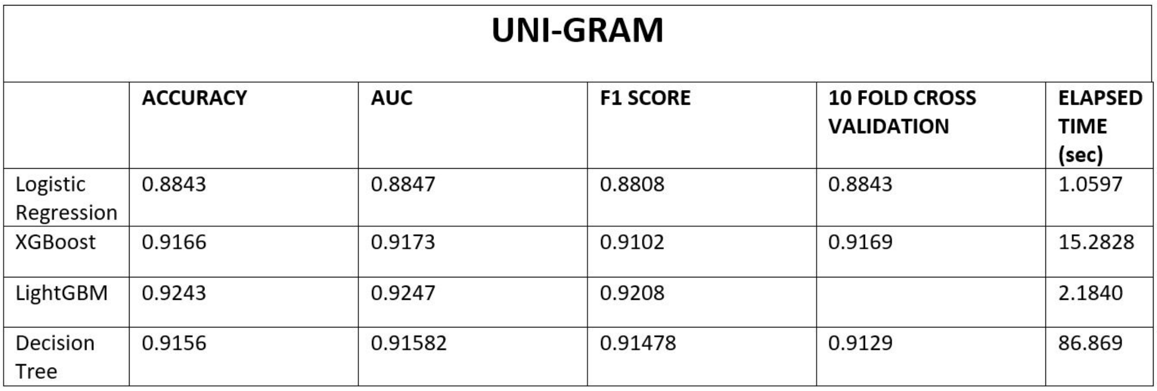
6.2 F1 SCORE:

F1 Score is the advanced value of precision and recall. The two values that F1 score delivers are 0 and 1 where 1 being the best value and 0 being the worst value. It is calculated as follows: F1 Score= 2*(Precision*Recall)/ (Precision+Recall)

6.3 AUC:

Area Under Curve also known as AUC gives a the output by combining all separate variables across all the models. This is a prediction technique usually used in the classification problem for a machine learning model. This is one of the most important evaluation metrics as it does not incline towards the size of the evaluation data. Like F1 Score, AUC gets the value ranging between 0 and 1 and is calculated using sensitivity and specificity

6.4 RESULTS:



UNI-GRAM

	ACCURACY	AUC	F1 SCORE	10 FOLD CROSS VALIDATION	ELAPSED TIME (sec)
Logistic Regression	0.8843	0.8847	0.8808	0.8843	1.0597
XGBoost	0.9166	0.9173	0.9102	0.9169	15.2828
LightGBM	0.9243	0.9247	0.9208		2.1840
Decision Tree	0.9156	0.91582	0.91478	0.9129	86.869

Figure 3: UNI-GRAM

The below results are interpreted using evaluation metrics and summarised in a tabular form.

6.5 Discussion

The project was carried out on basis of understanding whether a machine learning model can get an accurate results with low latency. The major complications that came across while doing this project was the curse of dimensionality. As a result the data was indeed reduced and could not work on a very large dataset. The next question was using the open source cloud platform to make the project. This was another tough situation when doing document term matrix. As the space consumed for bi-gram and tri-gram were too large and as a result the system crashed and a result could not implement them. The boosting techniques used in this project are relatively new and to understand the parameters and tune them was difficult hence it was eliminated in this project.

7 Conclusion and Future Work

This project on Fake news detection was carried out with a motive to help the social sites in understanding the difference between fake news and genuine news by using machine learning models and receiving high accuracy with low latency. As a results the machine learning models used in this projects outperformed on than the deep learning models used in the previous projects in terms of low latency and high accuracy. To achieve the said goal, initially the data was checked for class imbalance and dimensionality problem. PCA dimensionality reduction techniques were used to eliminate the curse of dimensionality. By this step the unwanted columns that produce noise and time complexity is removed for performance efficiency. Then the machine learning models were implemented. The models used in this project were Logistic Regression, XGBoost, LightGBM and the results were then compared with the traditional machine learning model like Decision tree for a better understanding.

The results were noted and tabulated. T The results were tested with various evaluation metrics and they were then compared. The results show that Logistic regression performed better with low latency but LightGBM performed better both in Low latency and high accuracy and XGBoost performed better than the traditional machine learning model Decision tree.

LightGBM showed an accuracy of 0.92 with elapsed time of 2.1 seconds followed by Logistic regression with accuracy of 0.88 with elapsed time 1.05 and XGBoost with 0.91 as accuracy level with elapsed time as 15.28 seconds.

FUTURE WORK:

a. In future, this research can be carried out with new set of data with large volume or with live data. By this way it can be known to understand the computational speed and latency level in an advanced way.

b. In this project the research did not carry out with hyper-parameter tuning. In future this can be a scope to understand the best performing parameters for the same machine learning models.

c. The next future work suggestion would be to research on new machine learning models and try implementing them with this dataset to compare which model outperformed.

References

Abu-Salih, B., Wongthongtham, P. and Chan, K. Y. (2018). Twitter mining for ontology-based domain discovery incorporating machine learning, *Journal of Knowledge Management* **22**(5): 949–981.

Alurkar, A. A., Ranade, S. B., Joshi, S. V., Ranade, S. S., Sonewar, P. A., Mahalle, P. N. and Deshpande, A. V. (2017). A proposed data science approach for email spam classification using machine learning techniques, *2017 Internet of Things Business Models, Users, and Networks*, IEEE, pp. 1–5.

Anggraina, A., Primartha, R. and Wijaya, A. (2019). The combination of logistic regression and gradient boost tree for email spam detection, *Journal of Physics: Conference Series*, Vol. 1196, IOP Publishing, p. 012013.

Aphiwongsophon, S. and Chongstitvatana, P. (2018). Detecting fake news with machine learning method, *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, IEEE, pp. 528–531.

Aswani, R., Kar, A. K. and Ilavarasan, P. V. (2018). Detection of spammers in twitter marketing: a hybrid approach using social media analytics and bio inspired computing, *Information Systems Frontiers* **20**(3): 515–530.

Fang, Y., Gao, J., Huang, C., Peng, H. and Wu, R. (2019). Self multi-head attention-based convolutional neural networks for fake news detection, *PloS one* **14**(9).

Ganaa, E. D., Abeo, T. A., Mehta, S., Song, H. and Shen, X.-J. (2019). Incomplete-data oriented dimension reduction via instance factoring pca framework, *International Conference on Image and Graphics*, Springer, pp. 479–490.

Gupta, M., Bakliwal, A., Agarwal, S. and Mehndiratta, P. (2018). A comparative study of spam sms detection using machine learning classifiers, *2018 Eleventh International Conference on Contemporary Computing (IC3)*, IEEE, pp. 1–7.

Helmstetter, S. and Paulheim, H. (2018). Weakly supervised learning for fake news detection on twitter, *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, pp. 274–277.

Jang, B., Kim, I. and Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets, *PloS one* **14**(8).

Katsaros, D., Stavropoulos, G. and Papakostas, D. (2019). Which machine learning paradigm for fake news detection?, *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, IEEE, pp. 383–387.

Klyuev, V. (2018). Fake news filtering: Semantic approaches, *2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, IEEE, pp. 9–15.

Kotteti, C. M. M., Dong, X., Li, N. and Qian, L. (2018). Fake news detection enhancement with data imputation, *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, IEEE, pp. 187–192.

Lam, K. C., Palaneeswaran, E. and Yu, C.-y. (2009). A support vector machine model for contractor prequalification, *Automation in Construction* **18**(3): 321–329.

Mahir, E. M., Akhter, S., Huq, M. R. et al. (2019). Detecting fake news using machine learning and deep learning algorithms, *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, IEEE, pp. 1–5.

Ozbay, F. A. and Alatas, B. (2019). A novel approach for detection of fake news on social media using metaheuristic optimization algorithms, *Elektronika ir Elektrotechnika* **25**(4): 62–67.

Reis, J. C., Correia, A., Murai, F., Veloso, A., Benevenuto, F. and Cambria, E. (2019). Supervised learning for fake news detection, *IEEE Intelligent Systems* **34**(2): 76–81.

Shabani, S. and Sokhn, M. (2018). Hybrid machine-crowd approach for fake news detection, *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, IEEE, pp. 299–306.

Sohrabi, M. K. and Karimi, F. (2018). A feature selection approach to detect spam in the facebook social network, *Arabian Journal for Science and Engineering* **43**(2): 949–958.

Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X. and Vlachos, A. (2017). Fake news stance detection using stacked ensemble of classifiers, *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pp. 80–83.

Tyagi, S., Pai, A., Pegado, J. and Kamath, A. (2019). A proposed model for preventing the spread of misinformation on online social media using machine learning, *2019 Amity International Conference on Artificial Intelligence (AICAI)*, IEEE, pp. 678–683.

Yang, K.-C., Niven, T. and Kao, H.-Y. (2019). Fake news detection as natural language inference.