Efficient Machine learning techniques for hoax Content Classification

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE

OF

MASTER OF TECHNOLOGY

IN

INFORMATION SYSTEMS

Submitted By:

NEHA KIRAR

(2K17/ISY/11)

Under the supervision of

MR. JASRAJ MEENA

(Assistant Professor Department of CSE)

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

JUNE, 2019

CANDIDATE’S DECLARATION

I Neha kirar, Roll No. 2K17/ISY/11 student of M.Tech Information Systems, hereby declare that the project Dissertation titled “Efficient Machine learning techniques for hoax content classification” which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi Neha kirar

Date:

CERTIFICATE

I hereby certify that the Project Dissertation" Efficient Machine learning techniques for hoax content classification" which is submitted by Neha kirar, Roll No 2K17/ISY/11 Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi **Mr. Jasraj Meena**

Date: **SUPERVISOR**

ACKNOWLEDGEMENT

I express my gratitude to my major project guide Mr. Jasraj Meena, Assistant Professor Department of CSE, Delhi Technological University, for the valuable support and guidance he provided in making this major project. It is my pleasure to record my sincere thanks to my respected guide for his constructive criticism and insight without which the project would not have shaped as it has.

I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.

Neha Kirar

Roll No. 2K17/ISY/11

M.Tech (Information Systems)

E-mail: emailneha21@gmail.com

ABSTRACT

Hoax news has been floating all over the social media much faster than real news this creates diversity and confusion in the community. Whereas learning the context from a certain headline is very crucial but most challenging task would be to predict the intention of the user, this prediction would be a stepping stone to detect fake news in the field of natural language processing. In this study, experiments conducted aimed at selecting the best algorithm in classifying hoax and non-hoax news with the number of data in English language using news data from all over the world using text preprocessing methods and machine learning based approaches. Also it has vital applications nowadays at every online social media platform its essential to beware of false information because half or false knowledge is very dangerous and might be serving someone's corrupt intensions. This research includes comparison of existing models and the prediction of possibility of hoax content in a given statement.

Table of Contents

Title **Page no.**

List of Tables

List of Figures

Chapter 1
INTRODUCTION

Information technology is a structural frame of software hardware, users and methods for retrieve, store, transmit information (data).With the Advancement in technology the news network has been influenced the accessibility, its seems online social media are easy to welcome in. As online news networks grow, the quality of news distributed is less. The disseminated news is not all true news and eventually falls into category of Hoax content, which is a type of yellow journalism. Whereas news media are developed so much for a good reason that every one person in this universe gets updated of what is happening in each and every corner of the world. That why its network needs to be wide open and complex so that it covers each and every node of this earth. But this network is also being used to serve some people's bad purposes, for their personal benefits. Hence we have to focus on the classification of hoax content because the lack of correct knowledge might not lead to misleading conclusions and opinions.

1.1 Motivation

Hoax is an information or news that contains uncertain things or which really are not facts that occur [1]. Hoax can also be identified with the following [2]: the news comes from uncertain / untrusted sources. Hoax elements such as Images, photos or videos used are the result of engineering, which uses offensive sentences, and contain political and racial elements. Certain examples of hoax news are news article got viral on Facebook that coconut oil should not be eaten it's dangerous for health and also causes cancer. Another Circulating report describes the mysterious discovery of a dead whale in a Utah dairy farmer's field. Also Circulating report claims that when Jamaican authorities conducted DNA tests on the body of a homeless man found behind a fast food restaurant, they discovered that the man was actually reggae star Bob Marley.

Question is why hoax news comes into existence and why they are more prompt to be suspected and most of the people are getting victimized by these hoax news. So to answer this entire question is we need an understanding; if you are a seeker and you wish for limelight it’s very easy now a days, a very common way to trend over internet, this is deliberately floated and practiced to attract commoner’s attention via news media or online social media. That means hoax news definitely gives you fame whether it is good fame or bad fame that people have to understand. Drawing attention may have different reasons associated with it, it could be to spread misleading information or lie, or may be just to spice it up for people’s gossip**.** Whereas digital media has come back with increased amount of fake news content in it. The main concern of hoax news is that, it keeps the power to diverge society and their views, which leads to affection relations within a nation and internationally as well. Other bad consequences of fake news comprise losses from different aspects, ruining social relation, trolling public pain and sometimes may lead to cybercrimes as well. From the view of science and technology, in today’s era its very significant to make some substantial enhancement in hoax classification problem, A text analytical system can enhance human abilities to spot lies [3]. In text classification, clustering of text documents should be done on their parameters and we need to attain this for better structured information for all the trusted audience and their trust.

1.2 GOAL OF MASTER THESIS

This thesis provides us deep study of hoax content and also presents various classifying methods to distinguish the news statements before we are affected by hoaxes or even spread the hoax. Aiming to progress in identifying hate news pattern classification, in the hope that it will distinguish hoax content from the remaining contents and their sources also plays an essential role source whether it’s from trusted network or not and its certification ensure the originality of the online content. Our anticipated methodology can make optimal collaboration of problem caused by hoax news as it takes place nowadays. While in the education field or general knowledge to be more specific everything is getting compromised by hoax content, so we need to be more concerned about the problem we are facing with and increasing with time. Bringing out the improvement abilities of hoax content classification by researchers in the text mining field with natural language processing on the urgent need of improvement in its performance. The techniques which are used in this study are the natural language processing along with classification system using a machine learning based approach.

1.3 ORGANIZATION OF DISSERTATION

Chapter two give you brief about the topic and how its need is generated also highlighting how useful it could be it also includes literature review of hoax news content. Basic definition of hoax content is presented and is followed by how it is affecting the diversity.

Chapter three includes different ways of hoax content classification. After this different types of preprocessing phases are discussed.

Chapter four presents the processing of research methodology.

The fifth chapter contains the results of our approaches on a benchmarks dataset and validates the results showcasing the probability of truthfulness of a particular statement.

Chapter sixth concludes the thesis and further ideas for future work has been proposed.

Chapter 2
BACKGROUND

With the increased number of users over internet, data has also increased over period of time. One of the widely used Natural Language Processing task in diverse trade problem is text classification. We do classify text to automatically classify text documents into one or many well-defined classes. For example Categorization of news articles into defined topics.

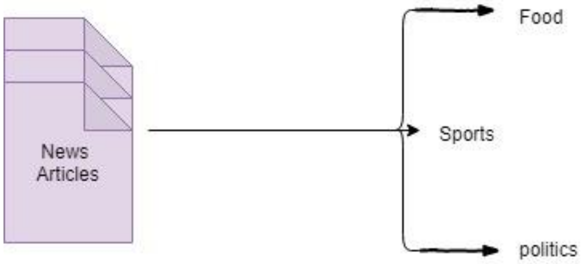


Fig 2.1: Depicting text Classification

Whereas text classification comes under supervised learning process, because the dataset we have is in the form of text with labels of it, and models get trained over those labels to classify. To build a model which can classify an end to end text efficiently should composed of pipeline which consists of following processes under on bundle:

a) Data preprocessing
b) Feature Extraction and engineering
c) Model Selection and training
d) Testing and improving the model performance

2. Literature Review

2.1 News

Reporting of an event or opinion which just has happened, and which should be in public interest or knowledge of information in the form of online social media, newspaper, television and radio. With rapidly technology advancement increments the sources of news, the unclear sources of news also increased with uncertainty whether news is from the fact or just a hoax. The hoax is a deliberate manipulation of news and aims to provide false recognition or understanding [4].Hoax news often found not only in the form of print but also in social media. There are various reasons of intentional hoax divergence, it might varied from influencing speech for a particular community, hate speech, earning money through drawing people attention on a particular medium and so on; hoax news with hate speech has a dissimilar way of writing it generally. The authors use different names to define the same concept that can be observed in our works reviewed. They call it misinformation, rumor, hoax, malicious trend, spam or fake news, but all converge to the same semantic meaning, that is of an information that is unverified, of easy spread throughout the net, with the intention of either block the knowledge construction (by spreading irrelevant or wrong information due to lack of knowledge of the theme) or either manipulate the readers opinion [5][6][7][8].

2.2 Text Mining

Data mining is used for finding the useful information from the large amount of data [9] while text mining is one of division of data mining. In text mining the outcome of mining is in the form of text only that aims at defining sample space of words that represents contents of document so that associativity between documents can be easily carried out. Text mining is very useful now in the field of biomedicine, marketing, online media etc. In text mining there are four interconnected stages because outcome of one stage becomes input for the next stage.

In the text demonstrating data is collected into a collection of data with text type and unknown structures existing in it. Preprocessing will clean up the text and make it free from words and characters which are not adding on into any meaning to the text. After this process text becomes more structured as it clarifies the documents better. At the feature selection stage, features are selected at the time of classification process into desired category.

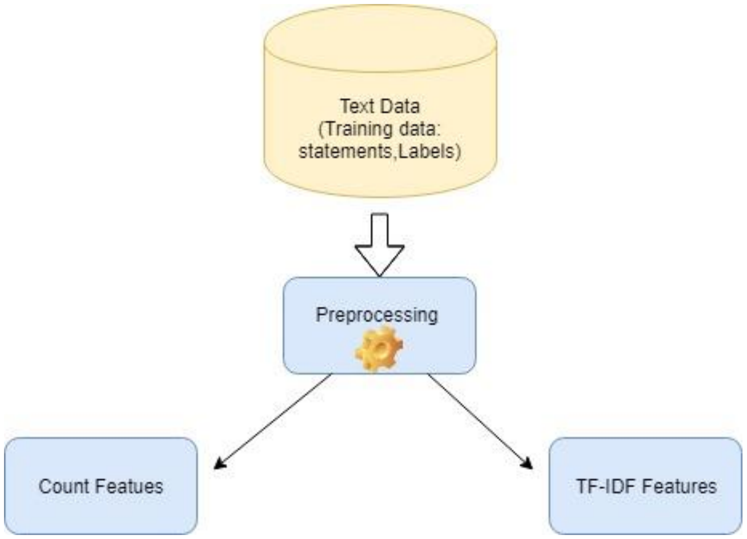


Fig 2.2: Preprocessing techniques on training data

2.3 Text Preprocessing

The text preprocessing of training data is essential to transform data into more meaningful and more understandable form via removing insignificant characters and processed until its set to ready to process further. Commonly, documents can be characterized on variety of basis and have various dimensionality, the characteristics of which are documents contains noise, and unwanted parts, and which gives unstructured layout of the text on the documents. To characterize a text into a particular class, the most fundamental factor that helps in such classification decision is the pattern of the text and that pattern can only be found out in a structure documents.

2.4 Research and improvement of feature words on Count Functions

For text preprocessing we broadly have used two ways one is count features which is used to count the number of occurrences of a substring from a given string. This count will basically is calculated to get an idea how many unique words we have and which words adds the most of the meaning to the article. The better way to enhance the performance of the model is by using term frequency concept in it.

2.5 Research and improvement of feature words weight based on TFIDF algorithm[14]

According to Aizhang Guo, Tao Yang among the various weights measuring algorithm in a document such as entropy function frequency function, Boolean function, The Term Frequency Inverse Document frequency (TFIDF) [15] gives better results over Term Frequency (TF). TF supposed to get a count of number of times a word is repeating in a document to the total number of words in a document, higher the TF count more important that word would be.

$$\text{TF}=\frac{n_{i,j}}{\sum_k n_{i,j}} \qquad (2.5.1)$$

TF is useful somehow but become a major drawback for used helping words like “is”,”am”,”the” because these words are not drawing any meaning to any sentence but their TF count would be highest but of no use.

Then TFIDF is introduced, the main goal of the TFIDF[16] model is to find the important word that can be evaluated by taking the record of the number of documents divided by the number of documents that contain the word w. The inverse data rate determines the weight of the rare words in all documents in the corpus and can be calculated by:

$$\text{idf}(\omega) = \log \frac{N}{df_t} \qquad (2.5.2)$$

Lastly, the TF-IDF in equation 2.5.2 can also be written as the TF multiplied by IDF

$$\omega_{i,j} = tf_{i,j} \times \log \frac{N}{df_i} \qquad (2.5.3)$$

2.6 Bag-of-words(Bow) with n-gram using TFIDF n-gram

Bag of words is a way of representing text data in the form of vector while modeling text with machine learning algorithms. Since dealing with text is messy and machine understands numbers only so we need to fix the unanimous length of input and output. Bag of words is the simple to understand and has been a great success for dealing with problem like text classification and language modelling. The output of bag of words is called a Frequency Vector, and bundle of all extracted distinct words from our document organized as a vector. Bow with n-gram can be represented as follow:

1-gram= [‘is’,’am’,’the’,’jack’]

2-gram= [‘is the’,’jack cleaned’,’by car’]

So on….

4-gram= [‘car was cleaned by’]

Using Bow with tfidf weighing to provide equal importance to each word. Also we can improvise the results by normalizing the TFIDF values for that we use Euclidean normalization method which is represented as:

$$v_{norm}= \frac{v}{\sqrt{v_1^2+v_2^2+\ldots+v_n^2}} \qquad (2.6.1)$$

Chapter 3

RESEARCH METHODOLOGY

3.1. Data Analysis

Data analysis is the acute stage with the objective of keep best out of waste since data analysis consists of the sequence of the process of examining, cleaning, transforming and modeling data with the aim of discovering useful information, reporting conclusions and support decision making. One of the main reasons for analyzing the data to determine the complexity of the data, how data is looking, is that the data is authentic, and to ensure that the data contains the required field.

The data source that we used for this project is LIAR dataset which is a benchmark dataset for hoax as well as real news or statements. It contains 3 files with .csv format for test, train and validation. Below is some description about the data files used for this project.

LIAR: A BENCHMARK DATASET FOR FAKE NEWS DETECTION [19]

This dataset contained 13 variables/columns for train, test and validation sets as follows:

- Column 1: statement of ID
- Column 2: the label. (Label class contains: True, Mostly-true, Half-true, Barely-true, FALSE, Pants-fire)
- Column 3: the statement.
- Column 4: the subject(s).
- Column 5: the speaker.

- Column 6: the speaker's job title.
- Column 7: the state info.
- Column 8: the party affiliation.
- Column 9-13: the total credit history count, including the current statement.
- 9: barely true counts.
- 10: false counts.
- 11: half true counts.
- 12: mostly true counts.
- 13: pants on fire counts.
- Column 14: the context (venue / location of the speech or statement).

To make things less complicated we have chosen only 10 variables from this original dataset for this classification, this will help to categorize true and false values into further categories and others are filtered out for clarity. Also, the other variables can ignored because they are not adding any additional information relevant to news statements.

Below are the columns used to create 3 datasets that have been in used in this project.

- Column 1: Statement (News headline or text).
- Column 2: Label (Label class contains: True, False)
- Column 3: speaker
- Column 4: speaker's job title
- Column 5: location
- Column 6: barely true counts
- Column 7: false counts
- Column 8: half true counts.
- Column 9: mostly true counts.
- Column 10: pants on fire counts

3.2. Data Preprocessing

3.2.1. Handling Missing Fields

Data analysis ensure that no missing data fields are left in the dataset, So there is no need to tackle missing values or fields, but not all datasets are identical, like this research text dataset, the necessary requirement is to delete the whole line or statement because it might give incorrect results.

3.2.2 Text transformation

a) Tokenizing

The tokenizing process is where text data is broken down into specific parts from paragraph, sentence, or words. These specific parts are called as token. Tokenizing terminology is a process of replacing the value of the original value into a value called "token" [10]. For instance if a dataset contains a sentence “It’s raining today” it will generate a set of words consists of “Its”,”raining”,”today”.

b) Case folding

In documents there different ways of writing and different pattern follows hence to equate every kind of writing style we convert every letter of a word to lowercase so that it’s easier to evaluate. For example: “THESIS”, “Thesis” or “THeSis” would be read same and hence converted into “thesis”.

c) Filtering

In Filtering process, we tend to filter out some words from the set of text in an order to filter out the non-making sense words out. The removal of such words is called stop word removal, and it’s a standard operation carried out on data in the form of text that targets to take the words result of tokenizing. Also the representation of document is known a s set of features. Using this stop list model it permits user to create its own word dictionary which contains unused words relevant to a particular document.

d) **Stemming**

In stemming it's a word altering process, where actual word is acquire by eliminating the suffixes of prefixes, suffixes, infixes and combinations of prefixes and suffixes.

e) **TF - IDF weighting**

Term Frequency (TF) is the count of how frequent a particular term has appeared in the document. The higher the number of occurrence of a term in the document, the heavier the weight or provide a greater value of conventionality. Whereas Inverse Document Frequency (IDF) gives you a measure of how widely the term is distributed in the collection of documents. For instance the terms like helping verb "is" would appear at a lot of places and its distribution would be quite wide but does it adds on any sense to the existing meaning of a text? so what IDF suggest greater the distribution value the lesser the IDF value.

3.3 Step Analysis

For analyzing data we have several steps to be taken care of as following:

a) Collect the data on news articles from different time-lapse in desired format we have taken .csv format files for testing training and validation purpose. For deliberate fake news data addition to the original news dataset and labeling the data on the scale of how much event happened likely to be true or false. There are online sites available like theonion.com which helps you to get fake classified data or news.

b) Then the data which is managed in CSV format will passes on to the different stages of pre-processing the text. And the first stage is called tokenizing stage, where all the tokens (a token is also known as individual word of a sentence) of the sentences from the news statements is collected separately.

c) After tokens are formed, case folding process is carried out in which all the letters in each word have a keen focused on, making sure each letter of a word should be in a lowercase only if not its converted into lowercase alphabet, this is done because everyone has different way of writing to normalize these words so that a equalized effect is created by the program on same set of words.

d) The third phase is called filtering process after tokenizing and case folding processing. The filtering process is an elimination process where to have to find some special characters like punctuation mark, commas if any binary address or URL for such matter anything which adds on no sense or improve any meaning of the word is discarded from the dictionary of tokens.

e) Next is stop word removing phase for those words which are repetitive and appears in most of the places in the article is called as stopword here and we have to remove it. The aim is to reduce the load of processing by removing non-impactful words of the sentences.

f) Then next proceeding phase is stemming process in which we get the basic words after removal of its affixes that is suffix or prefix. The aim is to make very basic word to ease out the complexity of the words meaning.

g) The next and last process in the text preprocessing depends on the frequency of occurrence of a word in a document. Using IDF concept we have to calculate weight of each word in the news data to signify the importance of that particular word in the article.

h) There are three separate files of data namely train, test and validate which are explicitly imported in the program. Train data consider 80% data and the rest files have 20% data for testing and validation of data. There are statements which are labeled on the basis of high likely they are true or barely true, false count, half true, mostly true etc. with some more relevant factors as well.

i) Text classifications using machine learning algorithms namely Logistic Regression, Naïve Bayes, support vector machine, Random Forest, stochastic gradient descent classifiers.

j) Next step is to compare the performance between all the classification algorithms on the basis of their f1 score and accuracy. Examples of the application of the text preprocessing process to classify news data can be seen in table (1) below, namely as follows:

Table3.3: Text preprocessing process

Classification Process	**Resultant**
Preliminary Data	Research shows that a vast majority of arriving immigrants today come here because they believe that government is the source of prosperity, and that's what they support.
Tokenizing	Research shows that a

	vast majority of arriving immigrants today come here because they believe that government is the source of prosperity , and thats what they support.
Case folding and normalization	research shows that a vast majority

	of arriving immigrants today come here because they believe that government is the source of prosperity, and thats what they support.
Filtering	research shows that a vast majority of arriving immigrants

	today come here because they believe that government is the source of prosperity and thats what they support
Stopword removing	research shows vast majority arriving immigrants today come here because government source prosperity

	thats what they support
Stemming	research show vast majority arriving immigrant today come here because government source prosperity that what they support

3.4. Corpus Representation of text data

Corpus is basically a demonstration of the pool of texts, characteristically labelled with text interpretations labelled corpus.

3.3.1. TFIDF Matrix

Since we cannot work with text directly, we need to manipulate the text into numeric form for which we have TFIDF [20] methods but for simplicity of computation, corpus need to be converted into the matrix form. TFIDF is the fastest way to convert a document into matrix. Whole process of Term Frequency Inverse Document Frequency is described in chapter 2 in detail. But corpus matrix don't keep the semantic relationship between words, it only meant the highest frequency word in the corpus matrix it's like one hot encoding.

CHAPTER 4
PROPOSED WORK

In this work to design a model for hoax content classification, we have performed data preprocessing using techniques like tokenization, stemming, filtration etc. After preprocessing next task was feature extraction and selection methods like simple bag-of-words and n-grams using term frequency like tf-tdf weighting, to build the classifiers based on five different classification Algorithms. For building block of text classification we need to use prior classification models and need to validate which classification methods has outperformed the rest of the methods to predict the percentage of hoax content in any entered statement. Individual classifier models are fed with extracted features.

4.1 Models

4.1.1 Naïve Bayes

The Naïve Bayes algorithm is developed and named after the mathematician Thomas Bayes, it's a very famous method and widely used in text classification which has a simple chance. This algorithm uses prior experience to predict future probability and statistics. It is not the only algorithm, but a family of algorithms where everyone shares a mutual norm, that is, each pair of features being classified is independent of each other.

The fundamental Naive Bayes assumption is that each feature makes an:

i) independent

ii) equal

Contribution to the outcome.

The equation of the Naïve Bayes theorem can be written as follows (Mitchell, 1997) [11]:

$$P(X/Y) = \frac{P(Y/X) \times P(X)}{P(Y)} \qquad (4.1.1)$$

Where:

P(X|Y) = Probability of X based on the conditions of Y.

P(Y|X) = Probability of Y based on the hypothesis of X.

P(X) = Probability of X.

P(Y) = Probability of Y.

4.1.2 Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm [12] which is suitable for classification and as well as regression. Nevertheless, It is mostly used to solve classification problems. In SVM what we do we analyze and plot data in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate. Then to perform classification we need to search for a hyper-plane which can equally differentiate two classes, either class 0 or class 1 and the distance between hyperplane an successive classes should be uniform that is the best classification hyperplane solution we would prefer to achieve that.

The training data contains forecaster variables and observed response values. In two-dimensions we can visualize hyperplane as a line and let's assume that all of our input data points can be entirely separated by this line. For example:

$$\boldsymbol{B_0 + (B_1 * X_1) + (B_2 * X_2) = 0} \tag{4.1.2}$$

Where the coefficients (B_1 and B_2) that determine the slope of the line and the intercept (B_0) are found by the learning algorithm, and X_1 and X_2 are the two input variables.

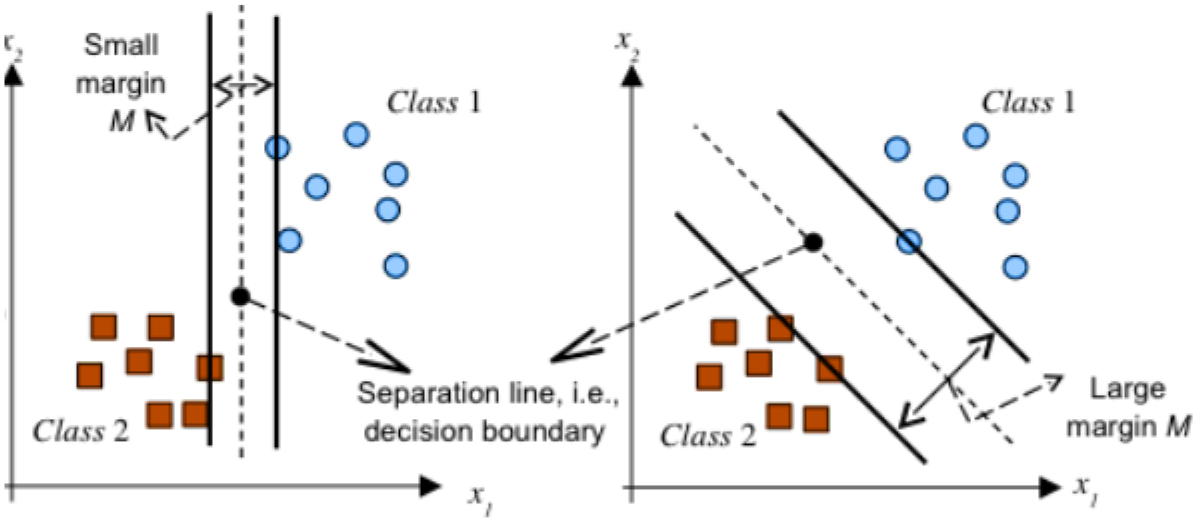


Figure 4.1.2: SVM Hyperplane

4.1.3 Random Forest (RF)

Random forest is "Ensemble Learning" in which taking Decision Tree algorithm multiple times [13]. RF is called collaborative learning approach because it uses multiple decision trees known as Forest and takes a decision based on majority. Decision tree regression algorithm splits the node on the basis of the intensity level. The general idea of the bagging method is that a combination of learning models increases the overall result.

$$Info(D) = \sum_{i=1}^{c} -p_i \, log_2 \, p_i \qquad (4.1.3)$$

Entropy is the measure of homogeneity in the data. Pi is the probability of arbitrary tuple in D belongs to ClassC_i.

$$Gain(A) = Info(D) - \sum_{j=0}^{v} \frac{|D_i|}{|D|}| * \; Info(D) \qquad (4.1.4)$$

Information Gain measure the reduction in entropy by classifying the data on a particular attribute. Classify the tuple from D based on the partition by A.

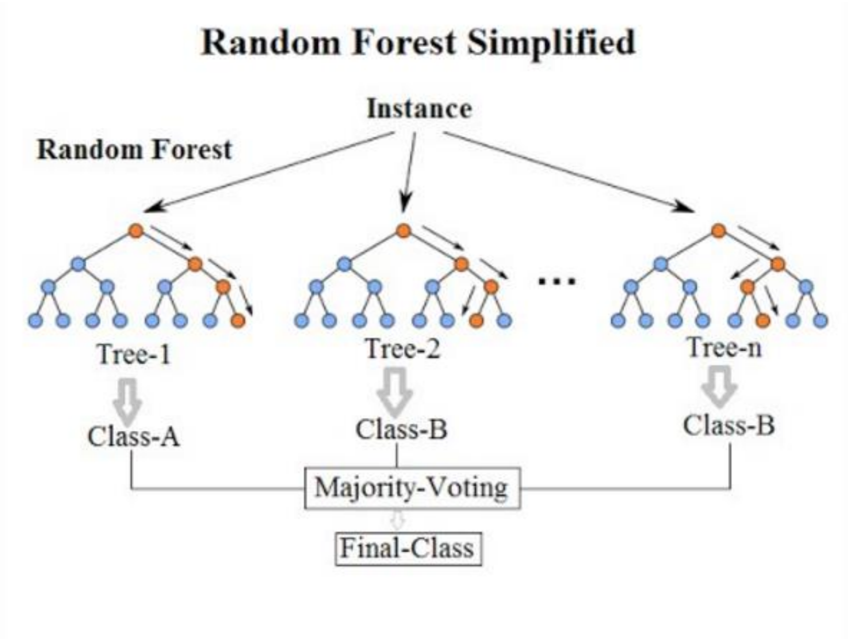


Fig4.1.3: Random Forest

4.1.4 Stochastic gradient decent (SGD)

To understand stochastic gradient decent, firstly we need to understand what gradient decent is. Gradient decent is basically the slope of the function. It is a optimization technique and used with most of the algorithms. The more steeper the slope if gradient is more. Gradient Descent can be described as an iterative method which is used to find the optimum values of the parameters of a function that minimizes the cost function as much as possible using calculus.

In SGD the word stochastic means the system or the process associated with random probability. Since it is an expensive process only a randomly selected batch from dataset is taken for each iteration to calculate gradient till minima is reached. SGD representation is:

For i in range (m):

$$\theta_j = \theta_j - \alpha(\widehat{y^i} - y^i)x_j^i \qquad (4.1.4)$$

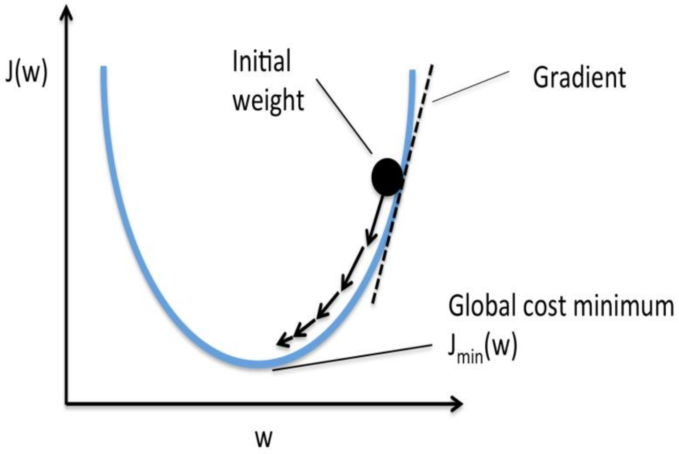


Fig 4.14 : Stochastic gradient decent minima curve

4.1.5 Logistic Regression

Logistic regression models the probabilities for classification problems with two possible outcomes. It’s an addition to the linear regression model for classification problems. It involves dependent variable which is used to find probability of success or failure of an event. LR doesn’t attempt to fit a straight line or hyperplane, the logistic regression model uses the logistic function (sigmoid function) to squeeze the output of a linear equation between 0 and 1.

Mathematically it can be represented as:

$$\text{Logistic }(\eta) = \frac{1}{1+\exp(-\eta)} \qquad (4.1.5)$$

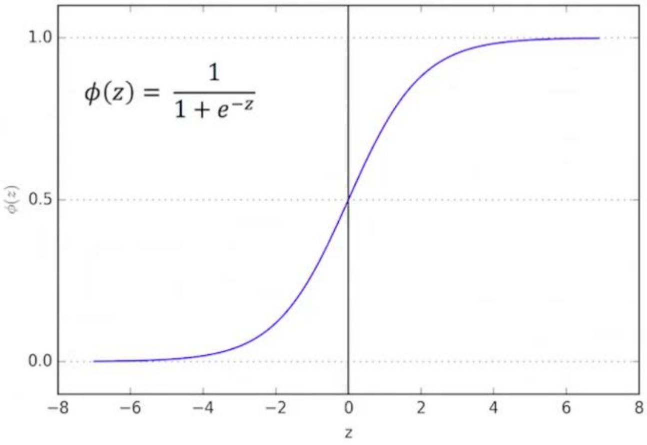


Fig 4.1.5: Logistic Regression curve

4.2 Training model

Next step is to train the model of classification. Train the model by different classification models.

4.3 Model Prediction

Deduce,

y_predict = classification_technique(X_test)

Chapter 5

EXPERIMENTAL RESULTS

The following system configuration has been required while conducting the experiments:

- Processor: Intel Core i5
- Main Memory: 8 GB
- Hard Disk Capacity: 1 TB (for faster processing)
- Software Used: Spyder 3.6 and Anaconda.

5.1 Data Representation

Data consist of 10241 training news statements, 2552 testing news statements and 2571 validatio n news statements which are categorized with 6 different labels indicating value from truth to fal se categories. For the below given graph representing data in the form of histogram with differen t categories of it.

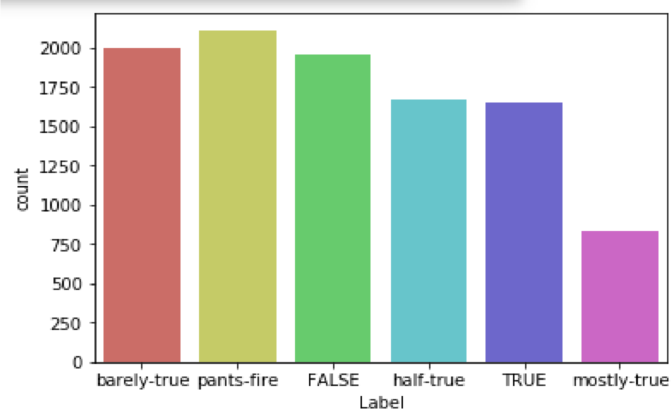


Fig 5.1: Data Representation

5.2 Model Evaluation

For prediction evaluation we are using the following measurements from confusion matrices:

5.2.1 Precision

Precision value is the level of measure of precision between what information system is generating as an output and what user actually demanded from system. Calculation of precision can be mathematically written as follows:-

$$Precision = \frac{TP}{(TP+FP)} \qquad (5.1.1)$$

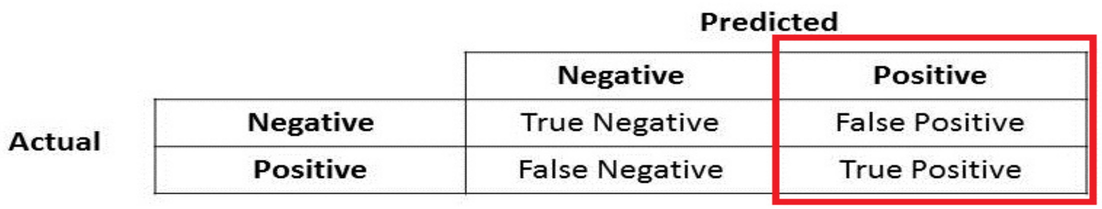
Equation (11) can be written as follows:

$$Precision = \frac{Relevent\ data\ found}{All\ data\ found} \qquad (5.1.2)$$

Where: *TP* = True Positive is the number of relevant data that is correctly classified as matches data by the system.

FP = False Positive is the number of irrelevant data, but classified as matches data by the system.

Also how to estimate precision from confusion matrix,



		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Algorithm	Precision
Naïve Bayes	0.406360424
Logistic Regression	0.3753688172
Stochastic gradient decent	0.4360189573
Linear SVM	0.3591885442
Random Forest	0.3760479042

Fig 5.1.1: Precision values of different classifiers

5.2.2 Recall

The rate of achievement of the system to rediscover information is called recall. Therefore, Recall really calculates how many of the real positive aspects that our model captured through labeling it as positive.

Calculation of recall values can be written in form:

$$Recall = \frac{TP}{(TP+FN)} \qquad (5.2.1)$$

Equation (13) can be written as follows:

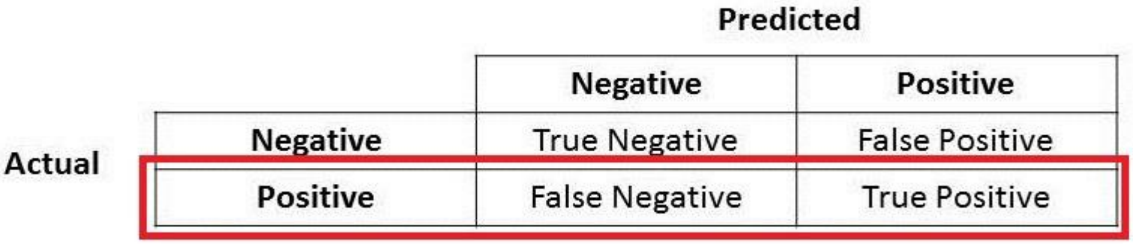
$$Recall = \frac{Relevent\ data\ found}{All\ Relevant\ data\ in\ database} \qquad (5.2.2)$$

Where: *TP* = True Positive is the number of relevant data that is correctly classified as matches data by the system.

FN = False Negative is the number of relevant data, but isn't classified as matches data by the system.

We have acquired different recall values for each algorithm used, which is depicted in the table below:

Also how to estimate Recall from confusion matrix



		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Algorithm	**Recall**
Naïve Bayes	0.9411764706
Logistic Regression	0.8137254902
Stochastic gradient decent	0.9972067039
Linear SVM	0.7521008403
Random Forest	0.7296918768

Fig 5.2.1 : Recall values of different classifiers

5.2.3 F-1 Score

The F-1 score is a performance measurement value made to see the results obtained from the classification process based on the accuracy and recall values that have been obtained. In other words, F-1 Score is the balance between precision and recall value when there is uneven distribution. Calculations for F-1 scores can also be written as follows:

$$F1\text{ Score} = \frac{2\times Precision\times Recall}{(Recall+Precision)} \qquad (5.3.1)$$

We have acquired different F-1 Score values for each algorithm used, which is depicted in the table below:

Algorithm	**F1-Score**
Naïve Bayes	0.7068126520681265
Logistic Regression	0.7280606717226435
Stochastic gradient decent	0.7212121212121213
Linear SVM	0.6920103092783506
Random Forest	0.6726920593931568

Table 5.3.1 : F-1 Score of different classifiers

5.2.4 Accuracy

Level of closeness between predictive values and real values is the Level accuracy. Accuracy can be calculated and written as follow:

$$Accuracy = \frac{TP+TN}{(TP+FN+FP+FN)} \qquad (5.4.1)$$

Calculation of accuracy in equation (5.4.1) can also be written as follows:

$$Accuracy = \frac{\text{Data that is correctly classified}}{\text{Total data tested}} \qquad (5.4.2)$$

Where: *TP* = True positive is the number of relevant data that the system correctly classifies as match data.

TN = True Negative is the number of irrelevant data that the system has correctly classified as unmatched data.

FP = False Positive is the number of irrelevant data, that the system has classified as matches data by the system.

Algorithm	**Accuracy**
Naïve Bayes	0.6034755134
Logistic Regression	0.6232227488
Stochastic gradient decent	0.5639810427
Linear SVM	0.6122432859
Random Forest	0.6058451217

Table 5.4.1 : comparative accuracy analysis of different classifiers on testing data

Comparison among precision, recall, F-1 score and accuracy values of different classifier

Algorithm	**Precision**	**Recall**	**F-1 Score**	**Accuracy**
Naïve Bayes	0.406360424	0.9411764706	0.7068126520681265	0.6034755134
Logistic Regression	0.3753688172	0.8137254902	0.7280606717226435	0.6232227488
Stochastic gradient decent	0.4360189573	0.9972067039	0.7212121212121213	0.5639810427
Linear SVM	0.3591885442	0.7521008403	0.6920103092783506	0.6122432859
Random Forest	0.3760479042	0.7296918768	0.6726920593931568	0.6058451217

Table 5.4.2: comparison Table

Chapter 5
CONCLUSION

In the work presented**,** Efficient Machine learning techniques for hoax content classification has been proposed with their performance analysis on the world wide based benchmark news data. We have built these namely classifiers Naive-Bayes, Logistic Regression, Linear SVM, Stochastic gradient decent and Random forest classifiers for predicting the hoax news detection. All the dataset features are fed into different classifiers. And each and every feature is used in all the classifier. Total 10241 news statements were used to train each classifier. Each statement has associated with 6 parameters which measures possibility score to being called either call it as mostly-true, barely-true, fire-pants, true and false. This research concludes that the best classification results are achieved by Logistic Regression if compared with Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient decent and Random forest, and the highest accuracy obtained by logistic regression of 62.322 % and highest F-1 score is 72.806% and highest accuracy indicates higher precision to predict the hoax content in news.

REFERENCES

[1] Juditha. C., (2018): Interaksi Simbolik dalam Komunitas Virtual Anti Hoak suntuk Mengurangi Penyebaran Hoaks, Jakarta: Jurnal PIKOM, vol. 19, no. 1, Kementerian Komunikasi dan Informatika RI.

[2] Monohevita. L., (2017): Stop Menyebarkan *Hoax*, Depok: Universitas Indonesia.

[3] Vuković M., Pripužić K., Belani H. (2009) An Intelligent Automatic Hoax Detection System. In: Velásquez J.D., Ríos S.A., Howlett R.J., Jain L.C. (eds) Knowledge-Based and Intelligent Information and Engineering Systems. KES 2009.

[4] Dahlan. M. A., (2017): Ahli: "*Hoax*" MerupakanKabar yang Direncanakan, Jakarta: ANTARA News.

[5]E. C. Tandoc Jr, Z. W. Lim, and R. Ling, "Defining fake news a typology of scholarly definitions," Digital Journalism, pp. 1–17, 2017.

[6]L. Zheng and C. W. Tan, "A probabilistic characterization of the rumor graph boundary in rumor source detection," pp. 765–769, IEEE, July 2015.

[7] S. Ahmed, R. Monzur, and R. Palit, "Development of a Rumor and Spam Reporting and Removal Tool for Social Media," pp. 157–163, IEEE, Dec. 2016.

[8] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "Study of hoax news detection using nave bayes classifier in Indonesian language," pp. 73–78, IEEE, Oct. 2017.

[9] Dr.S.Vijayarani et al , International Journal of Computer Science & Communication Networks," Preprocessing Techniques for Text Mining - An Overview "Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Vol 5(1),7-16 9 ISSN:2249-578.

[10] R. Tokens, "Tokenization Product Security Guidelines– Irreversible and Reversible Tokens," no. April, pp.1–84, 2015.

[11] Mitchell. T. M., (1997): *Machine Learning*, *Singapore*: *McGraw-Hill*.

[12] oachims T. (1998) Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg

[13] K. Torizuka, H. Oi, F. Saitoh and S. Ishizu, "Research of Text Categorization Model based on Random Forests," *IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pp. 487 - 491, 2018.

[14] T. Y. Aizhang Guo, "Research and improvement of feature words weight based on TFIDF algorithm," *IEEE Information Technology, Networking, Electronic and Automation Control Conference,* pp. Pages: 415 - 419, 2016.

[15] F. E. A. W. SALTON G, "Extended Boolean information retrieval," *Communications of the ACM*, vol. 26 (11), pp. 1022 - 106, 1983.

[16] J. K. S., “A statistical interpretation of term specificity and itsap-plication in retrieval,” *Journal of Documentation*, pp. 11-21, 1972.

[19] William Yang Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, to appear in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), short paper, Vancouver, BC, Canada, July 30-August 4, ACL.

[20] Z. X. CongyingShi, “Review TFIDF algorithm,” *Computer Applications*, pp. 167 - 170, 2009.