FAKE NEWS DETECTION

A

Project Report

submitted in partial fulfilment of the

requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE

Specialization in

Telecom Informatics

By:

Name	Roll No	Branch
Abhay Chauhan	R106216001	CSE TI
Divya Tyagi	R106216009	CSE TI
Rajnish Kr Gupta	R106216014	CSE TI
Dikshit Sharma	R106216020	CSE TI

Under the guidance of

Mr. Ahatsham

Assistant Professor

Department of Informatics

School of Computer Science

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES

Bidholi, Via Prem Nagar, Dehradun, Uttarakhand

2019-20

CANDIDATES DECLARATION

We hereby certify that the project work entitled **Fake News Detection** in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering with Specialization in Telecom Informatics and submitted to the Department of Informatics at School of Computer Science, University of Petroleum and Energy Studies, Dehradun, is an authentic record of our work carried out during a period from **January, 2020 to April, 2020** under the supervision of **Mr. Ahatsham**, **Assistant Professor**.

The matter presented in this project has not been submitted by us for the award of any other degree of this or any other University.

Abhay Chauhan	**Divya Tyagi**	**Rajnish Kr Gupta**	**Dikshit Sharma**
(R106216001)	(R106216009)	(R106216014)	(R106216020)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: 30 April 2020

(**Mr. Ahatsham**)
Project Guide

Dr. T.P. Singh
Head, Department of Informatics
School of Computer Science
University of Petroleum and Energy Studies
Dehradun - 248001 (Uttarakhand)

ACKNOWLEDGEMENT

In endeavoring to present this project, we would like to appreciate **School of Computer Science** for giving us an opportunity to get acquaintance with the fields of Artificial Intelligence and Machine Learning and its features for enhancing our skill set as software developers.

We are deeply indebted to **Mr. Ahatsham**, Assistant Professor and mentor, for his continued support and guidance throughout the project. He encouraged us to be innovative and to strive towards making a more efficient project.

We would also like to take this opportunity to thank **Dr. T.P. Singh**, Department Head, School of Computer Science for giving us a chance to be a part of this program and proving our skills.

We extend our gratitude to **Dr. Manish Prateek**, Professor and Dean SCS, UPES for giving us the necessary facilities to carry out our project work successfully.

Finally, we would like to thank all our **friends** for their help and constructive criticism during our project work. We have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

Abhay Chauhan	**Divya Tyagi**	**Rajnish Kr Gupta**	**Dikshit Sharma**
(R106216001)	(R106216009)	(R106216014)	(R106216020)

ABSTRACT

The authority of news in penetrating and altering the opinions and decisions of the countrymen is undeniably significant. People tend to believe what they see and read, and fundamentally base their opinions on the randomly perceived knowledge. To prevent the commotions rendered by the news of no relevance and fraudulent origins, this project presents a solution to detect fake news. The implementation of Natural Language Processing techniques to identify the origin or source which could possibly be producing fake news has been the primitive goal. This project emphasizes on recognizing the fake news assets relying on the authenticity of numerous articles produced by a single source. If an entity is recognized as a fraudulent source, with high confidence one can affirm that any further article or news-supply from it would be fake. Stressing the classification on the source widens the tolerance of wrong deductions because the source classification is based on the statistical observations of their previous supplies.

The inference laid after the result analysis clearly suggests that the first phase of this project's development, in all the four cases of TF-IDF vector levels—the Count Vector, the Word level, the N-Gram level, and the Character level—the Support Vector Machine had consistently shown the highest accuracy, that is, around 71.58%. And, in the second phase, the incorporation of Deep Neural Network Models—CNN, LSTM, and GRU—higher accuracy was furthermore achieved. LSTM showed a phenomenal accuracy hike around 91.25%. Hence, among all the applied algorithms, in order to efficiently determine hoax or fake news, the implementation of LSTM is preferable for the model's desirable credibility.

Keywords: NLP, Linear Regression, Naïve Bayes, SVM, TF-IDF Vectors, ANN, Deep Neural Networks, CNN, LSTM, GRU, Fake news.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

This project aims to detect and bring to periphery the much prevalent scams—in form of diplomatically deceiving news—using machine learning algorithms. The significance of read/heard news in recasting the beliefs of the people and their reactions is undisputable. People tend to believe what they perceive from images or writings, and form superficial opinions with no stringent base. Now, how wrong would the whole course of such a perception go if the acquired-knowledge over which the opinions were established came out to be nothing but a lie? In recent years, fake news has been reported for having material impact on significant local and global issues leaving people in a dire commotion. Fraudulently, this concept has been taken as an expanding issue to put heat over the delicate news topics and seems inevitable by its course of action.

Evidently, in the year 2016, the notion of demonetization preluded by the Prime Minister of India, Mr. Narendra Modi, coursed its fate in many wrong directions where many superfluous news spread across the country. People had to get the currency changed from the banks within a month-duration, meanwhile the hoaxes were on spread that the new currency notes have Nano-chips installed in them to detect the hoarding of black money. The rumors were widely spread and extensively believed. It was so much on spread that it led the Ministry of Finance to come out to deliver official statements regarding the irrelevance of such news. This is just a minor movement that such fraudulent news can cause; the cardinal aspects still lie untouched!

Likewise, any burning topic is expected to get dissected towards a scam phase. This unethical practice of airing the fraud information is done mostly for political, communal, economic, or personal interests. Such randomly spread fake news largely affects the viewers' ideas about the concerning rationality of the prevalent topic. This can again be inferred from the US Presidential election of 2016. Trump's candidature was well-rooted yet was doubted because of the prevalence of the fabrication of irrelevant news such as the involvement of Hillary Clinton in selling weapons to ISIS. The rumors against Trump and Clinton were repeatedly and rapidly aired and are believed to have changed the outcome the 2016 election. This was again a concerning case where fake news played diabolically and had an enormous impact on people's opinions and decisions.

To deal with the unwelcoming impacts of fake news, many techniques can be deployed for detecting and reporting the fake news. For the start, in this project, the experiment was done on a predefined dataset. The dataset that has been used here comprises of two columns. The first column—Statement—contains around 10,241 news article headlines, and the second column—Label—contains the corresponding status of these news articles being either T (True) or F (False) by the origin. The benefit of administering a dataset as humongous as this is that the machine learning models that are implemented will get sufficient amount of data to get trained on and to be tested against.

In the earlier phase of this project, three different and independent approaches were implemented, namely—Linear regression, Support Vector Machine, and Naïve Byes—for the purpose of detection and prediction of whether the provided news is fake or not. However, a much strong and influential version of this project is worked upon where the implementation of Neural Networks is incorporated to furthermore improve the accuracy of the project. A comparison will be established between both the approaches to identify the best of both.

The very first model that has been implemented is based on the supervised learning algorithm known as Linear Regression. In the Linear regression concept, the association between an independent (predicting) and a dependent (response) variable is sought out, that's why the name. This predicts the dependent value which is based on independent variable. In this, values are predicted within a linear range rather than classifying them into different categories. It is called Linear Regression because it sets linear relation between both the variables. In this algorithm predication error is obtain, which should be as small as possible, where predication error is the distance between the point and regression line. Through Regression Line, this algorithm predicts minimum error. We have used the Logistic Regression approach where we take the dependent variable (to be predicated) as a binary number indicating the outcome as either 1 (meaning True) or 0 (meaning False). In Logistic Regression, the value is expected to lie between the range of 0 to 1 and is therefore fit for labelling, unlike Linear Regression where there lie infinite numbers between a certain range.

Naïve Bayes is a conventional, plane and effective machine learning classifier technique which has its base on the Bayesian Theorem and is customarily used in large number of dataset problem. Since it's based on the Bayesian Theorem, one can comprehend the association of this technique with probability. It is the way of classifying conditional probability, where it determines that an attribute in a class is not related to other attribute in the same class. Naïve Bayes classification is widely used in text classification because of its independence rule, and has higher success rates in comparison with a few existing algorithms.

The Support Vector Machine algorithm has been one of the most popular classifier algorithm till now. The reason behind this is the fact that it is a supervised algorithm which is used for Regression and Classifier both. Supervised approach works on labelled dataset, which identify the input; on the basis of input it provides output. This algorithm creates hyperplane between two class. It is the best splitting boundary between the data. There is a widest margin between the support vector. When compared, the accuracy offered by Naïve Bayes was around 63.15%, by Linear Regression was around 63.50% and that offered by Support Vector Machine was around 71.58%. It can be easily inferred that SVM is more efficient among these three models.

After successfully implementing the above algorithms, the strive to achieve more success in terms of accuracy led to the incorporation of the concept of Neural Networks. In that regard, the first Deep Neural Network model that is seen upon is the Long Short Term Memory model (LSTM). Just because Artificial Neural Networks had a short hand in terms of memory, Recurrent Neural Networks (RNNs) were instituted. Conceptually, adding more to the feedforward neural networks, RNNs came up with an internal memory element by propagating the outputs of a node in two directions—one from inputs to outputs, the other from outputs to inputs—with the latter forming loops over the nodes and permitting a memory state at each node. LSTM is a superlative version of RNN and has enabled the memory storage of a node to store the information for short and long duration both, hence avoiding the Vanishing Gradient problem of RNNs. In this dire problem, a complex network with a heavy number of layers finds it difficult to retain the information floating at the distant layers. A classic LSTM network comprises various memory blocks known as cells. From each cell, two states are transferred to the neighbor cell. One is termed as the cell state and the other is termed as the hidden state. These cells are accountable for remembering data and the operations done on the data. This is feasible because of three features of LSTM, called the gates—Forget gate, Input gate and Output gate. Forget gate is accountable for expelling the information from the cells. The Input gate is accountable for the reception of information by the cell. The Output gate is accountable for selecting relevant information from a cell and displaying it out as an output.

A Convolutional Neural Network (CNN) is conceptually made up of two chief segments—Feature Extraction and Classification. Feature extraction includes the convolution layer where the term convolution corresponds to a mathematical operation. An input image is operated by a convolution using a filter, which is generally a small grid-matrix containing certain values. This matrix performs operations with the originally fetched input and produces an output matrix. The output thereafter is supposedly passed through an activation function. The other significant Feature Extraction stage is the pooling layer. When the output from the convolution layer is extracted, it is generally passed through a pooling layer which provides a summarized model of the input features. The pooling layer also reduces a networks overall number of parameters and computations and hence clips the time of model's training.

The Gated Recurrent Units (GRUs) came up as a superlative of the primitive Recurrent Neural Networks, and have featured two gates, namely—Update gate and Reset gate. One of the prime aim for the introduction of GRU is to solve the vanishing gradient problem. The model retains the potential to manipulate and control the flow of data thereby refining the output using these two vectors. It is almost similar to the LSTM model, but it has some parameters which are even lesser than the LSTM model itself. It also doesn't have an output gate. Networks like Gated Recurrent Units are preferable because like other types of Recurrent Neural Networks, they possess the ability to retain the information for a fine duration of time, and also they can control the data flow because of the gates.

In the end, the comparative analysis will see that the first phase of this project's development, in all the four cases of TF-IDF vector levels—the Count Vector, the Word level, the N-Gram level, and the Character level—the Support Vector Machine has consistently shown the highest accuracy. And, in the second phase, the incorporation of Deep Neural Network Models—CNN, LSTM, and GRU—higher accuracy was furthermore achieved. LSTM showed a phenomenal accuracy hike.

2. Literature Review

Shivam B. Parikh, Pradeep K. Atrey et al. in the paper of theirs emphasized that fake news had been around us since decades, but because of the rapid and worldwide adoption of social-media is nowadays on the peak. That is why it is a popular problem. They analyzed on the basis of Headline or Images people believe on the fake news. In the later part of the paper, they incorporated the research problems in the detection of fraud news and identified variegated platforms which are typically utilized to spread any news material over the globe efficiently. They have analyzed and put the classes of data that are contained in a news article and have discussed the affect that all these classes renders to the reader, and also they made the comprehension of fake news categories significantly easy. Ahead to this, they provided with a brief about the existing detection techniques of the fake news prevalence contrasting them on the basis of their performance. They have effectively described the datasets which are available for the experimental purpose by the fake news analysts and researchers. At last they summed up concluding about the challenges that this field of research revolving around the fake news has [1].

Jiawei Zhang, Bowen Dong et al. Their views on this paper take into consideration that excessive people are on social Media. People blindly believe whatever news is shared on it. Sometime it is the main reason of environment misbalance, which can disturb the political peace. They are trying to detect the fake news by analyzing the data. In this paper there had been introduced a new model to automatically detect the fake news which is named as- FAKEDETECTOR. From the existing information, this model had taken a few necessary attributes to put into use while producing a relevant network model which can comprehend the pattern and perspective of the articles and their sources. Alleged conducts and examinations have been carried over this model based on the real world witnessed dataset of fraud news to check the performance of various existing models with FAKEDETECTOR. The results incurred after the experiments proved the efficiency of this model and its practical use [2].

Marian Cristian Mihăescu explained about the concept of Linear regression, where he explained the application of classifier of business. To conceive real-life knowledge about the reader, Mihăescu had used a typical data mining concept implementation. The prime motive was to come up with a cutting edge model which can be used for prediction of knowledge limits and reaches of a reader. The ultimate result of this procedure would be the production of a significant information intended for the reader. The production of this procedure is relevant in its delivery when tended for a virtual platform of E-Learning incorporating necessary amendments of the required attributes. Implementation of such some complex analytical proceedings needs many things. Initially, a virtual E-Learning platform was a requirement. Then the requirement of structured data pertaining the perceived activities by the reader was put-forth, which represented the required user input for the implemented modelling technique. This modelling technique has a significant role in the project. This paper had the implementation of the much common Linear regression technique. This process, in its end, will yield an attribute set—the set of coefficients contained by the linear equation of Linear Regression—which is utilized for producing the value of the dependent variable. As the model comes to its working, it was useful in the computation of dependent variable's value which further presents the values of all other required parameters. An absolute understanding of the parameters was done. The further scope of this paper was to put all the implementations and findings to a general implementation which would also allow the implementation of other efficient machine learning algorithms to improve the current E-Learning system [3].

Shivam Bansal, explained the basic use of NLP task in different business problem "Text Classification". In this Paper he explained the steps of implementation of "Text Classification". He explained four Steps: Dataset preparation, Feature Engineering, Model Training, Improve Performance of text classifier. In his Paper he gives the accuracy on the parameter TF-IDF vector as a feature. He had given the accuracy in Count vector, Word Vector, n-gram Vector, Character Vector [4].

Xinyi Zhou and Reza Zafarani, highlighted the considerable hike in the prevalence of fake news and also its internal effects on the democracy, decisions, and social well-being have invited the ever-demanded requirement for a detailed study, detection and a stringent action over such fraud news. In this research paper a brief analysis of fake news has been done in an extensive and consistent manner. The Analysis of the research determines and points out some of the fundamental ideas across different regulations, for instance, psychology and social science, to make it easier to distinguish between the real and fake news the idea of interdisciplinary research is promoted. Current fake news is analyzed and the evaluation is made for the reference. There are four aspects that are brought out from the analysis of fake news: (1) the wrong idea that it provides, (2) the way it is decrypting the news and how the scripting is achieved, (3) how the generation of the required arrangements are done, and (4) the reliability of the sources or the authors from where the news must have come. The information is differentiated in every aspect of the news from the source or the author on the account of how the information is beneficial and interpreted, all the suitable techniques which can be applied, and also the different schemes and approaches that are taken in account. After analyzing the concerns of the fake news that are happening and on the attributes of the fake news, they had listed a few possible further research divisions which might extend their project's limits [5].

Shloka Gilda, in his paper examined various techniques of Natural Language Processing and its applications in the field of fake news detection, which is, deceiving fraud information which are coming from an unreliable source and are totally fake leading to further misunderstanding. A dataset consisting of around 11,000 news articles was taken from certain online sources, then the terms that were applied on the data were- the context free grammar detection and the Tf-IDf (Term Frequency-Inverse Document frequency) of bi-grams. The Dataset then was tested on different algorithms such as Support Vector Machine (SVM), Random-Forests, Bounded-Decision-Trees, Gradient-Boosting and Stochastic-Gradient-Descent. At the end, it was concluded that the TF-IDF of bi-grams experimented into the model of Stochastic-Gradient-Descent gave an accuracy of 77.2%. [6].

Xiao-Xiao Niu and Ching Y. Suen, took the vision of implementing a handwritten-digit identification model to the reality. To receive better results, two prominent classifiers— *Support Vector Machine (SVM), and Convolutional Neural Network (CNN)*— that are known to present remarkable results in their operational fields of pattern recognition are used. This model assigned distinct responsibilities to both the classifiers, where the feature extraction part is designated to the CNN and the recognition of the digits is handled by the SVM. The ultimate combination of these two models enables the feature extraction from the input images followed by the provisions of predictions. The precision rate given by the model was around 99.81% which is a material achievement [7].

Ben Athiwaratkun and Jack W. Stokes, Considering the glitches presented by the malicious software, this project concludes the potential of the semi supervised learning in the identification of such software. Various substantial models such as Long Short Term Memory (LSTM), Convolutional Neural Network (CNN), and Gated Neural Network (GRU) are used in this project. At last the conclusion laid states that the implementation of Long Short Term Memory with the temporal-max-pooling had rendered better performance of about 31.3%.

3. Statement of Problem

In recent years, fake news has been reported for having significant impact on local and global issues leaving people in a commotion. 70 percentage of the youth are totally dependent on Social Media and are willing to base their opinions on whatever they are fed on such platforms. This can create a problem larger than racism, climate-change, or terrorism itself.

4. Objectives

1. To implement various algorithms of Deep Neural Network

   1.1. Convolutional Neural Network (CNN)
   1.2. Long Short Term Memory model (LSTM)
   1.3. Gated Recurrent Units (GRUs)

2. To compare and analyze the accuracy rendered by the Natural Processing Model and the Deep Neural Network.

5. Design Methodology

5.1 Agile Model

Agile ideology has the mix attributes of incremental and iteration models of software development.

Agile model has its say because of its ability to deliver a working product rapidly and incorporate the requirements during the next iteration. Both development and testing activities are concurrent, and it breaks the product into small incremental builds. One of the chief feature of agile is its pair-programming strategy.

Each Iteration involves simultaneous working on these areas–

a. Planning
b. Requirement Analysis
c. Designing
d. Coding and Building
e. Testing

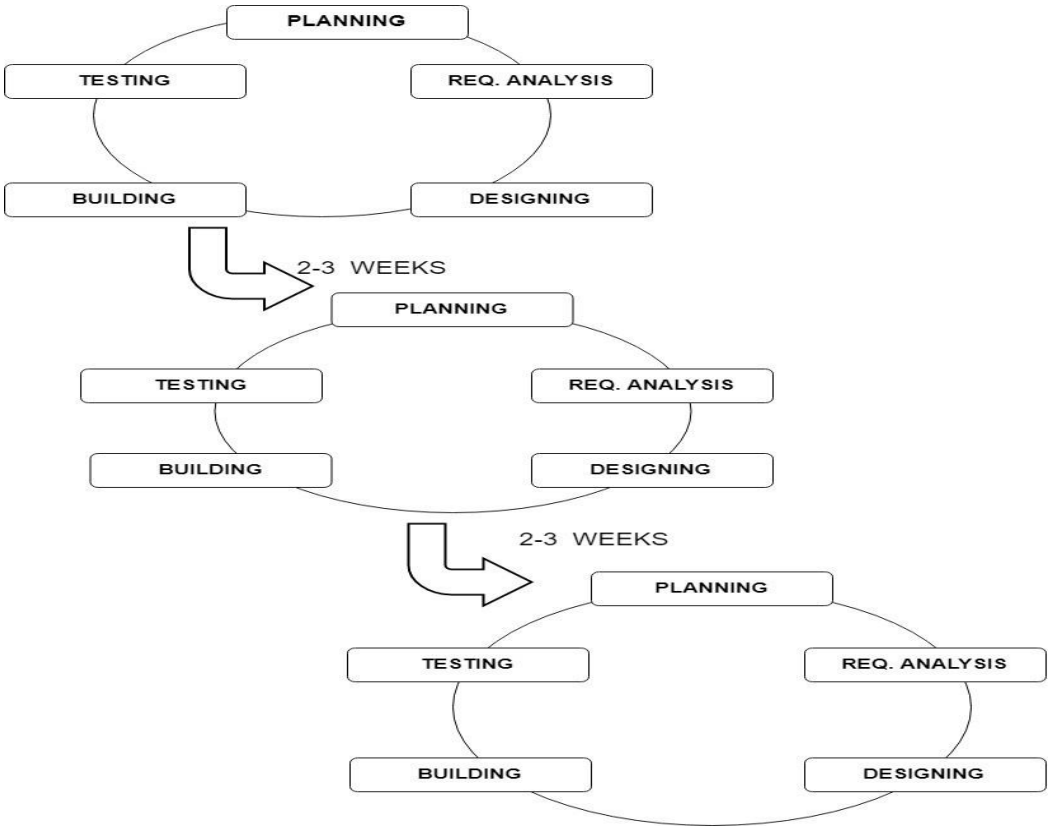

Figure 1: Agile Model

5.2 Dataset

For the start, the experiment was done on a predefined dataset. The dataset that has been used here comprises of two columns. The first column—Statement—contains around 10,241 news article headlines, and the second column—Label—contains the corresponding status of these news articles being either T (True) or F (False) by the origin. The benefit of administering a dataset as humongous as this is that the machine learning models that are implement will get sufficient amount of data to get trained on and to be tested against.

	Statement	Label
2	Says the Annies List political group supports third-trimester abortions on demand.	F
3	When did the decline of coal start? It started when natural gas took off that started to begin in (President George W.) Bushs administration.	T
4	Hillary Clinton agrees with John McCain "by voting to give George Bush the benefit of the doubt on Iran."	T
5	Health care reform legislation is likely to mandate free sex change surgeries.	F
6	The economic turnaround started at the end of my term.	T
7	The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW) faculty fired during the last two decades.	T
8	Jim Dunnam has not lived in the district he represents for years now.	F
9	I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate.	T
10	However, it took $19.5 million in Oregon Lottery funds for the Port of Newport to eventually land the new NOAA Marine Operations Center-Pacific.	T
11	Says GOP primary opponents Glenn Grothman and Joe Leibham cast a compromise vote that cost $788 million in higher electricity costs.	T
12	For the first time in history, the share of the national popular vote margin is smaller than the Latino vote margin.	T
13	Since 2000, nearly 12 million Americans have slipped out of the middle class and into poverty.	T
14	When Mitt Romney was governor of Massachusetts, we didnt just slow the rate of growth of our government, we actually cut it.	F
15	The economy bled $24 billion due to the government shutdown.	T

Figure 2: Dataset- comprise of two columns—Statement and Label.

5.3 Linear Regression

In Linear regression concept, the association between an independent (predicting) and a dependent (response) variable is sought out, that's why the name. This predicts the dependent value which is based on independent variable. In this, values are predicted within a linear range rather than classifying them into different categories. It is called Linear Regression because it sets linear relation between both the variables. In this algorithm predication error is obtain, which should be as small as possible, where predication error is the distance between the point and regression line. Through Regression Line, this algorithm predicts minimum error. We have used the Logistic Regression approach where we take the dependent variable (to be predicated) as a binary number indicating the outcome as either 1 (meaning True) or 0 (meaning False).

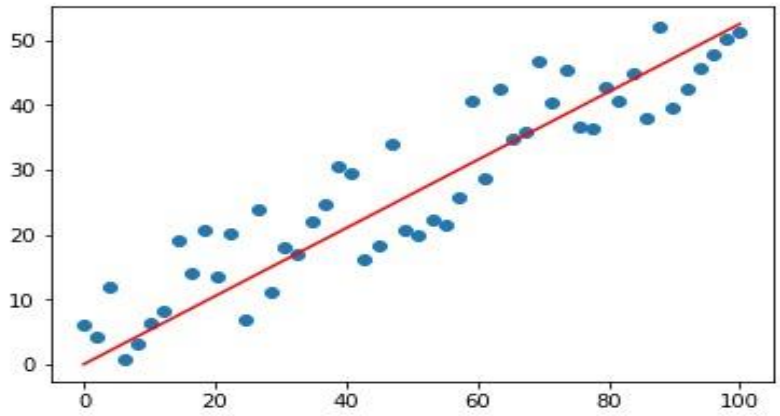

Figure 3: Linear regression [12]

In Logistic Regression, the value is expected to lie between the range of 0 to 1 and is therefore fit for labelling, unlike Linear Regression where there lie infinite numbers between a certain range.
It is given by:

$$\boldsymbol{g(E(y)) = ¢ + £x1 + ¥x2} \quad (1)$$

Here, the function g( ) is called as link function. Function E(y) is the expectation of the response variable and the α, β, γ are to be predicted. The link function links the expectation of y to the linear predictor (α + βx1 + γx2).

5.4 Naïve Bayes

Naïve Bayes is a conventional, plane and effective machine learning classifier technique which has its base on the Bayesian Theorem and is customarily used in large number of dataset problem. Since it's based on the Bayesian Theorem, one can comprehend the association of this technique with probability. It is the way of classifying conditional probability, where it determines that an attribute in a class is not related to other attribute in the same class. Naïve Bayes classification is widely used in text classification because of its independence rule, and has higher success rates in comparison with a few existing algorithms.

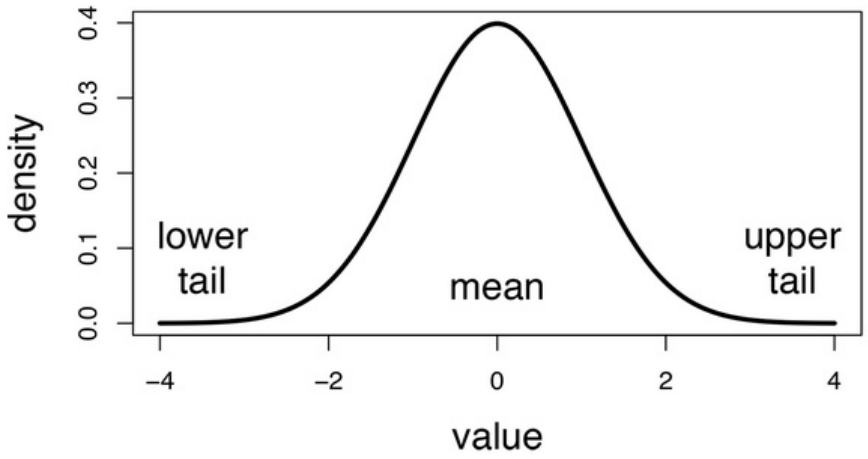

Figure 4: Naïve Bayes [14]

It is given by:

$$\boldsymbol{p(\frac{c}{d}) = \frac{p(c).p(\frac{d}{c})}{p(d)}} \quad (2)$$

Here,

P(c|d) reflects a class posterior probability given the predictor.
P(c) reflects the class prior probability.
P(d|c) reflects the predictor probability given the class.
P(d) reflects the predictor prior probability.

5.5 Support Vector Machine

The Support Vector Machine algorithm has been one of the most popular classifier algorithm till now. The reason behind this is the fact that it is a supervised algorithm which is used for Regression and Classifier both. Supervised approach works on labelled dataset, which identify the input; on the basis of input it provide output. This algorithm creates hyperplane between two class. It is the best splitting boundary between the data. There is widest margin between the support vector.

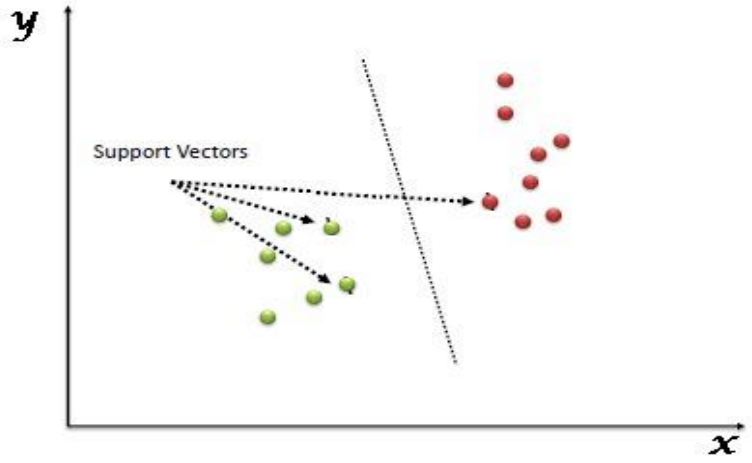

Figure 5: Support Vector Machine [11]

For Hyperplane, it is given by:

$$\boldsymbol{w^t x = 0} \tag{3}$$

For line, it is given by:

$$\boldsymbol{y = ax + b} \tag{4}$$

5.6 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is conceptually made up of two chief segments—Feature Extraction and Classification. Feature extraction includes the convolution layer where the term convolution corresponds to a mathematical operation. An input image is operated by a convolution using a filter, which is generally a small grid-matrix containing certain values. This matrix performs operations with the originally fetched input and produces an output matrix. The output thereafter is supposedly passed through an activation function. The other significant Feature Extraction stage is the pooling layer. When the output from the convolution layer is extracted, it is generally passed through a pooling layer which provides a summarized model of the input features. The pooling layer also reduces a networks overall number of parameters and computations and hence clips the time of model's training.

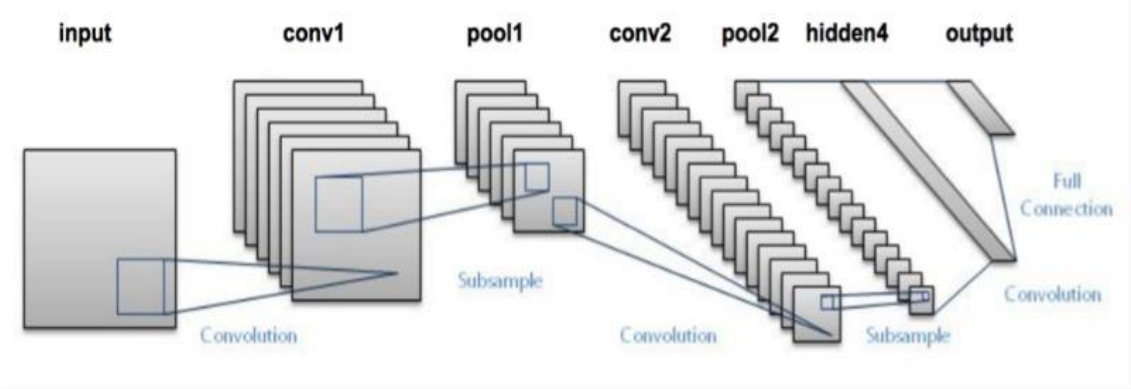

Figure 6: Convolutional Neural Network [11]

5.7 Long Sort Term Memory

LSTM is a superlative version of RNN and has enabled the memory storage of a node to store the information for short and long duration both, hence avoiding the Vanishing Gradient problem of RNNs. In this dire problem, a complex network with a heavy number of layers finds it difficult to retain the information floating at the distant layers. A classic LSTM network comprises various memory blocks known as cells. From each cell, two states are transferred to the neighbor cell. One is termed as the cell state and the other is termed as the hidden state. These cells are accountable for remembering data and the operations done on the data. This is feasible because of three features of LSTM, called the gates—Forget gate, Input gate and Output gate. Forget gate is accountable for expelling the information from the cells. The Input gate is accountable for the reception of information by the cell. The Output gate is accountable for selecting relevant information from a cell and displaying it out as an output.
It is given by:

$$\bar{c}t = \tanh(Wc[at - 1, xt] + bc) \quad (5)$$

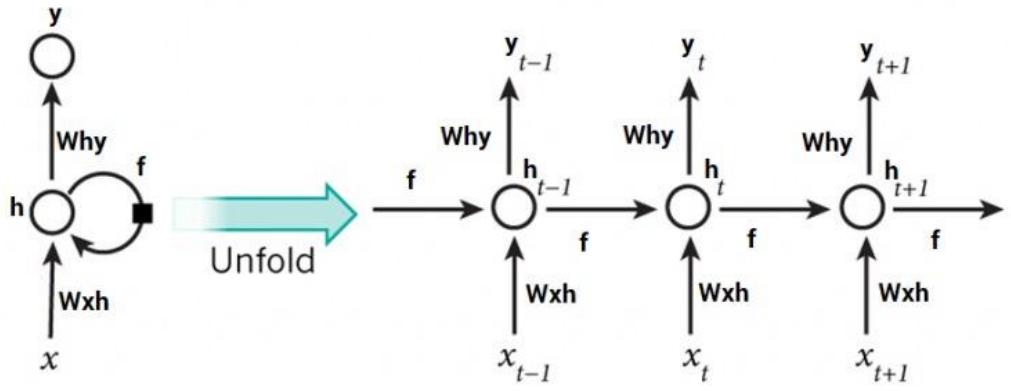

Figure 7: Long Sort Term Memory [9]

5.8 Gated Recurrent Units

The Gated Recurrent Units (GRUs) came up as a superlative of the primitive Recurrent Neural Networks, and have featured two gates, namely—Update gate and Reset gate. One of the prime aim for the introduction of GRU is to solve the vanishing gradient problem. The model retains the potential to manipulate and control the flow of data thereby refining the output using these two vectors. It is almost similar to the LSTM model, but it has some parameters which are even lesser than the LSTM model itself. It also doesn't have an output gate. Networks like Gated Recurrent Units are preferable because like other types of Recurrent Neural Networks, they possess the ability to retain the information for a fine duration of time, and also they can control the data flow because of the gates.
It is given by:

$$\bar{c}t = tanh(Wc[Gr * ct - 1, xt] + bc) \quad (6)$$

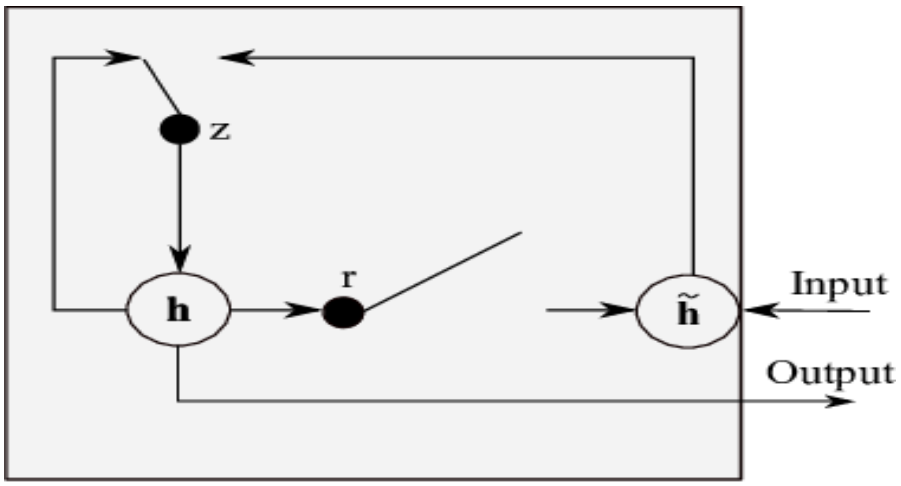

Figure 8: Gated Recurrent Units [13]

5.9 Activation Functions

Activation functions are basic mathematical equation which oversee the overall output of a network. Each neuron uses this function to activate or deactivate its participation in the network so as to achieve the required prediction of the model.

5.9.1 Sigmoid Function

A Sigmoid Function is a non-liner activation function which has an S-letter-shaped curve and has been accustomed because of its range, which lies between (0-1). Because of its range, this function is widely used in the cases where probability is to be predicted.

$$\boldsymbol{A = 1/(1 + e^{-x})} \tag{7}$$

5.9.2 ReLU Function

Rectified Linear Unit or ReLU function gained its popularity because it is comparably faster and involves less sophisticated mathematical operations. In this, if at any input state the value is less than zero, it takes it to be zero and the value can be above or equal to zero.

$$\boldsymbol{A(x) = max(0, x)} \tag{8}$$

5.10 Process Flow

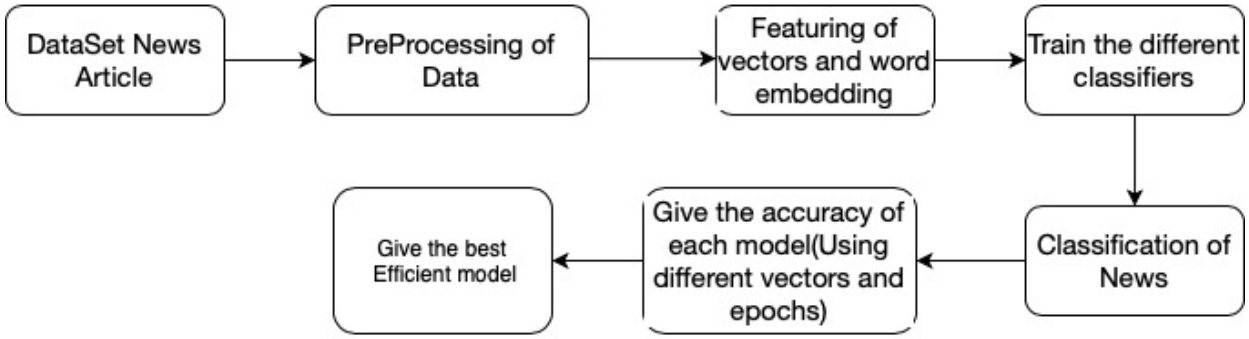

Figure 9: Process Flow

- **Dataset News Article:** For the start, the experiment was done on a predefined dataset. The dataset that has been used here comprises of two columns. The first column—Statement—contains around 10,241 news article headlines, and the second column—Label—contains the corresponding status of these news articles being either T (True) or F (False) by the origin.

- **Preprocessing of Data:** Data preprocessing involves steps such as data encoding or tokenization. In data encoding, words are transformed into numeric representations which are machine comprehendible. In tokenization, the text is split into a desired set of relevant tokens.

- **Featuring of Vectors and Word Embedding:** Term Frequency-Inverse Document Frequency (TF-IDF) is a widely accepted text transformation algorithm which ciphers a text into coherent numeric representations. The type representation of words and documents making use of a dense vector representation is word embedding.

- **Train the different Classifier:** The chief focus of machine learning is to train the implemented algorithms over the training data, to learn its behavior from the data and then make predictions from that data. The implementation of a usual machine learning algorithm in Python using Sci-kit learn, a machine learning tool for Python.

- **Classification of News**: This term will focus on training a supervised learning text classification model in Python**.**

- **Give the Accuracy of Each Model:** Individually evaluate the performance of the algorithms under the various levels of Tf-IDf. This shall serve as the basis of the result analysis and will help in the comparative analysis.

- **Give the Best Efficient Model:** In the last phase, the best machine learning algorithm will be chosen based on the accuracy offered by each of them.

6. Implementation

6.1 Pseudo Code:

1. Start.
2. Create a file pointer (dataset).
3. Create two variables to save text and labels.
4. Locate and save the text in 1st variable.
5. Locate and save the label in 2nd variable.
6. Store the length of the dataset in any variable x.
7. Create a new local Pandas Dataframe.
8. Traverse and append each text and replace the Label “T” with 1 and “F” with 0.
9. Split the Dataset into Training and Testing Data.
10. Encode the data for pre-processing.
11. Create different vectors using built in functions (count, word, char, n-gram).
12. Call the train_model() function and pass the different vectors.
13. Train using all the six models (Linear Regression, SVM, Naive Bayes, CNN, GRU, LSTM).
14. Get the accuracy of each model using different epochs (1, 3, 5, 10, 25).
15. Give the best accuracy model.

6.2 Result Analysis:

6.2.1 Phase One: Natural Language Processing Models

i. **Count Vectors:** Count Vectorization involves counting the frequency of each word appearing in a given document (an article, a book, or even a paragraph). Python’s Sci-kit learn library has a tool called CountVectorizer to accomplish this.

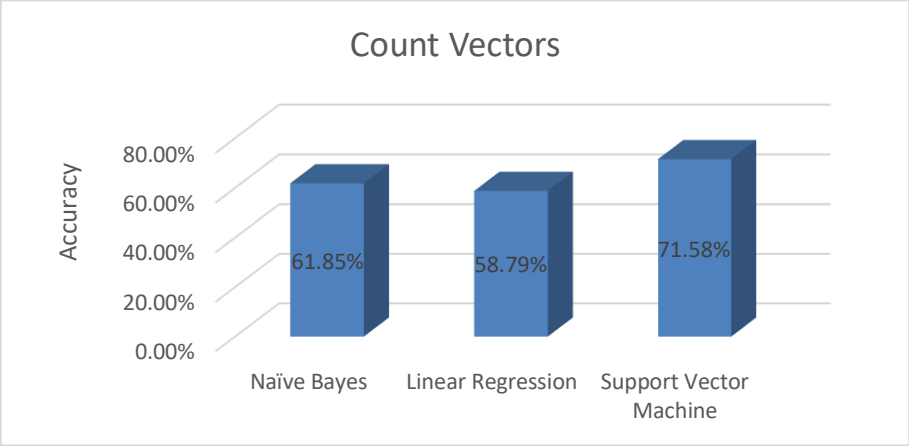

Figure 10: Count vector: Accuracy comparison.

The figure depicts a comparative analysis of the Accuracies of all three models that were implemented in the project taking count vector as the base. Clearly, the accuracy offered by Naïve Bayes is 61.85%, by Linear Regression is 58.79% and that offered by Support Vector Machine is 71.58%. We can infer that SVM Algorithm is comparatively more efficient.

ii. **WordLevel Tf-IDf:** It has the matrix which represents the Tf-IDf scores of each term in distinct documents. This algorithm counts the frequency of words appearing in a document.

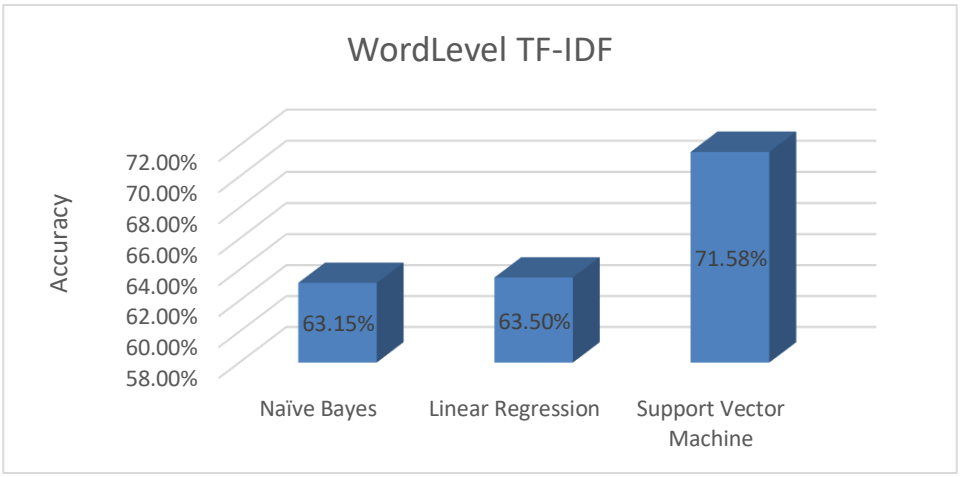

Figure 11: WordLevel TF-IDF: Accuracy comparison.

The above figure depicts a comparative analysis of the Accuracies of all three models that were implemented in the project taking WordLevel TF-IDF as the base. Clearly, the accuracy offered by Naïve Bayes is 62.60%, by Linear Regression is 63.40% and that offered by Support Vector Machine is 71.58%. We can infer that SVM Algorithm is comparatively more efficient.

iii. **N-Gram Vectors:** N-grams are supposed to be a combination of N terms together in a sequence. The N-gram matrix represents the Tf-IDf scores of N-grams.

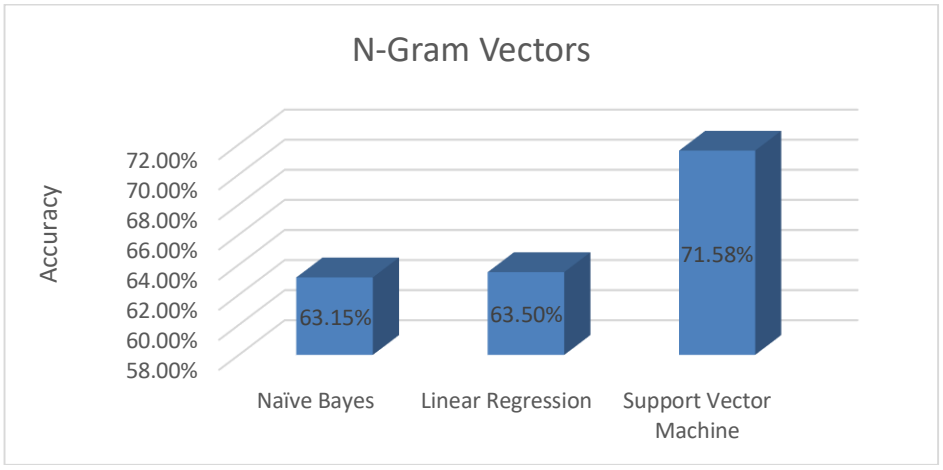

Figure 12: N-Gram Vector: Accuracy comparison.

The above figure depicts a comparative analysis of the Accuracies of all three models that were implemented in the project taking N-Gram Vector TF-IDF as the base. Clearly, the accuracy offered by Naïve Bayes is 62.68%, by Linear Regression is 62.93 and that offered by Support Vector Machine is 71.58%. We can infer that SVM Algorithm is comparatively more efficient.

iv. **CharLevel Vectors:** The matrix which represents the Tf-IDf scores at the character level of n-grams in a given document**.**

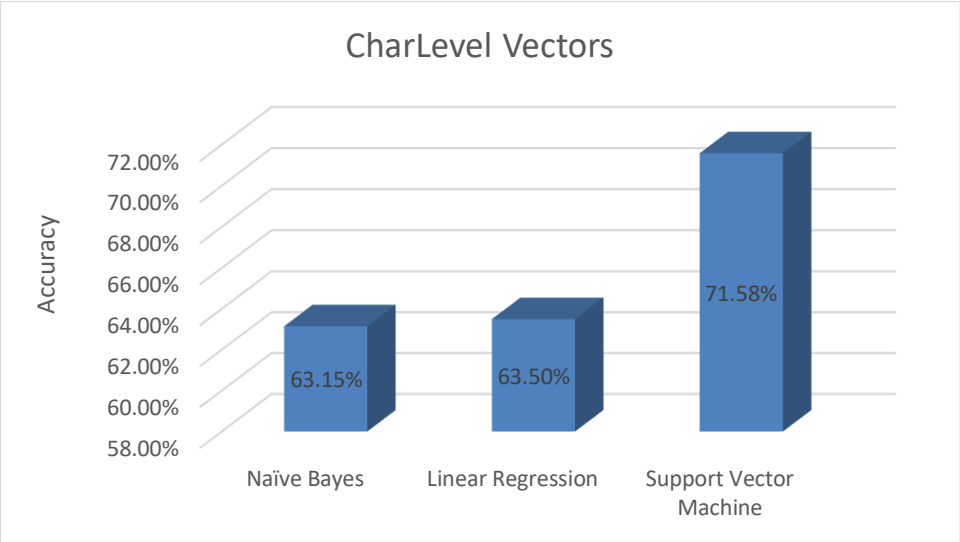

Figure 13: CharLevel Vector: Accuracy comparison.

The above figure depicts a comparative analysis of the Accuracies of all three models that were implemented in the project taking CharLevel Vector as the base. Clearly, the accuracy offered by Naïve Bayes is 63.15%, by Linear Regression is 63.50% and that offered by Support Vector Machine is 71.58%. We can say that SVM is more efficient.

6.2.2 Phase Two: Deep Neural Network Models

i. Epoch vs Accuracy

CNN: It can be easily deduced from the below line graph that the gradual increase in the epochs led to the increase in the accuracy offered by the CNN model, with the epochs ranging from 1 to 25, and the corresponding accuracy ranging 68.67% to 81.77%.

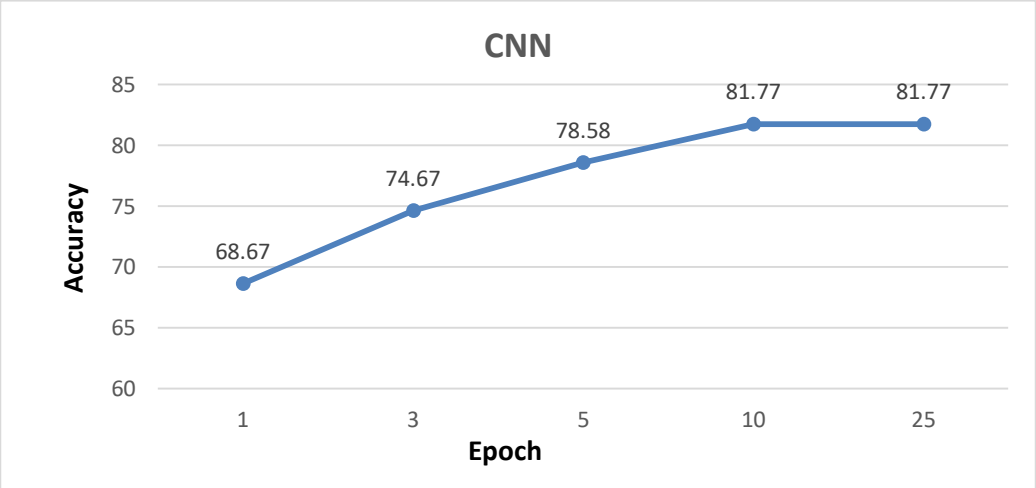

Figure 14: Epoch vs Accuracy- CNN

LSTM: It can be fairly inferred from the below line graph that the gradual increase in the epochs led to the increase in the accuracy offered by the LSTM model, with the epochs ranging from 1 to 25, and the corresponding accuracy ranging 78.67% to 91.25%.

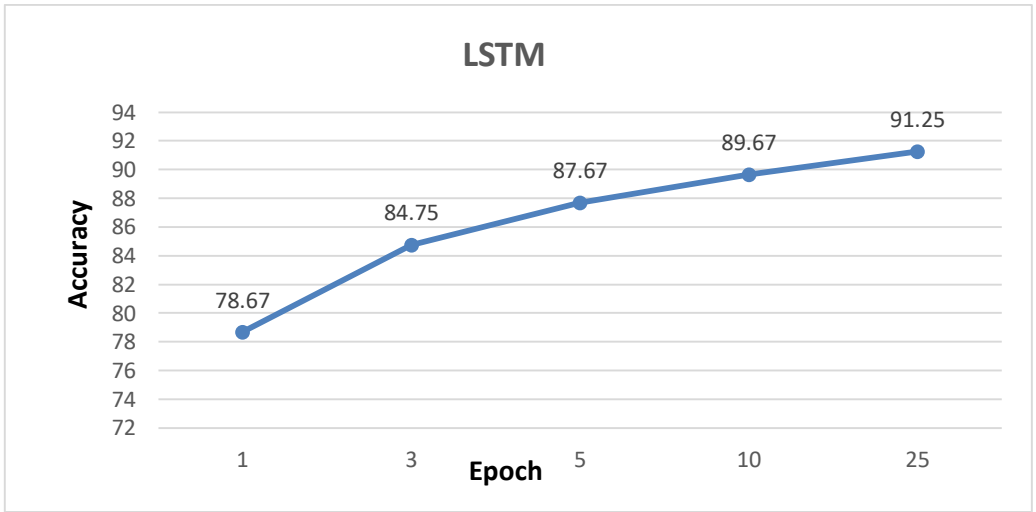

Figure 15: Epoch vs Accuracy- LSTM

GRU: It can be fairly inferred from the below line graph that the gradual increase in the epochs led to the increase in the accuracy offered by the LSTM model, with the epochs ranging from 1 to 25, and the corresponding accuracy ranging 78.67% to 91.25%.

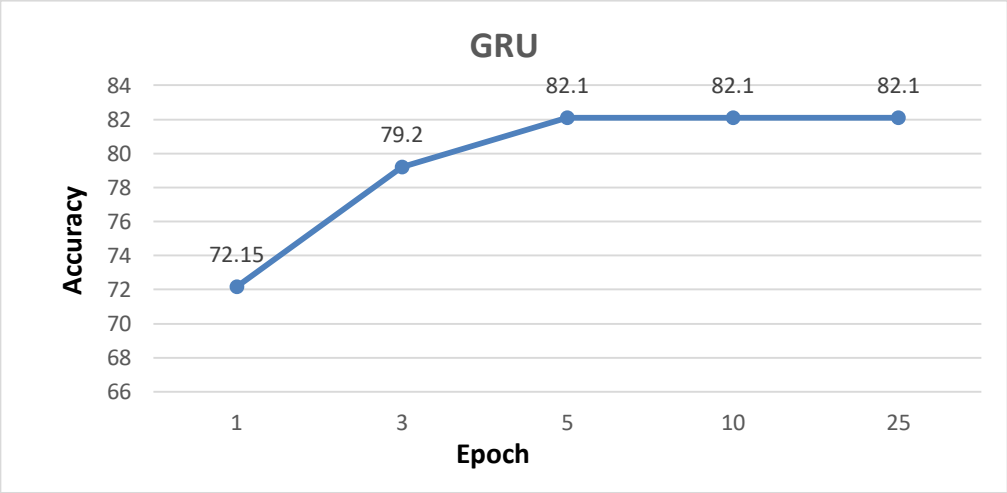

Figure 16: Epoch vs Accuracy- GRU

ii. Accuracy Comparison on the basis on epochs

1-Epoch: It is evident from the below graph that CNN offered 68.67% of accuracy, whereas, LSTM offered the highest accuracy that is of 78.67%, and GRU offered 72.15% of accuracy at the first epoch.

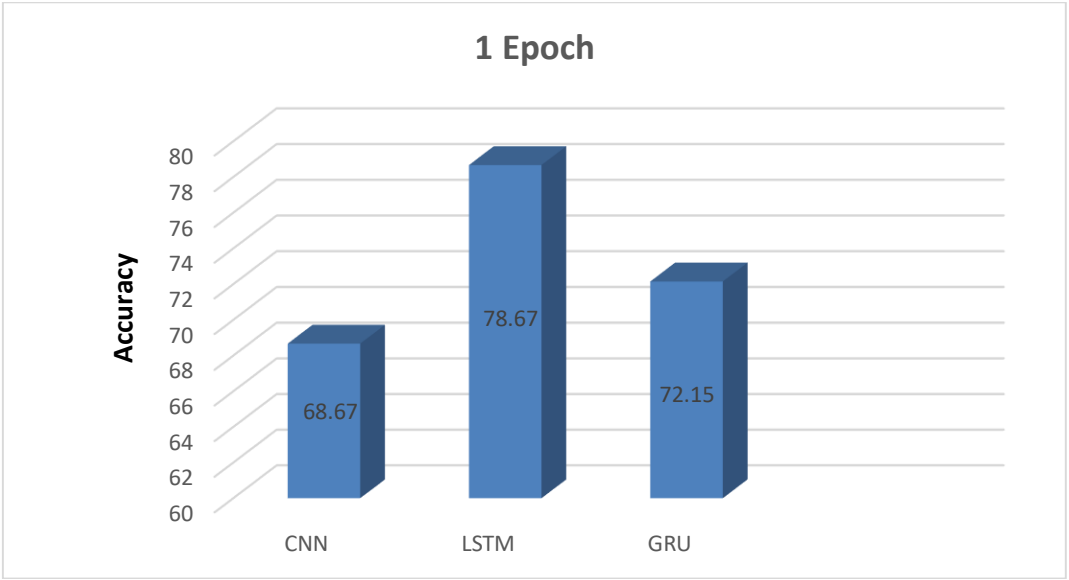

Figure 17: 1-Epoch: Accuracy Comparison

3-Epoch: It is evident from the below graph that CNN offered 74.67% of accuracy, whereas, LSTM yet again offered the highest accuracy that is of 84.75%, and GRU offered 79.2% of accuracy at the third epoch.

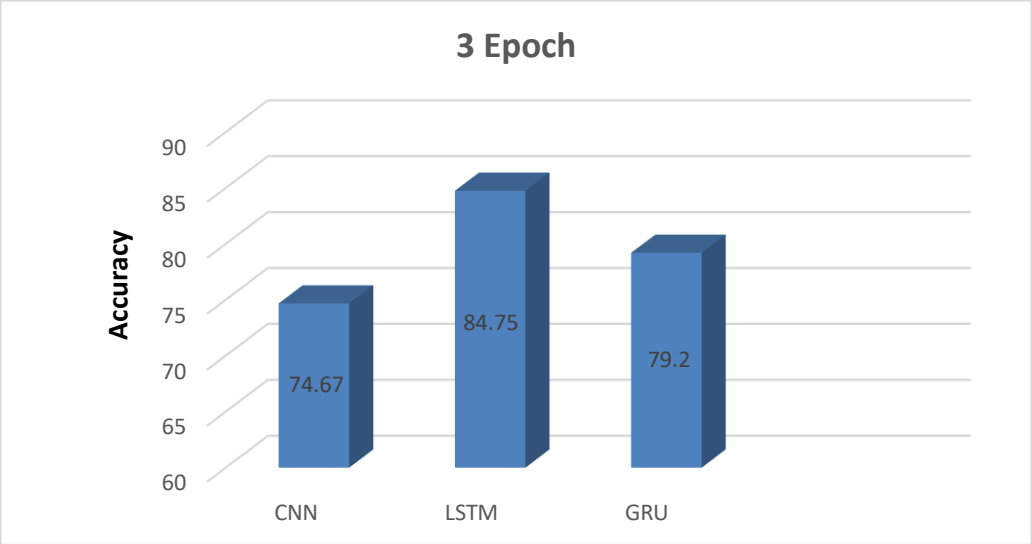

Figure 18: 3-Epoch: Accuracy Comparison

5-Epoch: It can be concluded from the below graph that CNN offered 78.58% of accuracy, whereas, LSTM yet again offered the highest accuracy that is of 87.67%, and GRU offered 82.67% of accuracy at the fifth epoch.

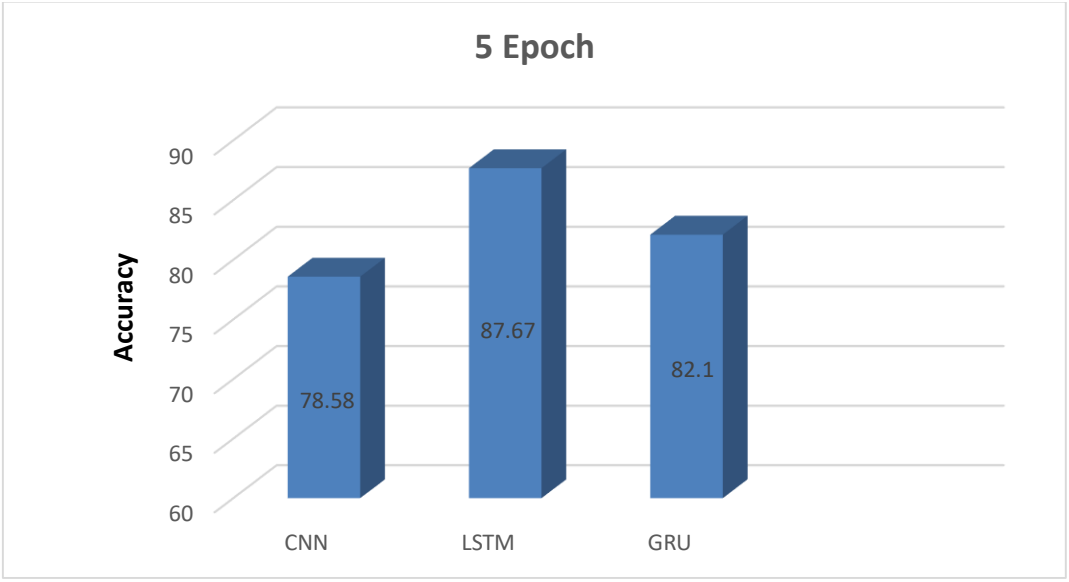

Figure 19: 5-Epoch: Accuracy Comparison

10-Epoch: It can be concluded from the below graph that CNN offered 81.77% of accuracy, whereas, LSTM yet again offered the highest accuracy that is of 89.67%, and GRU offered 82.1% of accuracy at the tenth epoch.

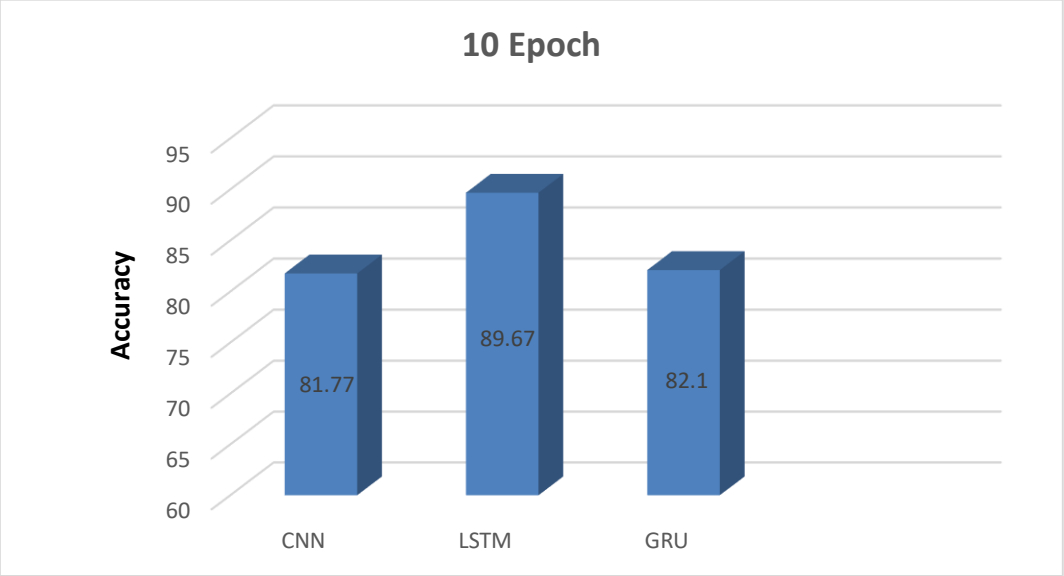

Figure 20: 10-Epoch: Accuracy Comparison

25-Epoch: It can be concluded from the below graph that CNN offered 81.77% of accuracy, whereas, LSTM yet again offered the highest accuracy that is of 91.25%, and GRU offered 82.1% of accuracy at the tenth epoch.

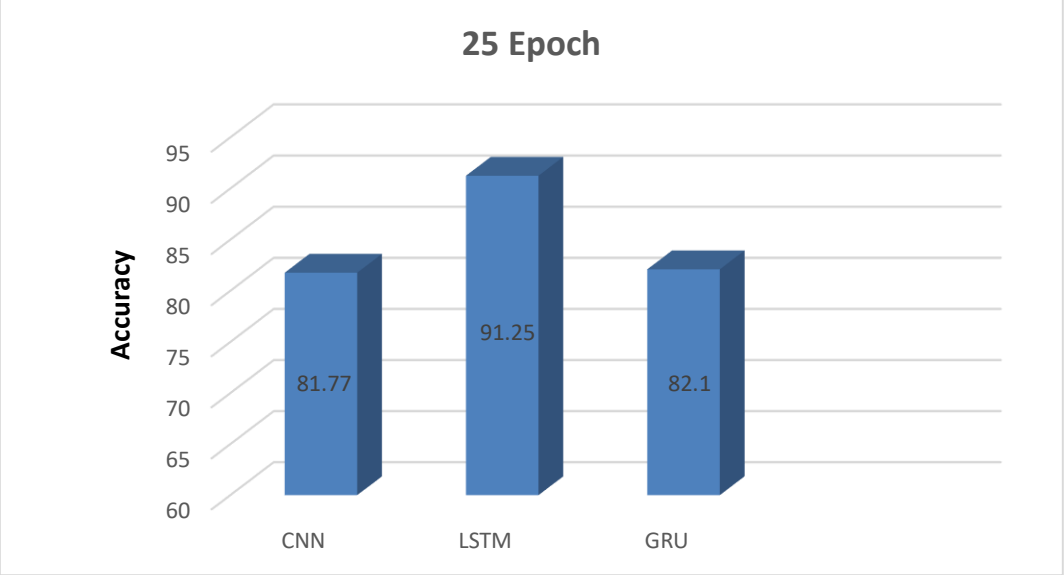

Figure 21: 25-Epoch: Accuracy Comparison

7. Pert Chart

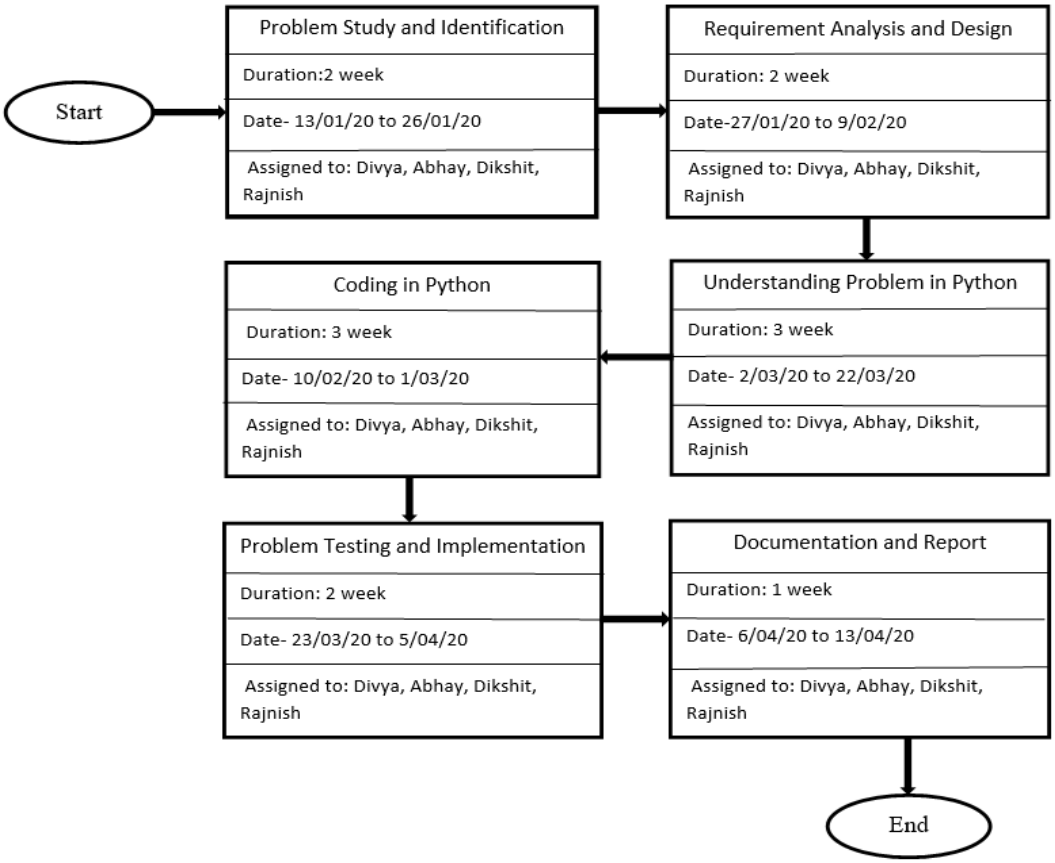

Figure 22: Pert Chart

8. Conclusion and future scope

This could be clearly deduced from the result analysis that in the first phase of this project's development, in all the four cases of TF-IDF vector levels—the Count Vector, the Word level, the N-Gram level, and the Character level—the Support Vector Machine has consistently shown the highest accuracy, that is, around 71.58%. And, in the second phase, the incorporation of Deep Neural Network Models—CNN, LSTM, and GRU—higher accuracy was furthermore achieved. LSTM showed a phenomenal accuracy hike around 91.25%.

Hence, among all the applied algorithms, in order to efficiently determine hoax or fake news, the implementation of LSTM to get maximum robustness in the outcomes will be preferred.

REFERENCES

[1] Shivam B. Parikh, Pradeep K. Atrey, *"Media-Rich Fake News Detection: A Survey"*. https://www.albany.edu/~sp191221/publications/Fake_Media_Rich_News_Detection_A_Survey.p, [Accessed: August 28,2019]

[2] Jiawei Zhang, Bowen Dong, *"Fakedetector", Effective Fake News Detection with Deep Diffusive Neural Network*, August 10, 2016. [Online], https://arxiv.org/pdf/1805.08751.pdf, [Accessed: August 28,2019]

[3] Marian Cristian Mihăescu, *"Classification of Learners Using Linear Regression"* [Online], http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.231.7320&rep=rep1&type=pdf, [Accessed: August 28,2019]

[4] Shivam Bansal, *"A Comprehensive Guide to Understand and Implement Text Classification in Python"* August 10.2016. [Online], https://arxiv.org/pdf/1805.08751.pdf, [Accessed: August 28,2019]

[5] XINYI ZHOU, *"Fake News: A Survey of Research, Detection Methods, and Opportunities"*, [Online], https://arxiv.org/pdf/1812.00315.pdf [Accessed: August 30, 2019]

[6] Shloka Gilda, *"Evaluating Machine Learning Algorithms for Fake News Detection"*, 2017 IEEE 15th Student Conference on Research and Development (SCOReD) [Online]. [Accessed: - September 3,2019]

[7] Xiao-Xiao Niu, Ching Y. Suen, *"A Novel Hybrid CNN–SVM Classifier for Recognizing Handwritten Digits," Elsevier,* vol. 45, no. 4, pp. 1318-1325, April 2012.

[8] Ben Athiwaratkun and Jack W. Stokes, *"Malware classification with LSTM and GRU language models and a character-level CNN".* In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pages 2482–2486. IEEE, 2017.

[9] SUNIL RAY, "https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/" MAY 29, 2017.

[10] SUNIL, "https://www.analyticsvidhya.com/blog/2017/06/architecture-of-convolutional-neural-networks-simplified-demystified/" MAY 29, 2017.

[11] SUNIL, "https://www.analyticsvidhya.com/blog/2017/05/neural-network-from-scratch-in-python-and-r/" MAY 29, 2017.

[12] "*https:// en.wikipedia.org/ wiki/ Linear_regression*".

[13] "https:// www.researchgate.net/ figure/ A-gated-recurrent-unit-GRU-The-update-gate-z-decides-if-the-hidden-state-is-to-be_fig4_322243714".

[14] "https:// www.quora.com/ What-is-the-difference-between-Gaussian-Naive-Bayes-and-Bayesian-networks".

A APPENDIX I PROJECT CODE