An Ensemble Based Classification Approach for Credibility Analysis of Online News by Detecting Clickbait News Headlines

THESIS

Submitted to the

G. B. Pant University of Agriculture and Technology
Pantnagar-263145, (U. S. Nagar), Uttarakhand, INDIA

By

PARUL AGARWAL
B. Tech.
(Computer Science & Engineering)

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Technology

(COMPUTER ENGINEERING)

August, 2017

ACKNOWLEDGEMENT

Pursing M. Tech research is just like climbing a high peak, step by step, accompanied with hardships, encouragement, and trust and with so many people's kind help. First of all I bow my head before 'God' who inspired me to face challenges of uneven times. The authoress expresses her deep sense of reverence and heartfelt gratitude to ***Prof. S. D. Samantaray****, Professor, Department of Computer Engineering, and Chairman of Advisory Committee for his invaluable guidance, and constructive suggestions throughout the investigation.*

With profound sense of gratitude the authoress expresses her warmest thanks to the members of the Advisory Committee, ***Prof. B. K, Singh****, Associate Professor, Department of Computer Engineering, and* ***Dr. Sanjay Mathur****, Professor, Department of Electronics & Communication, for their inspiring and constructive suggestions at every stage of this study.*

It is privilege to express my heartiest regards & sincere thanks to the entire faculty members ***Prof. P.K, Mishra, Dr. Rajeev Singh, Prof. Chetan Negi, Prof. Sunita Jalal,*** *and* ***Prof. Jalaj Sharma****, for their cooperation and support throughout the degree programme.*

The authoress tenders her sincere thanks to ***Dr. Devendra Kumar****, Dean, College of Post Graduate Studies,* ***Dr. H. C. Sharma****, Dean, College of Technology, and and* ***Dr. S.P. Singh,*** *Dean, Student Welfare, for their keen interest in providing the necessary facilities.*

My vocabulary fails to accentuate my profound reverence and sincere regards to my parents, ***Mr. Mahesh Agarwal*** *and* ***Mrs. Rachna Agarwal****. And also thanks to my all family members, for their generosity, everlasting inspiration, abundant love, encouragement, and sacrifices, and also for guiding me to achieve success at every step in life.*

Appreciations are also extended to my friends Manisha, Tejasvee, Priyanka, Medha, Simali, Deepak, Danish, Himanshu, and Rishab for their encouragement and helping hands at various stages of the work.

Pantnagar

August, 2017

(Parul Agarwal)

Authoress

CERTIFICATE - I

This is to certify that the thesis entitled **"An Ensemble Based Classification Approach for Credibility Analysis of Online News by Detecting Clickbait News Headlines"** submitted in partial fulfilment of the requirements for the degree of **Master of Technology** in Computer Engineering with major in **Computer Engineering** of the College of Post-Graduate Studies, G. B. Pant University of Agriculture and Technology, Pantnagar, is a record of bona fide research carried out by **Ms. Parul Agarwal,** Id. No. **49406** under my supervision and no part of the thesis has been submitted for any other degree or diploma.

The assistance and help received during the course of this investigation and source of literature have been duly acknowledged.

Pantnagar
August, 2017

(S.D. Samantaray)
Chairman
Advisory Committee

CERTIFICATE - II

We, the undersigned, members of the Advisory Committee of Ms**. Parul Agarwal,** Id. No. **49406**, a candidate for the degree of Master of Technology in Computer Engineering with major in **Computer Engineering**, agree that the thesis entitled **"An Ensemble Based Classification Approach for Credibility Analysis of Online News by Detecting Clickbait News Headlines"** may be submitted in partial fulfilment of the requirements for the degree.

(S.D. Samantaray)
Chairman
Advisory Committee

(B. K. Singh)
Member

(Sanjay Mathur)
Member

TABLE OF CONTENTS

3. MATERIALS AND METHODS

- 3.1. Materials
  - 3.1.1. Hardware used
  - 3.1.2. Software used
    - 3.1.2.1. Python 2.7
    - 3.1.2.2. PyCharm community edition
  - 3.1.3. Dataset used
- 3.2. Techniques used
  - 3.2.1. Natural Language Processing (NLP)
  - 3.2.2. Proposed ensemble classifier
    - 3.2.2.1. Why ensemble classification increase accuracy
    - 3.2.2.2. Selecting ensemble size
- 3.3. Proposed Methodology
- 3.4. Algorithm for Proposed System
- 3.5. Evaluation Metrics
  - 3.5.1. Accuracy score
  - 3.5.2. Confusion matrix
  - 3.5.3. K-fold cross validation
  - 3.5.4. Precision, recall and f1 score
  - 3.5.5. ROC curve and ROC-AUC score

4. RESULTS AND DISCUSSION

- 4.1. Experimental Setup
  - 4.1.1. Dataset and its description
  - 4.1.2. Feature set used for ensemble based classification
- 4.2. Experimental Results
  - 4.2.1. Command line interface
  - 4.2.2. GUI for the proposed system
- 4.3. Evaluation
  - 4.3.1. Accuracy
  - 4.3.2. Classification report containing precision, recall and f1 score
  - 4.3.3. Confusion matrices
  - 4.3.4. ROC curves and ROC-AUC

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

NLP	Natural Language Processing
Bagging	Bootstrap Aggregation
WWW	World Wide Web
NLTK	Natural Language Processing Tool Kit
AdaBoost	Adaptive Boosting
DT	Decision Tree Classifier
SVM	Support Vector Machine
POS	Part-Of-Speech
KNN	K - Nearest Neighbour
PAC	Probably Approximately Correct
ADDEMUP	Accurate and Diverse Ensemble-Maker giving United Predictions
ROC	Receiver Operating Characteristics
AUC	Area Under Curve
RF	Random Forest Classifier
NB	Naïve Bayes Classifier
RBF	Radial Basis Function
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GRU	Gated Recurrent Unit
NN	Neural Network Classifier
LTS	Long Term Support
IDE	Integrated Development Environment
CSV	Comma Separated Value
TPR	True Positive Rate
FPR	False Positive Rate

LIST OF SYMBOLS

Symbol	Meaning
$T(x)$	Function of Base classifier
$\in$	Element of a particular set
$\sum$	Summation/ sigma
$\int$	Integration
$\approx$	Approximately equal
y_{true}	true labels
y_{pred}	predicted labels
f_1	f1 score
w_i	weight of the i^{th} classifier
C_0	Constant
$I(p)$	Integrand
P_m	m^{th} point

Introduction

Chapter 1 INTRODUCTION

There are a number of pillars of the society and journalism is one of them. The role of journalism is to provide the news to the world without being emotional, biased, influential and opinionated. Many times, Journalism and democracy are paired together which reflects the power of journalism in a society. The quality of journalism can model the structure of the people, it can change the people's way of thinking toward anything. A journalist can portray a bad news as the good one and vice versa. A fake, misleading or false news may have many wrong effects on the society. For example, in a democracy, it can change the way of thinking of the voters who are going to elect their leader. In an article in Forbes "Can 'Fake News' Impact the Stock Market?" writer discussed the impact of fake news in the stock market. Many times some news headlines are just manufactured under a propaganda. So, not only the news but credibility of the news is also a factor to consider. The information which has been given by any news article is credible or misleading should also be a matter of consideration.

In this digital century, journalism experiencing a big change of digitalization. Presently, most of the media houses have become online. There are a number of mediums through which online media can reach to its readers. These mediums are Facebook/Twitter pages, websites and Android/IOS applications.

This modern online media has many differences from the traditional offline media. Media houses in old days have newspapers and magazines with subscriptions charges. They are static, present news on daily basis. Subscriptions charges are the main medium of their revenue. But the modern online media don't have this concept of subscription changes till now, they earn only from the advertisement on their page. Being on the Internet automatically refers to "It's free Phycology" of the human being, so it's very difficult to erase this from the human's mind. And that's the reason why don't online media start any type of subscription charges.

There is one more challenge, which online media houses face, is the level of competition. In the old times, people have given with very fewer choices of newspapers. So it was not difficult to survive because every media houses have a section of the audience as their readers. But today, we have hundreds of media houses most of them are online too.

Each media house has to compete with hundreds of others in order to survive in the market and in order to gain the attention of millions of users. This thrust of revenue and competition gave the birth to the concept of sensationalism in journalism or in other words **"Yellow Journalism"**. Although sensationalism is not a new concept in journalism but very common nowadays especially in online news.

1.1 Why Headlines Matters?

The main thing which heavily impacts yellow journalism is the Headlines. Headlines are the entry point of any news or article. As said that the first impression is the last impression, these headlines are for the first impression on the reader, so that they will read the article. So, online media houses create the eye-catching headlines, which immediately attract the readers. A headline determines the number of readers which are going to read the full news article and also influences the readers. And hence, headlines play a vital role in journalism and to handle the reader's psychology especially in this digital and social media era of journalism.

Research published in the Journal of Experimental Psychology confirms that misleading headlines affect reader's memory, their inferential reasoning, and behavioural intentions, even when the article itself corrects the misimpression the headline gives.

Take an example of how headline affects the full article. There was a visit of Indian Prime Minister to Germany, now think about these 2 headlines, first is "PM Modi visited Berlin and signed many MoU" and the second one is "Narendra Modi in Germany: 'Made For Each Other', Says PM to German Chancellor". It is understood that the second headline is exaggerated which will catch readers. But this is giving incomplete information and incomplete information is also misleading. This is how a headline can twist a news completely into another direction.

1.2 What is Clickbait?

The eye-catching, sensationalist headlines which are designed to lure readers to read or click on them and to encourage forwarding of the material over online social networks are known as **"Clickbaits"**. These headlines most of the time take the readers to some fraudulent links. The main reason for increasing clickbaits since these clicks will convert

into money. Some examples of clickbait headlines are “5 habits which will make you richer than anyone else in this world”, “Read to know what you will do in future based on your height and body shape”, and “This lady is thrown out of her house, what happened next will shock you”.

The main aspect which these clickbaits use is the curiosity gap. The headline generates curiosity by providing just enough information, but not enough to satisfy their curiosity without clicking through to the linked content. Readers will click and the article disappoints them because clickbait doesn’t contain the actual information and left users in between.

1.2.1 Impact of clickbait news:

1. False or misleading news may affect the society in many ways. Many times it creates havoc which may result in false information rumoured around a section of people.
2. Sole purpose of Clickbait news is only to make money which is to kill the spirit of journalism.
3. Clickbaits uses the curiosity gap of the users to generate the enough curiosity so that user will click on the link and mostly don’t live up to the expectation of reader and leave them in between (**Loewenstein, 1994)**
4. Many times Clickbait links tend to some fraudulent link which may ask credit or debit card details or other private information which may be harmful for the reader.
5. Some countries, such as India, Pakistan, there is less Digital literacy. In this case, people may easily got trapped in clickbait links.
6. News are the soft target for manufacturing clickbaits because most of the Internet users are interested in news more than any another domain in the Internet.

1.3 Natural Language Processing

Headlines are a piece of text. So, to process and analyze them, we need a method so that these headlines can be analyzed in order to detect the clickbaits. We have many technologies in computer science field and for processing the text, computer science provided with a powerful tool to handle the text written in a language i.e. Natural Language Processing (NLP).

NLP provides a number of tools and methods in order to analyze the text. Different methods and tool provided by the NLP can be used in order to work on the clickbait news

headlines. For doing this, first, we have to examine the clickbait headlines, their basic nature and then work into its analysis for produce implicit results.

NLP can be used to analysis of the text in syntactic, semantic, and pragmatic level. NLP provides tools for tokenizing by words or sentences, part-of-speech tagging of words, n-gram analysis, Entity named recognition, stop word elimination, stemming, chunking, chinking and lemmatizing. These tasks are very useful for any text classification task. NLP process the language, which may be in written (text) or spoken form, through different stages.

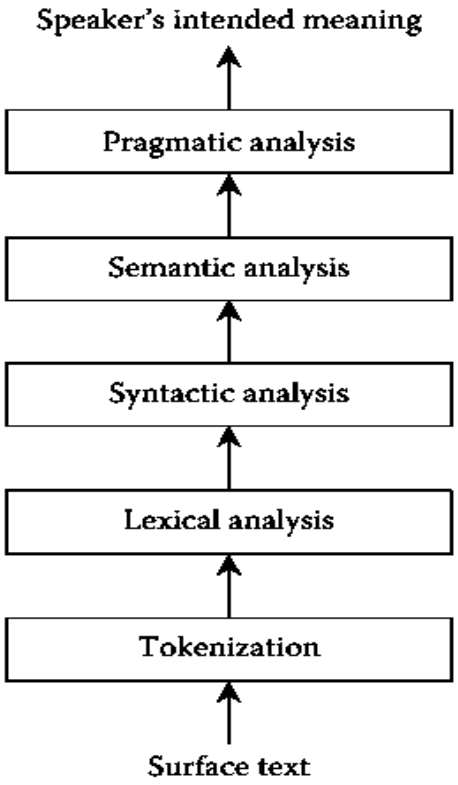


Figure 1.1: The stages of analysis in processing natural language

The text is first analyzed in terms of the syntax. Then it is analyzed in terms of semantic of the sentence, it depends more on the meaning of the sentence than the words itself. Next stage to analyze the meaning of the utterance or text in context is determined which is called pragmatic level.

Analysis of text at syntactic, semantic, and semantic is performed by a number of sub tasks like tokenizing into word and sentences, recognizing stop words, part of speech tagging, named entity recognition, stemming of words, and morphological analysis.

1.4 Ensemble Based Classification

An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. Ensemble based methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. An ensemble classifier consists of a number of base classifiers and makes the prediction by combining the results of individual predictions. Base classifier may be same or different according to the nature of the problem and method of making ensemble. It decreases the skewing nature of individual model by collecting multiple model. It also decreases variance which will make classifier less dependent on the single training set and hence produce better results.

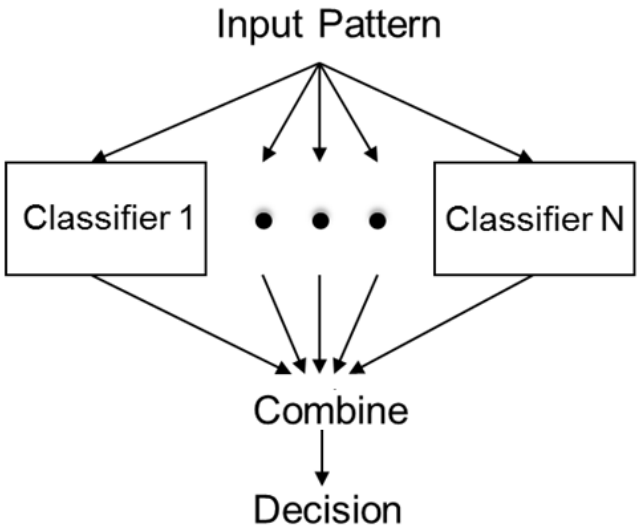


Figure 1.2: Ensemble based classification

If a classification problem have multiple weak classifiers, it would be a wise decision to make ensemble of them. In most cases, the ensemble of weak classifiers increases the accuracy. But, it may create problem when strong classifiers are collected to create ensemble. It may create the problem of overfitting which is undesirable in classification.

Ensemble classification produce better results for complex boundaries problem. Ensemble based methods are used for the several reasons like large dataset, too small dataset, generalization and complex class boundaries. Figure 1.3 demonstrated a complex boundary problem which cannot be handled by a single classifier. Due to the irregular boundaries, a classifier can work well on a particular region but not as a whole but ensemble of those classifiers can work well which is shown in Figure 1.4.

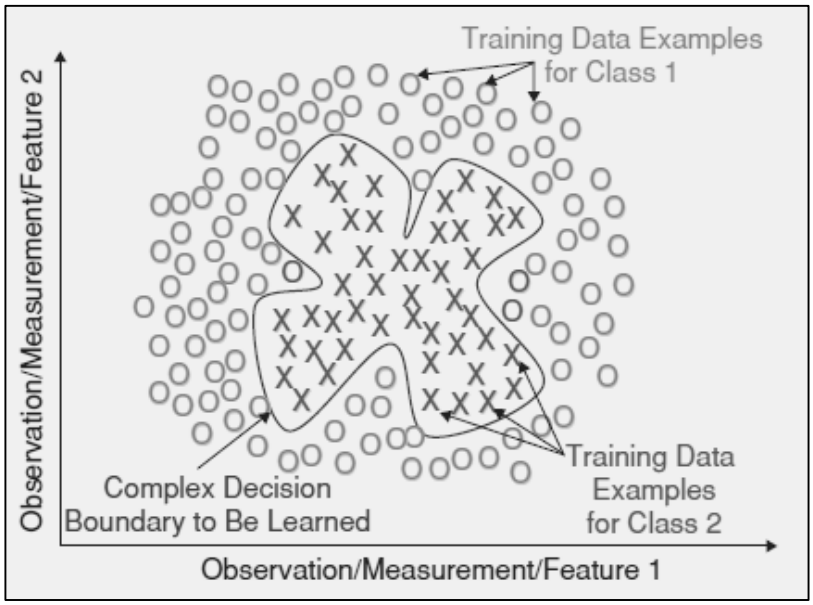


Figure 1.3: Classification problem with complex decision boundary

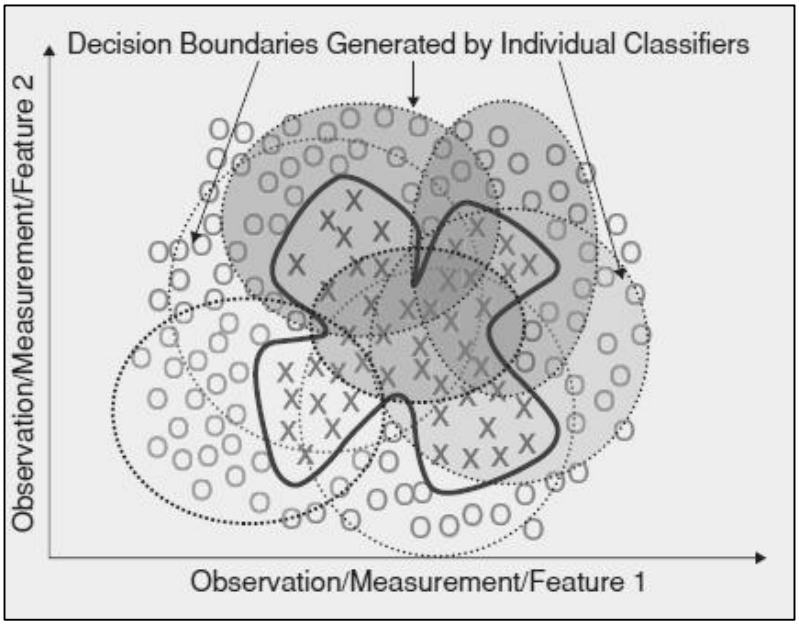


Figure 1.4: Ensemble based classification spanning complex decision space

Diversity among the classifiers and ensemble size are two main parameters which help to make a good ensemble classifier. If the classifiers in ensemble make errors on different regions of the input space, they will make more accurate ensemble classifier than any other individual classifier. Diversity in classifiers having errors on different data instances then they can make a good ensemble and increase the performance than that of individual classifiers. Classifiers selection for ensemble is also equally important because the diversity

dependent on it. Diversity can be achieved through training of classifiers with different training dataset which can be done with sampling.

Ensemble based classification can be of many types, these types differ in base estimator or classifier and the decision method which is used for combine the predictions from multiple estimators. Common types of ensemble based classification are as follows:

1.4.1 Bagging

Bagging is short form of Bootstrap Aggregating. In this method, multiple classifiers of same type are trained using different training data subsets which are randomly drawn from the entire training data with replacement. The final prediction is done by aggregating the decisions of individual classifiers. Bagging method is good even when dataset size is limited. Base classifier should be unstable and weak to maintain diversity which results is better performance.

1.4.2 Boosting

Boosting is similar of Bagging in terms of resampling of training data and same type of classifiers are used in both methods. In this method, base classifiers are built sequentially, next classifier tries to reduce the bias of previous one. First classifier is trained using random subset of the training dataset. Second classifier is trained on training data subset in which only half data samples are correctly classified by the previous classifier. This process continues till the ensemble size is reached. Results of all individual classifiers are combined with majority voting and a strong classifier is built from multiple weak classifiers.

1.4.3 Mixture of experts

Each conceptually different classifier is trained using same training dataset and then the final decision is taken by averaging their predicted probabilities or majority voting. This method is well performing only when there are many different classifiers which are all well performing. In this method, diversity is maintained due to the different types of classifier unlike bagging and boosting which differs in training samples.

1.4.3.1 Voted classifier with hard voting

In majority or hard voting, the final decision for a particular problem is the decision that represents the majority of the decision predicted by each individual classifier.

This method used majority voting for final prediction. The predication which is supported by majority of the classifiers is selected as the final decision.

1.4.3.2 Voted classifier with soft voting

In the soft voting method, unlike hard voting, specific weights are assigned to each classifier with the help of the weights parameter. When weights are provided, the predicted class probabilities for each classifier are collected, multiplied by the classifier weight, and averaged. The final prediction is then derived from the decision with the highest average probability.

The algorithms, which are mentioned, are widely used for the ensemble learning, one can select the algorithm according to the nature of the problem and a number of classifiers. Combining decision can also be done with different polling methods like majority win, weighted average, winner takes all, and stacked voting. Many problems where majority win method produce better result, in other problems it may fail bluntly.

1.5 Motivation

The number of people using the Internet to find and read news online is consistently on the rise. Some studies and surveys have shown that there is a drastic increase in the online news readers. Online news is becoming a powerful and popular tool for the journalism. Billions of people access news online on the hourly and daily basis. A news headline becomes viral over the Internet in a period of minutes and hours. A misleading news can affect the readers, even this effect may be very little.

Clickbaits are killing the noble purpose of journalism to make aware people of the activities happening all over the world, aware, and educate them. The number of clickbait headlines is also increasing drastically day by day as the readers are increasing. All of us encounter with a lot of clickbait useless headlines on every other day. Even, reputed online media houses are also taking help of clickbait headlines to increase the readers count. Many attempts have been taken to tackle this problem, but clickbait headline creators are using new ways to trap the readers. Our work will try to build a dynamic system with the use of Ensemble learning which can address the challenges in detecting these false, misleading, and sensationalist headlines.

1.6 Problem Statement

Clickbaits are affecting the journalism, society and other areas in many aspects. Sensationalism, the thrust of reader's clicks and viral headlines give rise to the creation of clickbait news headlines. Clickbaits compromise with the quality of the news, spread misinformation, and exaggerate the news events. There should be a way to detect these misleading headlines so that filtered and quality news can be preserved, and forwarded to the readers. Detection of these type of online news articles and headlines is necessary to preserve the spirit of true journalism.

1.7 Objective

- Analyse the credibility of online news.
- Study and compare the behaviour of clickbait and non-clickbait news headlines.
- Recognize the characteristics of clickbait news headlines.
- Propose and design a system which automatically classify online news headlines into clickbait and non-clickbait categories.

1.8 Thesis Outline

Chapter 1 describes the background information, motivation, problem statement, objectives and significance of the study.

Chapter 2 looks at the various studies and related work carried out in order to get the required information about the credibility analysis of online news, clickbait detection, Text-analysis through Natural language Processing, and ensemble based classification method.

Chapter 3 elaborates various techniques, hardware and software tools that have been used in the development of the proposed work. It also provides information about the data set used for this study and the adapted model.

Chapter 4 deals with the result and discussions of the proposed study. It also elaborates classification performance estimated with the help of a number of evaluation metrics.

Chapter 5 deals with the conclusions remarks and summary of the work. The future scope is also discussed that can be used for further improvement.

Review of Literature

The relevant literature pertaining to various aspects of the present study are reviewed in order to gain a thorough understanding of the subject. The review of literature is categorised with respect to the work done in the area of

- Credibility of Online news
- Natural language processing and Text Analysis
- Ensemble based classification
- Clickbait headlines detection
- Evaluation metrics for classifier's performance evaluation

At the end of this chapter research gap will be discussed and this research gap analysis is used to recognise the flaws in previous methods. This would be helpful for the proposed methodology in which we try to overcome these flaws and shortcomings.

2.1 Literature Review on Credibility of Online News

The number of people using the Internet to find and read news online is consistently on the rise. One national study by the Pew Research Centre reported that weekly use of online news tripled from 11 million to 36 million people in the United States between 1996 and 1998. This report is enough to know about the popularity and power of online news. Since online news is accessed by billions of people on daily basis, the loss of credibility of online news may affect badly to society in any from. So, it is necessary to be concerned about the credibility of online news.

Schweiger (2000) conducted a survey for credibility analysis of media on WWW in comparison to other media. He used three survey mode for improving the result that are face to face, telephonic and through emails. Total 540 respondents are participated in survey. As a conclusion, Schweiger assessed the credibility of web, newspapers and television under 11 factors seem of them are competent, completeness, partial or not, whether researched, balance, detailed or not, and contradiction etc. And he found that web credibility is satisfactory for only 3 or 4 factors whereas television and newspapers are not good for one

or two factors. This survey was conducted for both the categories web users and web non-users, but relative credibility remained same for both cases.

Flanagin and Metzger (2000) described about the popularity of Internet for accessing news. They stated that the Internet news growth tends to increase the online fraud, misleading content over social groups and misinformation. Newspapers, television and magazines undergo under some kind of editorial process like fact verification, content rechecking, and cross verification. But news on Internet don't have such scrutiny. Taking media credibility as a measure, 1041 surveys were completed. Principal components factor analyses were performed on the nine items constituting the Internet verification scale. They calculated the relative credibility of media channels of the Internet and assessed whether credibility ratings depended on the type of information sought. In this analysis, newspapers are rated high in credibility score and Internet news are rated least.

Garrison *et al.* (2002) has done a study on components that analyses the credibility of online news. A national probability sample from the 50 states and District of Columbia s used for the study. With the use of adult's telephone survey and 536 interviews, some factors are discovered which decided the credibility of news. These factors include trustworthy, current, biased, fair, report the whole story, objective, dishonest, up-to-date, believable, balanced, accurate and timely. Standard deviation for each factors of Newspapers, television and online news is calculated and it is noticed that it is highest in online news. They have also discussed about the importance of branding in online news, which will give rise to clickbait headlines.

Ha and Ahn (2011) explained about user's tweet sharing behaviour. They have proposed a research model in which this behaviour of retweeting considered to be dependent on argument quality, source credibility, information usefulness, and self-efficiency. They concluded that source and information credibility play a vital role in tweet sharing.

Castillo *et al.* (2011) studied about the credibility of information spread through social media networks, especially information credibility of news propagated through Twitter. Firstly twitter data is collected and filtering of tweets which contain some news in it are done. Chats and personal opinions are discarded. Then credibility assessment is performed considering some factors like the reactions and emotions by user, level of certainty of users, external source cited or not, and characteristics of the user. Message based, user based, topic

based and propagation based features are used to build supervised classifier for automatically tweet credibility assessment and for finding newsworthy topics. Out of 68 features, best 15 features are selected to increase the accuracy of classifier.

According to **Gupta and Kumaraguru (2012),** all content posted on Twitter is not trustworthy or useful in providing information about any topic After analysing twitter data they have concluded that on average 30% of total tweets posted about an event contained information about the event while 14% was spam. Hence, they proposed a system to assess the tweets credibility. First, tweets about some trending topics of that time are collected, then annotation the set of tweets for each event using human annotator is done. A combination of supervised machine learning and relevance feedback approach is used to rank tweets. Features are identified using regression analysis. Two categories of features are used in this classification, one for the message itself and other for user who posted that tweet.

Chung *et al.* (2012) studied about the factors that play a role in credibility for three news source categories that are, mainstream, independent and websites. In their study, 288 participants are there, most of them are online news consumer. Credibility scale included interactivity, hypertextuality, expertise, trustworthiness, and multimediality. 7-point Likert-type scales used to measure the credibility. They found that hypertextuality of online news is influential whereas multimediality and interactivity did not affect credibility.

Kawabe *et al.* (2013) proposed a method to assess the credibility of tweets of Twitter social sites. Sentiment analysis as a base method, additionally used sentiment orientation dictionary. Four modules are used that are tweet collector, tweet opinion classifier, tweet sender/receiver, tweet credibility calculator and tweet opinion database. To assess the information credibility, majority decision is used by comparing the number of contrast opinions about a topic. Credibility is the ratio of the same opinions to all opinions about a topic. Expertise Consideration also used as *EScore*. It this way, individual expertise can also be a part of credibility assessment and as a result accuracy increased.

2.2 Literature Review on Natural Language Processing & Text Analysis

Bird & Loper (2002) described about the Natural Language Processing Toolkit (NLTK) in Python. NLTK is The Natural Language Toolkit is a suite of tutorials, program modules, exercises, and data sets covering statistical and symbolic natural language

processing. NLTK is developed in 2001 and mainly for three purposes which include demonstrations, assignments and projects. NLTK contain a number of modules like tokenizer for generate tokens, corpus module which contain thousands of datasets, processing modules which have a number of parsers and taggers. NLTK is very easy to use the only step is to install it in to the system and use inbuilt modules to perform a number of NLP tasks in Python.

According to **Sebastiani (2002),** in this era of digitalized documents, it is convenient and necessary to classify the text. He reviews a number of method for Automated Text Categorization using machine learning. In this approach classifier is automatically build by learning, rather than using some predefined domain knowledge. Text classification using machine learning is more effective and productive rather than previous knowledge engineering approaches. It can also handle very noisy text. He also compared more than 10 classifier for this task including DT, SVM, AdaBoost, and others.

Nasukawa *et al.* (2003) presented a sentiment analyzer that extracts sentiment about a subject from online text documents. Sentiments can also thought as opinion. They performed three tasks with the help of Natural Language Processing that are, sentiment extraction, relationship analysis, and feature term extraction from the text document. Digital camera and music review articles are used as dataset and mixture algorithm approach is used for feature term extraction. For sentiment and semantic relationship analysis, sentiment pattern database is used with product review dataset.

Nasukawa and Yi (2003) proposed a sentiment analysis approach to deal with opinion extraction related with polarities of positive or negative for particular subject from a document, rather than characterizing the entire document into positive or negative. According to them Sentiment analysis is not a single task whereas it involves expressions identification, polarity and strength of the expressions, and also the relationship to the subject. For POS tagging, Markov-model-based tagger is used. Window analysis is used which included at least 5 words before and 5 words after the target subject. Identification of semantic relationships helped to achieve a high precision rate.

Kouloumpis *et al.* (2010) have conducted study on sentiment analysis of twitter's tweets. In their work, they followed supervised approach and collected data sets as training data. For analysing, they have used three types of data sets, because Twitter contains non-

uniform data like texts, acronyms, emoji etc. Hence they used different data sets for training purpose. Emoticons and abbreviations (e.g., OMG, WTF, and BRB) are identified as part of the tokenization process and treated as individual tokens. Hash tagged data set, the emoticon data set and a manually annotated data set are used.

Medhat *et al.* (2010) has done a survey on algorithms of sentiment analysis and its applications. They presented sentiment analysis process on product reviews. This process have several steps; Sentiment identification, feature selection, sentiment classification. Input of the process is product reviews and output is the polarity of the review. In recent year, many work has been done on this field. Some of the features which are important in sentiment analysis are terms presence and frequency, POS, opinion word and phrases, and negations. They described about some feature selection methods like Point-wise Mutual Information, Chi-square, and Latent Semantic Indexing. Some of the algorithms used for sentiment classification are rule base, web based, lexicon based, Graph based, semi supervised classification, random walk algorithm, ranking algorithm, multi-class SVM, NLP, corpus based, Naïve Bayes, context based, KNN and decision tree classification.

Collobert *et al.* (2011) proposed a neural network learning algorithm which can perform all the tasks related to NLP. NLP tasks includes POS tagging, semantic role labelling, chunking, named entity recognition, and stemming. They explained the necessity of NLP that convert English text to programmable data structure that can be used for further computing. A window approach network is presented in the paper for different NLP tasks. Wall street journal data is used for the training for POS tagger, chunking and other tasks. POS tagger label each word with its part of speech value. Chunking chunk the section of text by words or sentences which is required. Name entity recognizer labels atomic elements in the sentence into categories. Transforming Words into Feature Vectors is one of the main work which is required for classify the text. Various approaches are described for extracting high level features from the words and sentences like window approach, sentence approach, and tagging schemes.

Gokulakrishnan *et al.* (2012) performed opinion mining in Twitter data stream. Two disjoint datasets one containing 8500 manually annotated negative tweets and other containing 41,000 positive tweets are used. They have applied sampling techniques to reduce the skewness of dataset. They have applied some pre-processing step like uppercase identification, lower casing, replacing emotions, URL identification, punctuations and

hashtag detection, removal of hashtags and query terms, and compressions of word to avoid unnecessary noisy data. After getting pre-processed data, a number of classifier are trained and as a conclusion Naive Bayes Multinomial classifier is found to have highest average accuracy with 75% in positive/negative/neutral classification of tweets.

Manning *et al.* (2014) described the design and purpose of the Stanford CoreNLP Natural Language Processing Toolkit in their paper. It is open source NLP technology which provides complex language analysis tasks. It is pipelined framework and developed in Java programming language. The initial version was developed in 2006. Some of the annotators provided with StanfordCoreNLP are tokenize, cleanxml, ssplit, truecase, ner, pos, gemma, gender, parse and sentiment which are used for different NLP tasks. It provides support to six languages, Arabic, Chinese, English, French and German. Some of the annotators are missing in languages other than English. By using this toolkit, a complex task like sentiment detection can be done easily with two or three lines of code.

Kiritchenko *et al.* (2014) presented a sentiment analysis system for message-level analysis as well as term-level analysis of Tweets and SMS. The framework depends on a supervised statistical text classification approach utilizing surface form, semantic, and sentiment features. Message level analysis detect the sentiment of whole message, whereas term level analysis the sentiment of a particular word. Sentiment lexicons are used for the proposed system. An SVM using linear kernel is trained with a set of features like word ngrams, character ngrams, POS, all-caps, Hashtag, negation, sentiment lexicons, punctuation, emotions, elongated words, stop word, length and some others. This SVM classifier with linear kernel presented in outperform maximum entropy based classifier.

2.3 Literature Review on Ensemble Based Classification

Schapire (1990) proved that a strong classifier in probably approximately correct (PAC) sense can be generated by combining weak classifiers through boosting. He discussed about the power of weak learnability which can make a base of a strong learner. He has developed an algorithm which convert weak learner to strong one. It is one of the earlier ensemble based classification algorithm. He has shown that a model of learnability in which the learner is only required to perform slightly better than guessing is as strong as a model in which the learner's error can be made arbitrarily small. He also defined a set of general

upper bounds on the complexity of any strong learning algorithm as a function of the allowed error.

Breiman (1996) showed that bagging is effective on unstable learning algorithms, such as neural networks, where small changes in the training set result in large changes in predictions. He presented the idea of Bagging predictors which is a method for generating multiple versions of a classifier and using these to get an aggregated classifier. The averaging of these multiple classifiers are done to predict the final decision. The algorithm is also tested using classification, and regression trees and found to have satisfactory of better results. Bagging is the short form of Bootstrap Aggregating. It is also one of the earlier algorithms along with Boosting for ensemble based learning.

Opitz and Shavlik (1996) proposed a technique i.e. ADDEMUP. It used the genetic algorithm concept for making accurate and diverse ensemble of neural network. Diverse classifiers make a better ensemble. This algorithm can also include the prior knowledge which can increase the efficiency of ensemble. ADDEMUP initiated with an initial network then genetic algorithm works and new networks are formed by mutation and crossover operation. Fitness function include both diversity and accuracy.

Opitz and Maclin (1999) have done a comparative analysis of performance among single classifier, bagging, arcing and AdaBoost for neural network and decision tree classifier. For evaluating performance, they used average of five 10-fold cross validation which is run for ensemble of 25 classifiers. They used 23 datasets for their experiments. Ensemble size also affect the performance. It is stated that 10 or more is a good ensemble size for desired performance. Results demonstrated that bagging outperform a single classifier and boosting outperform bagging. Due to the presence of noise, boosting may suffer from overfitting.

According to **Hu (2001),** diversion in the classifiers result in the uncorrelated classifications, and this uncorrelated nature of classifiers increase the accuracy of ensemble. He used set theory and database set operations and developed a new rough set based approach. Reducts are created which contain all possible attributes then a reduct classifier is built which only contain non redundant and important attributes. The reduct classifier contain minimal set of rules, all redundant rules are pruned. This method is scalable but not performed well for large databases.

For improving the generalization performance, **Kim *et al.* (2002)** developed a classifier ensemble of SVM using bagging which is also known as Bootstrap aggregating. Each of the individual SVM classifier has trained using independently using the randomly chosen training samples from the original training set. After this, decisions of all classifiers are combined with a number of methods like majority voting and least square estimation based weighting. This ensemble classifier has applied on IRIS standard dataset and found that ensemble outperform the single classifier by 2% approximately. They also used this classifier to recognise hand-written digits and again ensemble outperform single SVM classifier by 1% (approx.).

Polikar (2006) studied about different type of ensemble based system and their design and implementation. He explained that ensemble based methods are used for the several reasons like large dataset, too small data, generalization and complex class boundaries. He explained about the necessity of diversity in classifiers having errors on different data instances then they can make a good ensemble and increase the performance than that of individual classifier. Classifiers selection for ensemble is also equally important because the diversity dependent on it. Diversity can be achieved through training of classifiers with different training dataset which can be done with sampling. He explained the necessity of weak classifiers in bagging, boosting and ensemble classifications.

Banfield *et al.* (2007) performed comparative analysis of eight approaches for creating ensemble of decision tree classifiers over 57 datasets which are available publicly. These approaches include bagging, boosting, and some randomized algorithms. They compared bagging against a number of methods with the help of 10 fold and 5 × 2 cross validation and found that bagging is better in 37 out of 57 dataset in both cross validation. And in other datasets, boosting and random forest worked well. For both the validation technique, bagging has highest average rank and random forests algorithms have low average rank.

Rokach (2009) reviewed different methods for ensemble classification. According to him, there are four basic building blocks of the ensemble learning that are training set, base estimator, diversity generator and combiner. There are basic two methods for making ensemble bagging and boosting. AdaBoost is an algorithm for boosting method. AdaBoost is reviewed in various situations and found that it generates a classifier with less misclassification rate and low variance. But it fails in the case of overfitting. Bagging also improves the accuracy of base classifier. There are also a number of methods for combining

the base estimators like majority voting, performance weighting, distribution summation, density based weighting etc.

Hagen *et al.* (2015) applied ensemble based classification technique for twitter sentiment analysis. Three classes are considered for sentiment polarity that are positive, neutral and negative. For the classification decision, average of individual classifiers confidence scores are used. They used three classifiers of top teams in SemEval 2013-14 event. The three classifiers which have been chose are SVM classifier with linear kernel, gradient decent classifier and maximum entropy based classifier. Feature set of SVM includes N-grams, all caps, Part of speech and polarity dictionary. Feature set gradient decent classifier includes normalised unigrams, stems, clustering, polarity dictionary and negation. The last classifier has length, polarity dictionary, emotion and negation as features.

2.4 Literature Review on Detecting Clickbait Headlines

According to **Blom and Hansen (2015),** forward referencing is widely used to build the clickbait headlines, so that curiosity will generate and readers will click on the news which is the sole purpose of clickbait headlines. News headlines are not just a piece of independent text whereas it is dependent on the full news article. 10000 headlines are used for news headlines analysis using multimodal reference model. It is observed that forward reference in headlines is expressed by many types such as adverbs, personal pronouns, definite article, and pronouns. Headlines are analyzed and it is found that hard news contains only 15% whereas soft news (online news) contain up to 47% forward reference dependent on the subject; for example, lifestyle, entertainment, and gadget sections contain most of the forward referencing.

Chen *et al.* (2015) reviewed a number of issues regarding online news as the increment of Internet of news affected the way news is presented. They felt that in the era of online news the gap between user-generated content and traditional news became negligible. Online news media are suffering from the cultural shift due to the thrust of clicks. They advocated for the awareness for the clickbait and misleading content which are present on the Internet and Digital literacy. There is a need for a news credibility assessment tool for news creator as well as for the readers so that they will become aware for the clickbait headlines. And this awareness will help to protecting the journalism as well as good for the readers.

According to **Conroy *et al.* (2015),** as the news has moved online, a new form of tabloidization has emerged i.e. click baiting. Clickbait news may be the reason for misinformation and rumor. Knowledge gap and curiosity may be used as a tool by the content creator to make clickbait headlines. They examined some methods that can be used for automatic detection of misleading online content. Certain levels of news analysis for clickbait detection are lexical, semantic, syntactic and pragmatic levels of analysis, image analysis (for non-textual news) and User Behaviour Analysis.

Potthast *et al.* (2016) developed an automatic system which detects the clickbait. Twitter corpus is used of 20 high profile news publisher. Top 20 publisher is selected with the help of a number of re-tweets. A random forest based classifier is designed with 215 features. These 215 features are categorized into three categories that are, teaser message, link webpages and Meta information. 203 features are under the teaser message category, 8 features for link webpage category and 4 in Meta information category. Twitter corpus is split into 2:1 ratio for training and testing purpose. Random forest classifier is trained with the help of 215 features values of the Twitter corpus. They compared Random forest, logistic regression, and Naïve Bayes classifier. RF performed well with 0.74 ROC-AUC whereas LR has ROC-AUC 0.72 and NB has ROC-AUC 0.69.

Chakraborty *et al.* (2016) described the curiosity gap, which is used to create clickbait headlines. They have collected thousands of news headlines and 7500 clickbait and 7500 non-clickbait news headlines chose for training the classifiers. By comparing both types of headlines, 14 features are selected. With these 14 features, click baits can be detected semantically, syntactically and pragmatically. SVM classifier with RBF kernel is designed which attained high accuracy rate. A browser extension is also developed which used personalized classification to automatically block the clickbait headlines appeared on any web page. They have used three blocking approaches; pattern based, topic based and hybrid approach. Pattern based approach achieved the highest accuracy of 81% among all three approaches. The browser extension had correctly blocked on average 89% of the links which contain clickbait news.

Agrawal (2016) argued about the effect and importance of news headlines. Headlines are the entry point of a news article and affect the whole article, hence headlines are taken into consideration. He parsed different social platforms such as Facebook, Twitter, and Reddit to create clickbait and non-clickbait corpus. He proposed a convolutional neural

network (CNN) model for clickbait detection. CNN is a type of deep learning technique. First textual headlines are converted into word embedding. These word embeddings are fed to CNN model with one layer of convolution. He developed two models, Click-Scratch and Click-Word2vec. Evaluation of both models are done using 5-fold cross validation. Click-Word2vec achieved 90% accuracy whereas Click-Scratch achieved 89%.

According to **Anand *et al.* (2016)** previous methods relies on feature engineering, hence they tried to develop Recurrent Neural Network (RNN) for detection of clickbait news which don't required heavy feature engineering. Distributed word and character level word embedding are used and the proposed model contain three layers; embedding, hidden and output layer. They introduced three different variants of Bidirectional Recurrent Neural Network model for detecting clickbaits. One is standard RNN, second is RNN with Long Short Term Memory (LST) and third is RNN with Gated Recurrent Unit (GRU). After 10-fold cross validation, accuracy of BiRNN is 96%, BiGRU is 97% and accuracy of BiLSTM is 98%.

According to **Chen and Rubin (2017),** in current time, clickbait news became a tool for spreading rumors and false information. They have used Q-methodology which is designed to study the human subjectivity. 30 participants are appointed to decide about 70 headlines whether they are definitely clickbait or definitely non-clickbait or uncertain. Information from these participants subjected to relationship and elements to distinguish comparable arranging designs among participants, demonstrating shared traits and showing subjective viewpoints. As a result, it is found that headlines containing profanity, forward-referencing, and colloquial phrasing are mostly clickbait. And also more clickbait news is from the entertainment, lifestyle, and sports news category.

2.5 Literature Review on Classification Evaluation Methodology

Kohavi (1995) reviewed and compared two accuracy estimation methods: cross-validation and bootstrap. The accuracy of a classifier is the probability of correctly classifying a randomly selected previously unknown sample. There are a number of ways to evaluate the performance of a classifier. In hold out method, two mutually exclusive data subsets are used which are known as training and testing set. Most commonly training and testing set are in 2:1 ratio. A classifier is trained using training set and test on the test set.

There is also a popular method called cross-validation, which can also be used when the dataset is of limited size. In k-fold cross-validation methods, the dataset is partitioned into k subsamples. In each kth iteration, the kth subsample is used for testing and rest k-1 subsamples for training. Performance is evaluated by averaging results of each iteration. He evaluates C4.5 and Naïve Bayes classifier over many datasets. K-fold cross validation is found better than the bootstrap method with moderate k value.

Fushiki (2011) presented a way to estimate the prediction error using k-fold cross validation. K-fold cross validation method have upward bias value whereas prediction error has downward bias value. It is tried to fill this gap between these two and hundreds of experiments are done. He concluded that the value of k should be 5 or 10 in the large dataset in order to reduce the computation burden.

Yadav & Shukla (2016) compared k-fold cross-validation to hold-out validation for quality classification. They have done a comparative analysis of both methods with 20 classifiers in four different problem datasets. In their comparison, it is found that in Page Blocks Classification problem, 90% times k-fold method gave better accuracy, in Handwritten Digits Recognition, it gave 80% times better accuracy, and in Letter Recognition Dataset, it gave 77.5% times better accuracy than hold out method. Based on these experiments, value of k can also be selected with reference to the number of instances of dataset. If data instances are 5000-10000, the value of k should be 5-6, if 100000-1000000, k should be 3-5, and if it is more than 1000000, k value should be small such as 2 or 3.

2.6 Research Gap

The concept of credibility of online news is quite complex and it can be measured considering a number of dimensions. Researchers have used a range of approaches to evaluate credibility. But none of them consider headlines or clickbait-ness their main concern of research. The trend of clickbait news headlines is increasing day by day, which makes clickbait-ness also a factor to assess the credibility of online news. Most credibility studies of online news have compared the credibility of online news with traditional media outlets but ignored the factor of clickbait headlines.

Online news media is mostly textual, a number of research papers relating to text analysis are reviewed and NLP is found to be one of the best method to work with the human

language either it is in textual form or in spoken form. It provides required tools and utilities to work on language (textual or spoken). A number of tools are also developed for implementing NLP. One of the powerful tool which is used by a number of researchers is Stanford coreNLP tool which is an integrated NLP toolkit with a broad range of grammatical analysis tools, a fast and robust annotator, and quality text analytics.

Other studies which are focused on the clickbait detection, either took too many features included redundant and low informatics features or apply classification techniques which are vulnerable to complex decision boundaries. Classification of clickbait and non-clickbaits are not a linearly separable, it can also have complex decision boundaries. In many cases, the same headline which looks like a clickbait may not actually clickbait. Because headline writing is an art, every journalist wants to write an effective headline and in the process of making headline effective, it may have some features which belong to clickbait headlines. Clickbait detection is a new area of research in recent years due to the increasing trend of online news and the vast increase in number of online news readers.

To fill this gap in the literature, the current study investigated clickbait headlines as a factor contributing to the credibility of online news. And for detecting these eye-catchy but misleading headlines, ensemble classification method, which is prone to complex decision boundaries, is used so that it can distinguish between effective and clickbait headlines. Feature selection is also used to opt out irrelevant and low informatics features, and by only selecting a subset of relevant features.

Materials and Methods

This chapter elaborates various techniques and tools that have been used in the development of the proposed work. First section presents materials, the hardware and software tools that have been used for the realization of proposed work and second section focuses on the comprehensive outline of the various techniques. Section three, four and five present the block diagram, flow chart and algorithm for the proposed work respectively.

3.1 Materials

This section deals with the brief introduction of the hardware, software tools that are used and the dataset which has been used in the proposed work.

3.1.1 Hardware used

All of the proposed work mainly held around text analysis, so that only a computer system as hardware tool is needed for realization of the proposed work. The system is developed and executed on a PC which has the following hardware specifications –

- Processor: AMD A6 3420m, 1.50 GHz
- Installed memory (RAM): 4 GB
- System type: Ubuntu 16.04 LTS, 64-bit operating system
- Hard disk: 512 MB

3.1.2 Software used

All programming of the proposed work has been done in Python programming language. It is convenient to use an IDE in order to work smoothly on any project. Hence, PyCharm IDE Software is used as a platform to work on python.

3.1.2.1 Python 2.7

Python is an interpreted general purpose programming language. It is created by Guido van Rossum and first released in 1991. It is open source and has a very strong community which increase its capabilities. Python have thousands of packages for different purpose. Anyone can freely install and use these packages with a simple command. It has easy syntax

and high code readability. Python files have .py, .pyc extensions. Python 2.7.12 is used for programming.

The proposed work deals with the text headlines and then classification, Hence it require features mainly related to these two fields:

a) Natural Language Processing
b) Machine learning Techniques for classification

For both field Python has a wide variety of features. There are a number of python packages which are useful in proposed work, some of them are:

1. **NumPy,** the fundamental package for scientific computing.
2. **SciPy,** package for mathematics, science, and engineering.
3. **NLTK (Natural Language Toolkit),** toolkit for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.
4. **Scikit Learn,** It is a very important Python library which is used for implementing Machine learning approaches in Python. It is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
5. **SocketServer,** module which simplifies the task of writing network servers.
6. **Matplotlib,** Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
7. **Tkinter,** short for Tk interface. It is the standard Python interface for providing graphical support using Tk toolkit.
8. **Urllib2,** it defines functions and classes which help in opening URLs (mostly HTTP) such as basic and digest authentication, redirections, cookies and more.

Basically Python works on command line but it can also be used with IDE for easy working. Python provides a default IDE for windows Operating System. Using an IDE would be a better decision for smooth environment. All features can be exploited at a single window with ease of programming.

3.1.2.2 PyCharm community edition

PyCharm Community is Lightweight open source IDE for Python & Scientific development. It is developed by Czech based JetBrains company.

It provides a user friendly graphical user interface to running, debugging python projects. It also provide code assistance, functionality for easy error detection for programmer, description of functions. It makes python programming easy. First version was released in July 2010. PyCharm Community Edition 2017.1.3 is used for the proposed work.

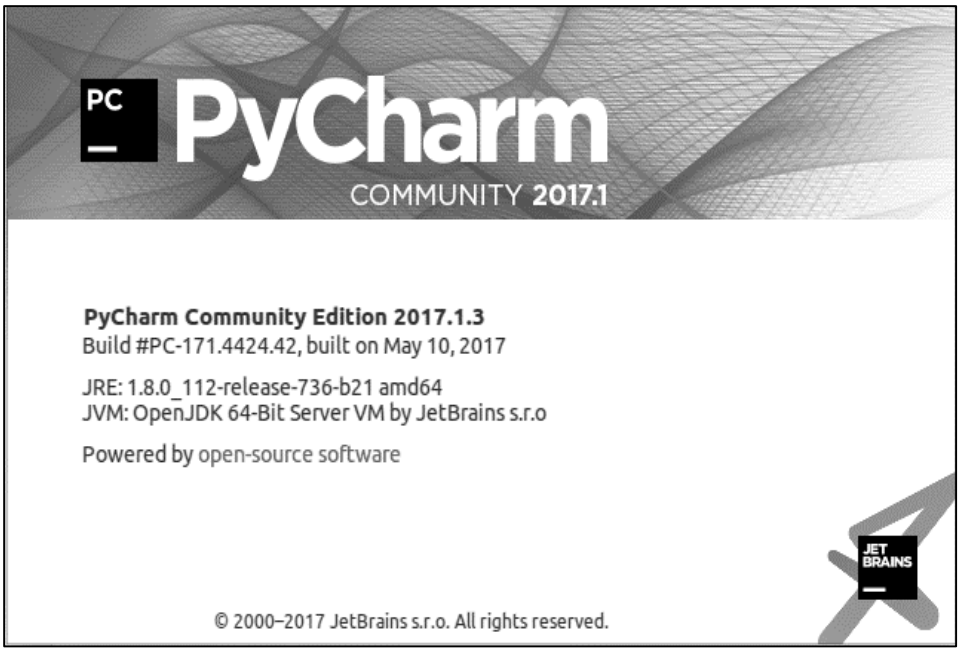


Figure 3.1: PyCharm community edition configuration

Minimum System requirements for PyCharm Community 2017.1.3:

- 512 MB RAM minimum
- 1 GB RAM recommended
- 1024x768 minimum screen resolution
- Python 2.4 or higher

Installation Instructions for PyCharm Community 2017.1.3 are:

- Copy the pycharm-2017.1.3.tar.gz to the desired installation location and rw permissions should be given for that directory.
- Unpack the pycharm-2017.1.3.tar.gz using the following command:
  `tar -xzf pycharm-2017.1.3.tar.gz`
- Run pycharm.sh from the bin subdirectory

3.1.3 Dataset used

Detection of clickbait headlines is the main purpose of the proposed work. Hence for convenience, headlines are collected for clickbait and non- clickbait headlines categories separately. Dataset which is used is in raw format which has an even distribution of 16000 clickbait and 16000 non-clickbait headlines and each headline in the dataset is separated by enter. Dataset is publicly available (Chakraborty *et al.,* 2016)

Non-clickbait dataset contains 16000 non-clickbait headlines which are collected by Newsreader from "TheHindu", "TheGuardian" and "Wikinews". These headlines are considered non-clickbait due to the fact that all of these are reputed media houses and have particular standard to form a headline, so that most of the headlines present by these media houses are not clickbait. And the clickbait dataset contains 16000 clickbait headlines which are collected from "UpWorthy", "ViralNova", "Scoopwhoop", "ViralStories" and "Buzzfeed". To decrease the false negative value Inter-rater agreement is used then 15800 headlines are selected for clickbait category.

10000 headlines from both the categories are used to train the clickbait classifier and rest 5800 headlines from both the categories are used for the testing of clickbait classifier.

3.2 Techniques used

This section and coming subsections discuss the different techniques which are being used in order to implement proposed work. We have raw dataset which is in form of text. Hence we required some technique for handling the text. Though human language either in written or verbal is not understandable by Computers, Natural language processing provides a way for text and language analysis. Other than NLP methods and techniques, Ensemble based classification is used for clickbait classification problem. In order to assess the credibility of headline, it is required to know whether the given headline is clickbait or not. For this purpose ensemble of classifiers is used.

3.2.1 Natural language processing (NLP)

NLP is used for human language analysis. It provides ways for syntax, semantic, lexical and pragmatic analysis of the language. Language is a medium for communication between two. The human language is not understood by the computer system. To solve a

number of real world problem, NLP helps to understand the human language either in written or verbal form to a computer.

News headlines are nothing but a piece of text. For analyse these textual news headlines, NLP is used. We have used Stanford coreNLP Natural language software which is written in Java for implementing NLP tasks in the proposed work. There are some NLP tasks which are useful for the proposed work:

1. **Tokenizing Words and Sentences:** It is the task of chopping it up a text documents into pieces, called "tokens". These tokens may be words or sentences.
2. **Recognizing Stop words:** Some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely.
3. **Stemming:** It is the process of reducing inflected words to their word root form.
4. **Part of Speech Tagging:** It is the process of assigning parts-of-speech to words.
5. **Named Entity Recognition:** It labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names.
6. **Lemmatizing:** It aims to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the "lemma".
7. **Text Classification:** Classify a text into desired category.
8. **Sentiment Analysis:** Classification of a text as positive, negative, or neutral. It tells the sentiment polarity.
9. **N-gram analysis:** N-gram is a contiguous sequence of n items from a given sequence of text. Low level n-gram is 1-gram, then bi-gram with two words.

3.2.2 Proposed ensemble classifier

There are a number of methods which are used to create ensemble, but for implementing ensemble based classification in proposed model we have used voted classifier with soft voting. In this, each classifier is designed independently with their own powers and limitations, then these independent classifiers are combined to form an ensemble classifier. We have used four classifier here to build the ensemble: Support Vector Machine (SVM), Decision Tree Classifier (DT), Random forest classifier (RF), and Neural Network Classifier (NN). The structure of proposed ensemble classifier is presented on Figure 3.2. The final decision is calculated considering the weight given to each classifier.

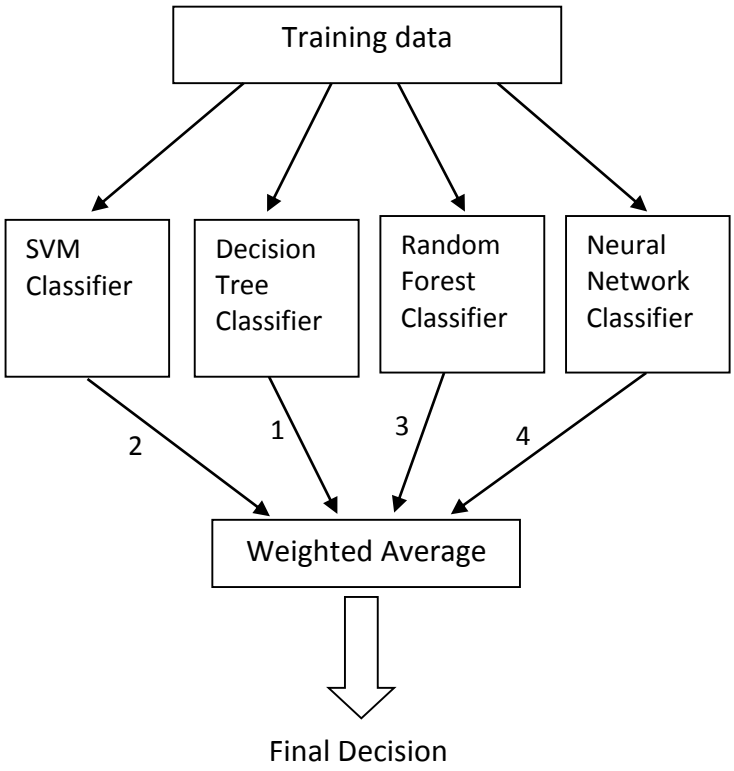


Figure 3.2: Proposed voted ensemble classifier

Each individual classifier is assigned a weight according to its capability. The one with lower error rate assigned with high weight and so on. As shown in Table 3.1, Neural Network (NN) classifier has highest accuracy hence assigned with the highest weight value and Decision Tree (DT) classifier with lowest weight due to its comparative lower accuracy. Here, we have used weighted average instead of majority voting, because majority voting many times failed to decrease the skewness of the final prediction. The model which is less powerful will be considered same as the stronger one in majority voting.

Table 3.1: Accuracy of individual classifiers used in ensemble

Classifiers	Accuracy	Weight
SVM	0.9142	2
DT	0.9008	1
RF	0.9238	3
NN	0.9253	4

Specific weights can be assigned to each classifier via the weights parameter. When weights are provided, the predicted class probabilities for each classifier are collected,

multiplied by the classifier weight, and averaged. The final class label is then derived from the class label with the highest average probability. For example, assume a sample case of class probabilities for class 1 and 2 for all four classifier, Here, the predicted class label is 1, since it has the highest average probability:

Table 3.2: Example for weighted average in voted classifier with soft voting

Classifier	Class 1	Class 2
SVM	2 * 0.7	2 * 0.3
DT	1 * 0.6	1 * 0.4
RF	3 * 0.8	3 * 0.2
NN	4 * 0.2	4 * 0.8
Weighted Average	**5.2**	**4.8**

3.2.2.1 Why ensemble classification increase accuracy

Ensemble classification is the collection of multiple model. Hence, Ensemble models can be considered as an additive expansion of the following form:

$$F(x) = c_0 + \sum_{n=1}^{M} \mathrm{c_m} T_m(x)$$

where the $\{T_m(x)\}, m = 1,2, \dots, M$ are known as base classifiers.

$T(x; p_m)_0^M$ is referring to each base classifier T_m. Here, each base classifier is described by a set of parameters or parameter vector p.

For example, if T_m is a neural net, p_m corresponds to the weights that define the neural net. If T_m is a tree, p_m corresponds to the splits that define the tree. Each base classifier can then be thought of as a "point" in a high-dimensional parameter space P. Ensemble classification problem can be stated as follows:

1. Find the points $p_m \in$ P
2. Find the constants $c_m \in$ R that minimize the average loss.

   The average loss can be described as follows:

$$\{c_m, p_m\}_0^M = \min_{\{c_m, p_m\}_0^M} \sum_{i=1}^{N} L\left(y_i, c_0 + \sum_{m=1}^{M} c_m T(x; p_m)\right)$$

The overall process can be thought as two-step process:

1. Choose the points p_m. Or we can say, "Choose a subset of M base classifiers out of the space of all possible base classifiers from a pre-specified family".
2. Determine the weights or the coefficients c_m.

How to judiciously choose the base classifiers or functions:

The goal is to find "good" $\{p_m\}$ so that the ensemble-based classification is "close" to the target function:

$$F(x;\{p_m\}_1^M,\{c_m\}_0^M) = c_0 + \sum_{m=1}^{M} c_m T(x;p_m) \approx F^*(x)$$

This function is analogous to a high-dimensional integral which is:

$$\int_P I(p)\ \partial p \approx \sum_{m=1}^{M} w_m I(p_m)$$

Common algorithms uniformly choose the points p_m at which the integrand $I(p)$ is evaluated. But certain values of these p_m variables have more impact on the accuracy of the integral being estimated, and thus these "important" values should be emphasized.

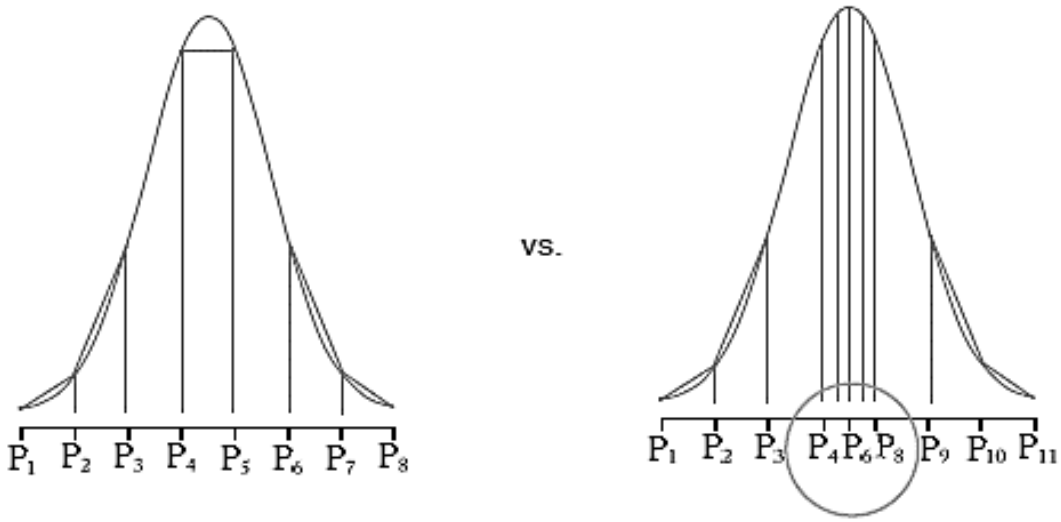


Figure 3.3: Numerical integration of an ensemble function

Figure 3.2 shows how accuracy of the integral improves when we choose more points from the circled region.

In the proposed model, four base classifiers are used:

$$T_1 = SVM\ , T_2 = DT\ , T_3 = RF\ and\ T_4 = NN$$

These four base functions are chosen due to their well performing and diverse as well. From the above description it is known that base function T_m play a vital role to increase accuracy.

Here, $T_1, T_2, T_3, and\ T_4$ are diverse due to their conceptual difference and decision boundaries, that's why these four are selected. All base functions are given with a weightage, and as fig 3.3 represents choosing important parameters increases the accuracy, neural network classifier parameters set is given higher weightage due to its higher accuracy.

3.2.2.2 Selecting ensemble size

Ensemble size refers to the base classifiers used for the classification. It is very crucial to choose a right ensemble size which influences the overall accuracy and also the computation cost. A widely held principle in Statistical and Machine Learning model inference is that accuracy and simplicity are both desirable. But there is a trade-off between the two: a flexible (more complex) model is often needed to achieve higher accuracy, but it is more susceptible to overfitting and less likely to generalize well.

To balance between two, an ensemble size should be selected which will not make the model more complex and also maintains accuracy. Generally, higher the ensemble size, higher will be the accuracy. But, if the ensemble size is too large, overfitting will occur.

Figure 3.4 represents a function which is under fitted and over fitted. The first graph is an example of under-fitted classifier. It is very far from the original decision boundaries, whereas in 3rd graph, classifier is over fitted on the true functions.

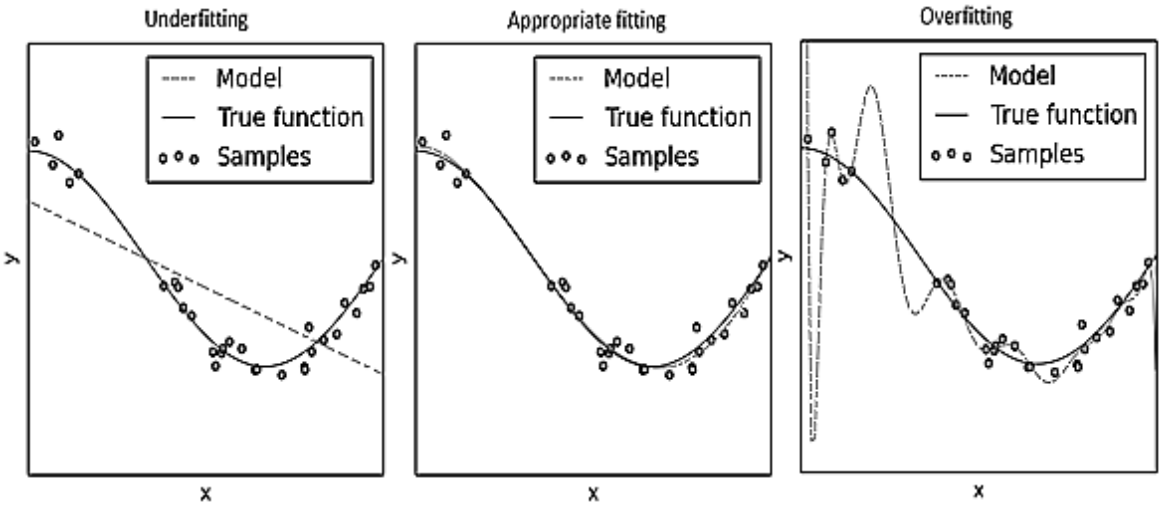


Figure 3.4: Underfitting and Overfitting

In fig 3.4, it can be concluded that there is an optimal ensemble size for which the problem of overfitting and underfitting can be balanced and the result of the classification is optimal due to minimum prediction error.

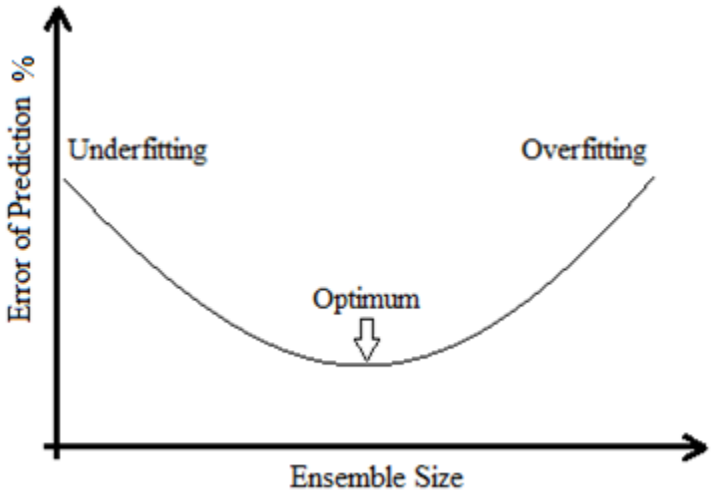


Figure 3.5: Ensemble size vs. prediction error

We have selected different sizes for our ensemble approach starting from 1, and calculate the respective approximate prediction error. In figure 3.5 containing graph in which ensemble size is on x-axis and prediction error is on y-axis, it can be observed that the optimal ensemble size for the proposed classifier is 4. That is the reason, why four base classifiers are selected for designing ensemble classifier.

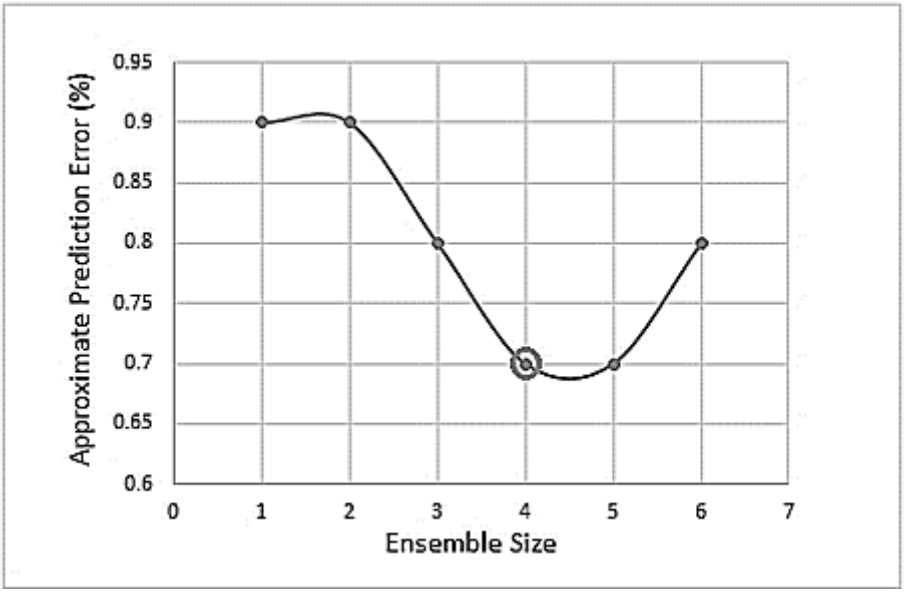


Figure 3.6: Ensemble size vs. prediction error plot for proposed classifier

3.3 Proposed Methodology

The whole process of detecting clickbait news headlines can be decomposed into several steps. First step is to acquire the data. Data is in raw format, hence it is necessary to pre-process it. Then by analysing both types of headlines, different features are extracted from them which may distinguish clickbait and non-clickbait headlines. These features are the base of classifier. At last, classifier is created and headline is classified into clickbait or non-clickbait.

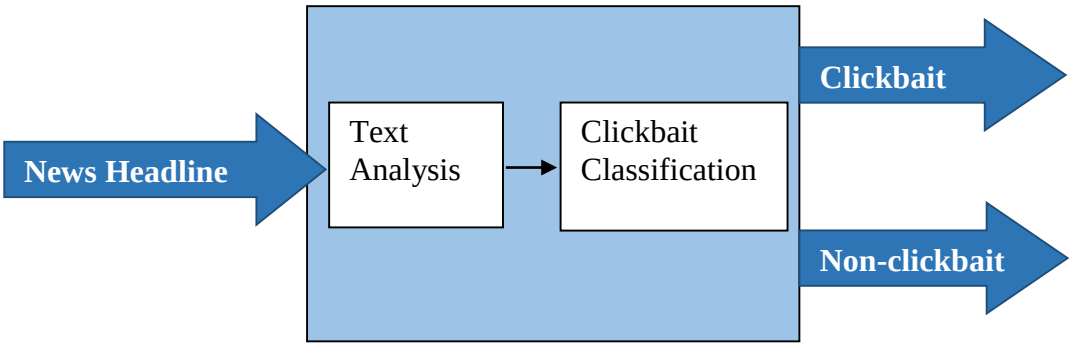


Figure 3.7: Block diagram of proposed system

The overall task can be subdivided into two subtasks: Text Analysis and Clickbait classification. Block diagram for proposed system is shown in figure 3.6. The whole system can think of as black box which takes news headlines as input and returns the status whether it is clickbait or not. Text analysis subpart uses NLP techniques in order to perform news headline analysis. Clickbait classification required Machine learning techniques to design the classifier. After text analysis, the feature vector for each data sample is generated which then fed to the clickbait classification.

The detailed view of both subtasks that are text analysis and classification are shown in figure 3.7 and figure 3.8 respectively. In text analysis, different tasks are performed in order to get the feature set value of the textual headlines.

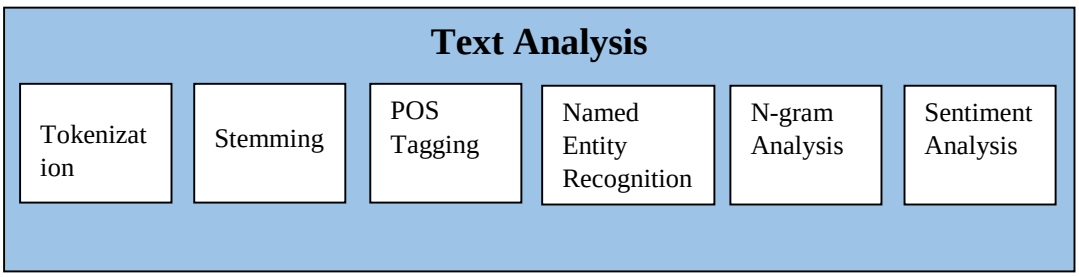


Figure 3.8: Detailed view of text analysis

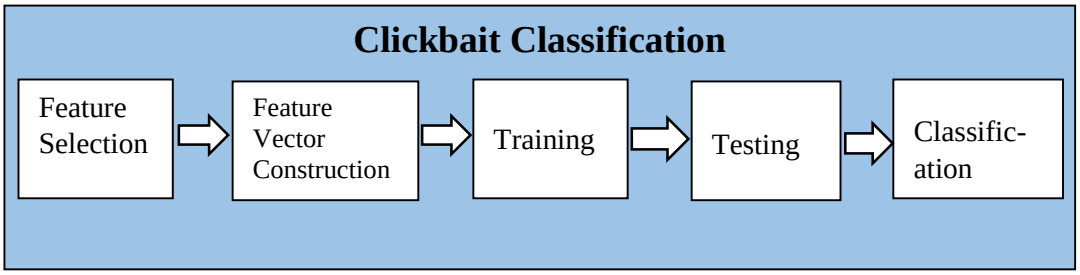


Figure 3.9: Detailed view of clickbait classification

Whole process, from textual headline analysis to clickbait classification can be considered as a flow of steps, which are shown in figure 3.9.

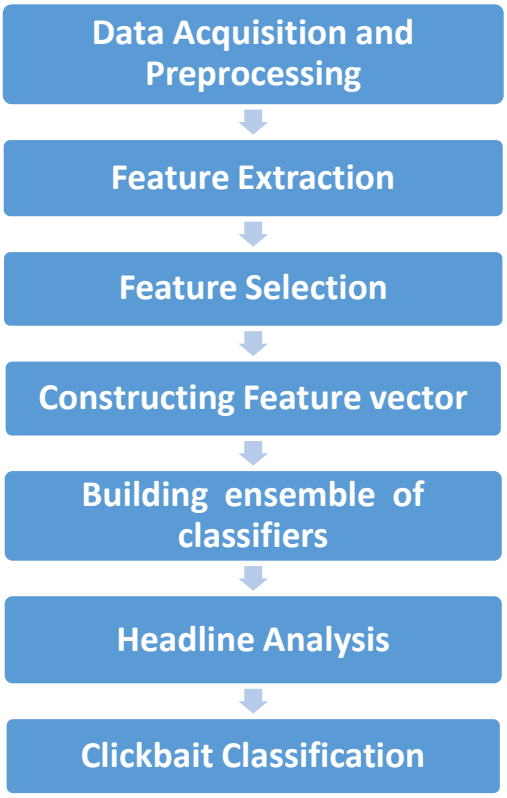


Figure 3.10: Steps used for implementing Proposed System

Step 1: Data acquisition and pre-processing

First step is to acquire the data i.e. online news headlines here. For proposed work, there is a need for clickbait headlines as well as non-clickbait headlines data to design a classifier which may distinguish between them. Firstly, headlines are gathered from online media houses. Clickbait news headlines are collected from different news sites which mostly produce clickbait news like “UpWorthy”, “ViralNova”, “Scoopwhoop”, “ViralStories” and

"Buzzfeed". And the non-clickbait news headlines are collected from some reputed online media houses like "TheHindu", "TheGuardian" and "Wikinews".

Headlines are arranged in the manner that each individual headline is separated with other one by enter (newline). For clickbait data, 16000 headlines are collected and after Inter-rater agreement some of the clickbait headlines are discarded to decrease false negatives.

Step 2: Feature extraction

Feature extraction is a very crucial step for any machine learning task. Whole classification task is dependent on the features selected for the exploit the behaviour of input.

Comparative analysis between clickbait and non-clickbait headlines is performed over different aspects like sentence structure, word structure, the type of language used, n-gram, sentiment polarity and some other. And it is concluded that clickbait headlines have some patterns. Clickbait and non-clickbait headlines are different in structural and linguistic manners. After studying the nature of clickbait headlines, a number of characteristics of clickbait headlines can be observed, these characteristics are as follows:

Characteristics of clickbait headlines:

1. Does not provide complete information, give the only hint about the topic and create fascination about the topic
2. Most of the times, clickbait headlines contain external web links, they encourage to click on the external link provided on the headline. For example, "A lady is thrown out of her own house, Click here to read what happened next".
3. They are formulaic, used some kind of formula, and fill the words in between that formula. For example, "10 facts which will blow your mind", "5 amazing things which will happen to you and will blow your mind". Both the example are structurally similar with some minor changes.
4. They are under deliverer, headlines often don't exactly tell what they are talking about, and they only give enough detail which will make readers to click on them. When they are clicked, no concrete information found on the whole article.
5. Contain numeral. For example, "15 amazing ideas which will make your skin clean and fair like the shining moon." (Chakraborty *et al.*, 2016)
6. Clickbait headlines revolves around some particular topics like lifestyles, entertainment, Bollywood, Hollywood etc.

7. Clickbait headlines are longer than non-clickbait news headlines. It is because clickbait does not contain actual information and by long sentences, they try to confuse readers (Chakraborty *et al.,* 2016).
8. Contains unexpected punctuation marks. Generally, an effective news headline does not have many punctuation marks or contains only '.' or '?' (In some cases only). But clickbait contain informal punctuation patterns like '!', '**', '@', '…' etc.
9. There is more separation between dependent words. For example, "A lady who found dead last week in New York has written a heart-wrenching open letter to the government". Here subject "a lady" and its verb "written" is separated by 9 words, which means it contains high syntactic dependency.
10. Clickbait headlines are very positive. It contain some words with very positive sentiment like super-awesome, heart-wrenching, and very inspiring etc.
11. Contain popular Internet abbreviations and slangs which are in trend because it attract attention of the readers especially youth. Some examples are 'LOL', 'WOW', 'LMAO', 'ROFL', 'FUCK', 'SHIT' etc. (Potthast *et al.,* 2016)
12. The main purpose of clickbait is to immediately gain attention of the readers, hence most of the times, these headlines contain more easy words to increase the understandability of headline.
13. High readability because it should be readable in seconds to make readers bind with the headline.
14. More stop words than an effective headline.
15. Contain general clickbait phrases like "will blow your mind", "what happened next", "things you should know" etc.
16. Clickbait headlines contain more adverbs and pronouns than effective ones.

These characteristics and differences are used to extract the features so that classifier can distinguish clickbait headlines from non-clickbait traditional headlines. Set of some features capture sentence structure and word patters, others capture clickbait language, and n-gram. Some features help in syntactic analysis, some features in semantic analysis and others in text analysis at pragmatic level.

Step 3: Feature selection

All features which are extracted may not produce best result. Many times, adding many features may decrease the accuracy of your model. Feature selection is applied to select only

important feature which participates more. In this phase, low informatics features are discarded in order to increase the overall performance of the clickbait classifier. After extraction and selection of the features, the classifier have total 18 features which are as follows:

Table 3.3: Selected features for ensemble classifier

	Features
Word Level	1. Headline title starts with number 2. Presence of word contractions 3. Presence of unusual punctuations 4. Length of the longest word
Semantic Level	5. Average length of words in headline 6. Length of the headline 7. Ratio of stops to other content words 8. Easy word to other words ratio
Clickbait Language	9. Number of common clickbait phrases 10. Number of common clickbait words and Internet slangs 11. Presence of determiners and pronouns
N-gram Features	12. Word N-grams 13. Part of speech N-grams 14. Syntactic N-grams 15. Longest dependency between words
Contextual Features	16. Subject (context) of the Headline 17. Readability (Flesch-Kincaid) 18. Sentiment Polarity

Step 3: Constructing feature vector

In this step, test is transformed into numerical features so that it can be used for machine learning. The feature vectors represent different feature values corresponding to the news headline. For each headline, value of features calculated numerically which is known

as “feature vector”. Two CSV files are created which contain feature vector for each headline in dataset, one for clickbait class (positive) and second for non-clickbait class (negative).

Step 4: Building ensemble of classifiers

The above generated feature vectors are used to build the classifiers. These vectors are used to train the classifier. Initial 10000 rows from both category are used for training the classifiers and rest 5800 are used for testing the classifiers. In this way, overall 20000 headlines are used as training set and 10600 headlines as testing set. Four classifiers (SVM, DT, RF, and NN) are designed then the Ensemble of classifiers is built with soft voting. Each classifier is assigned with a weight, the classifier which has high accuracy gets more weight and classifier with less accuracy given with less weight. After evaluating performance of all the classifiers individually and ensemble, it came to know that the ensemble increases the overall performance for detecting clickbait news headlines.

Step 5: Headline analysis

A news headline which is to be analysed is fed to the system, first corresponding feature vector is generated for the headline. Each of the feature, described in above section, is computed to form the feature vector out of headline. Then ensemble classifier take this feature vector as input for further prediction.

Step 6: Clickbait classification

In last step, the trained classifier predict the news headline whether it is clickbait or not. In this way, proposed system work and reach its goal to detecting clickbait or misleading news headline.

3.4 Algorithm for Proposed System

Following are the steps used for implementing the proposed system:

Input: Textual News Headline

Output: Is Clickbait? (YES/NO)

Algorithm:

Step 1: First collect clickbait and non-clickbait news headlines.

Step 2: Compute 18 features values for each headline. Headlines are converted to numerical features so that it can be used for further steps of classification.

Step 3: Create feature vector for each headline present in dataset and save it to CSV file. Two CSV files are created, one for clickbait class i.e. positive class and second for non-clickbait class i.e. negative class.

Step 4: Build four classifiers; SVM, DT, RF, NN.

Step 5: Make voted ensemble classifier from the classifiers created in step 4 with soft voting. Specific weights are assigned to each classifier according to its capability. Weights are:

$$w_{svm} = 2,\ w_{DT} = 1,\ w_{RF} = 3,\ w_{NN} = 4$$

Step 6: Fit and train the ensemble classifiers using into input and output vectors through training data present in two CSV files.

Step 7: Compute the Feature vector for the news headline which you want to check.

Step 8: Feed feature vector to ensemble classifier and predict the class.

Step 9: If prediction is 1 then the headline is clickbait. Otherwise, it is not a clickbait headline.

3.5 Evaluation Metrics

After building classifier, it is desirable to evaluate its performance. There are a number of evaluation metrics in Python. These metrics are used to assess the quality of prediction done by the model. Evaluation metrics may be a type of score or a matrix or a curve.

3.5.1 Accuracy score

Accuracy score is the measure of accuracy of the model. It is the fraction or the count of correct prediction. It can be calculated by comparing true table to predicted label in a classification. Formula for calculate accuracy is as follows:

$$accuracy(y_{true}, y_{pred}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(y_{pred_i} = y_{true_i})$$

Where y_{true} is indicating true labels and y_{pred} is indicating predicted labels. Here, conditional summation is used, if y_{true} is equal to y_{pred}, only when 1 got added for each iteration till it reaches to the one less of the sample size. If true and predicted labels are same only then it will added to the accuracy.

3.5.2 Confusion matrix

In confusion matrix, each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class. There are four parameters which are useful for confusion matrix:

1. **True Positive (TP):** The headlines sample which are predicted clickbait and they are really clickbait.
2. **True Negative (TN):** The headlines sample which are predicted non-clickbait and they are really non-clickbait.
3. **False Positive (FP):** The headlines sample which are predicted clickbait but they are not clickbait.
4. **False Negative (FN):** The headlines sample which are predicted non-clickbait but they are clickbait.

	Predicted: YES	**Predicted: NO**
Actual: YES	TP	FN
Actual: NO	FP	TN

3.5.3 K-fold cross validation

In K-fold cross validation, original dataset is partitioned into k subsamples randomly. After creating k samples, k-1 subsamples are used to train the model and one subsample is retained as validation data for testing purpose. This is done k times, in each iteration so that each of the k subsample used for testing and others three subsample for training. Each subsample is used for validation once. Cross validation score is calculated by calculating the mean of the scores for each iteration of testing.

3.5.4 Precision, recall and f1 score

Precision is the measure of how often the model predicted yes and it is correct. It can be calculated by the given formula:

$$precision = \frac{TP}{TP + FP}$$

Recall is the measure of when it's actually yes, how often does it predict yes?

$$recall = \frac{TP}{TP + FN}$$

The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

$$f1\ score = \frac{2 * precision * recall}{precision + recall}$$

3.5.5 ROC curve and ROC-AUC score

ROC curve refers to Receiver Operating Characteristic curve. And the area under ROC curve is known as ROC-AUC score. The ROC curve is plotted between the true positive rate (TPR) and the false positive rate (FPR). It demonstrates TPR on the Y axis, and FPR on the X axis. Top left corner of the ROC curve is the "ideal point" where false positive rate is 0, and true positive rate is 1.

Generally, larger the area under ROC curve (ROC-AUC) better the classifier. ROC plots are used to study the output and quality of the binary classifier. It can be used to evaluate model for decision making or cost-benefit analysis.

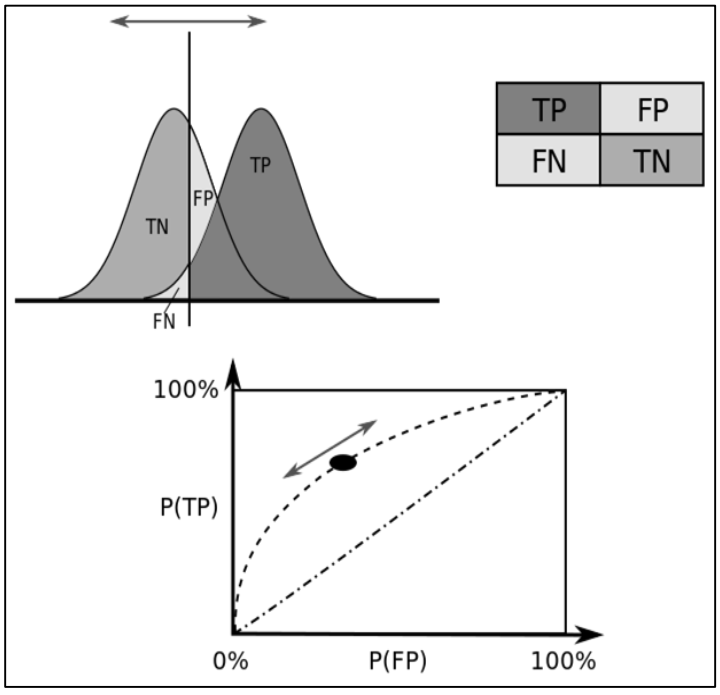


Figure 3.11: Receiver Operating Characteristic curve and its parameters

Results and Discussion

This chapter focuses on the results, implementation and evaluation of the proposed model. The algorithm used for proposed work implemented using Python 2.7.12 in PyCharm Community Edition 2017.1.3 IDE. A collection of clickbait and non-clickbait headlines are collected for the research work.

4.1 Experimental Setup

This section comprise of the important information related to dataset which is used for classification process. The experiment setup is prepared for a dataset containing clickbait and non-clickbait headlines. The performance of the algorithm depends on the used feature set. This section also describes the features which are selected for the ensemble based classifier.

4.1.1 Dataset and its description

Clickbait news headlines are collected from different news sites which mostly produce clickbait news like "UpWorthy", "ViralNova", "Scoopwhoop", "ViralStories" and "Buzzfeed". And the non-clickbait news headlines are collected from some reputed online media houses like "TheHindu", "TheGuardian" and "Wikinews" which are collected by Chakroberty *et al.,* 2014. Dataset have two classes named, clickbait and non-clickbait. Table 4.1 contains the description of dataset and its parameters.

Table 4.1: Dataset description for conducting experiments

Parameters	**Dataset**
Number of instances	31600
Number of classes	2
Total Number of features	18
Number of instances in each class	(15800,15800)
Training instances	20000 (10000, 10000)
Testing instances	11600 (5800, 5800)

4.1.2 Feature set used for ensemble based classification

Comparative analysis between clickbait and non-clickbait headlines is performed over different aspects like sentence structure, word structure, the type of language used, n-gram etc. It is found that clickbait and non-clickbait headlines are different in structural and linguistic manners. These patterns and differences are used for select features so that classifier can distinguish clickbait headlines from non-clickbait traditional headlines. After dimensionality reduction, only important features which contribute more on classification are selected. There are total 18 features in feature set of the proposed algorithm. Features and their types are defined in Table 4.2.

Table 4.2: Feature set used for classification

Feature Type		Features
Syntactical Level	Word Level	1. Headline title starts with number 2. Presence of word contractions 3. Presence of unusual punctuations 4. Length of the longest word
	Sentence Level	5. Average length of words in headline 6. Length of the headline 7. Ratio of stops to other content words 8. Easy word to other words ratio 9. Longest dependency between words
Semantic Level	Clickbait Language	10. Number of common clickbait phrases 11. Number of common clickbait words and Internet slangs 12. Presence of determiners and pronouns
	N-gram Features	13. word N-grams 14. Part of speech N-grams 15. Syntactic N-grams
Pragmatic Level	Contextual Features	16. Subject (context) of the Headline 17. Readability (Flesch-Kincaid) 18. Sentiment Polarity

4.2 Experimental Results

Proposed model takes news headline as input and classify it to clickbait and non-clickbait problem. The following sections contain the command line interface and graphical user interface of the proposed model.

4.2.1 Command line interface

In command line interface, it asks for the news headline which you want to check, one can enter the headline or can paste the news headline to the terminal. After reading and analysing headline, it returns the status. Any time you can press Q or q for exiting from the clickbait detector.

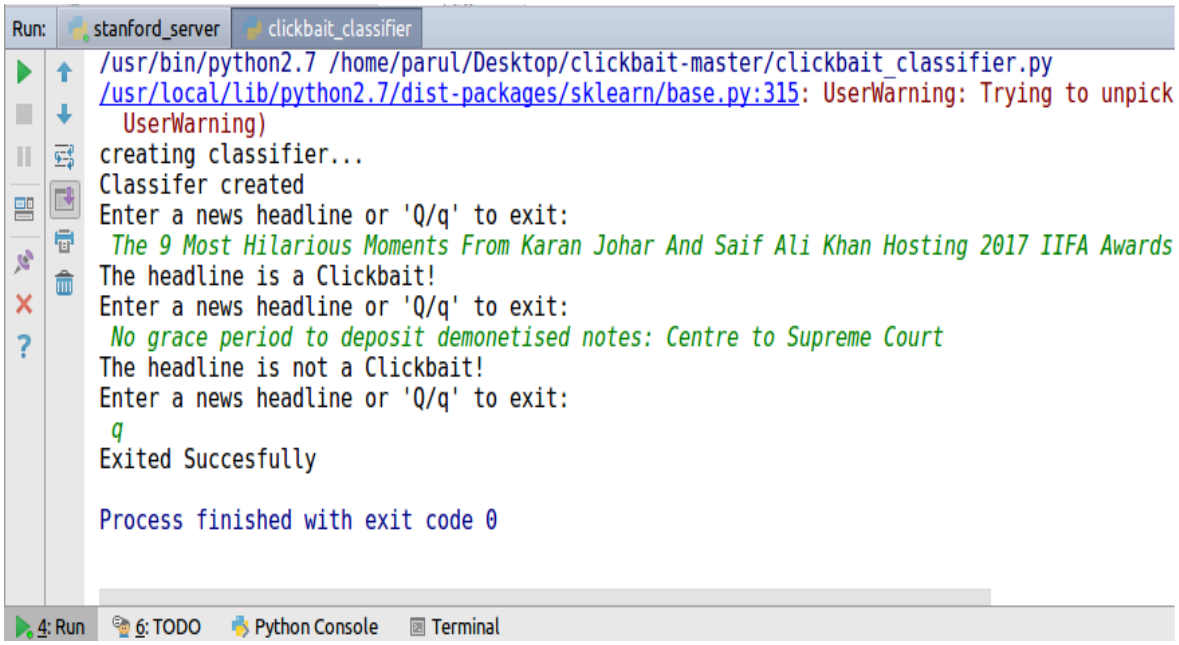


Figure 4.1: Command Line Interface for the proposed work

4.2.2 GUI for the proposed system

Graphical user interface of the proposed system has two option for detecting clickbait headlines. Users are provided with two text box, either they can enter/ paste the news headline or they can paste the news link. If news link is provided, it will automatically parse the web link to get the headline form the server and return it to the graphical window with the status of that headlines whether clickbait or not. Figure 4.2 and 4.3 show the GUI of the proposed model.

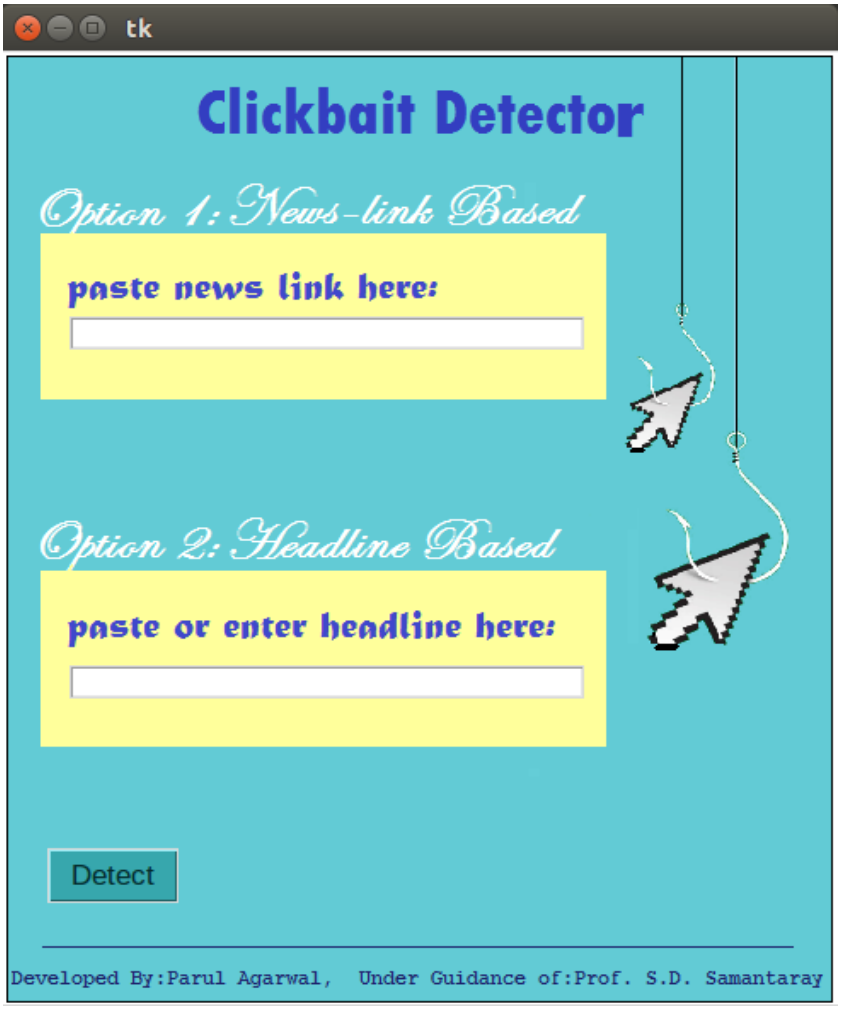


Figure 4.2: Graphical window for the proposed work

4.3 Evaluation

After designing the classifier, it is desirable to evaluate its performance. There are a number of evaluation metrics which are supported by Python. These metrics are used to assess the quality of prediction done by the model. Evaluation metrics may be a type of score or a matrix or a curve. This section presents the evaluation and validation of the proposed model using a number of metrics.

4.3.1 Accuracy

The accuracy of the proposed model is calculated separately for training and testing set separately. The training set contains the instances which are seen by the classifier but the testing set contains the unseen data samples. That is why the testing set accuracy drops.

The proposed model is also validated using 5-fold validation technique. It is an efficient technique to validate your model. Accuracy on training test, testing set and 5-fold cross validation score of the proposed model are presented in Table 4.3.

Table 4.3: Accuracy of the proposed model

Accuracy on training set	95.37%
Accuracy on testing set	93.13%
Accuracy with 5-fold Cross validation	93.08%

4.3.2 Classification report containing precision, recall and f1 score

Classification report contains the precision, recall and f1 scores for each class in it. In information retrieval, precision is a measure of result relevancy, while recall is a measure of how many truly relevant results are returned. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. Precision, recall and f1 score for each class and their average are shown in Table 4.4.

Table 4.4: Precision, recall and F1-score of the proposed model

	Precision	**Recall**	**F1-score**	**support**
Non- clickbait	0.92	0.95	0.93	5800
Clickbait	0.95	0.91	0.93	5800
Average/ Total	0.93	0.93	0.94	11600

4.3.3 Confusion matrix

Confusion matrix contains the values of true positives, true negative, false positive, and false negative. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions. Figure 4.4 and Fig 4.5 show the confusion matrixes of training and testing of the proposed work respectively. Each figure contains normalized and non-normalize matrixes. Normalized confusion matrix contains the percentage values of TP, TN, FP, and FN, while non-Normalized confusion matrix contains the samples counts.

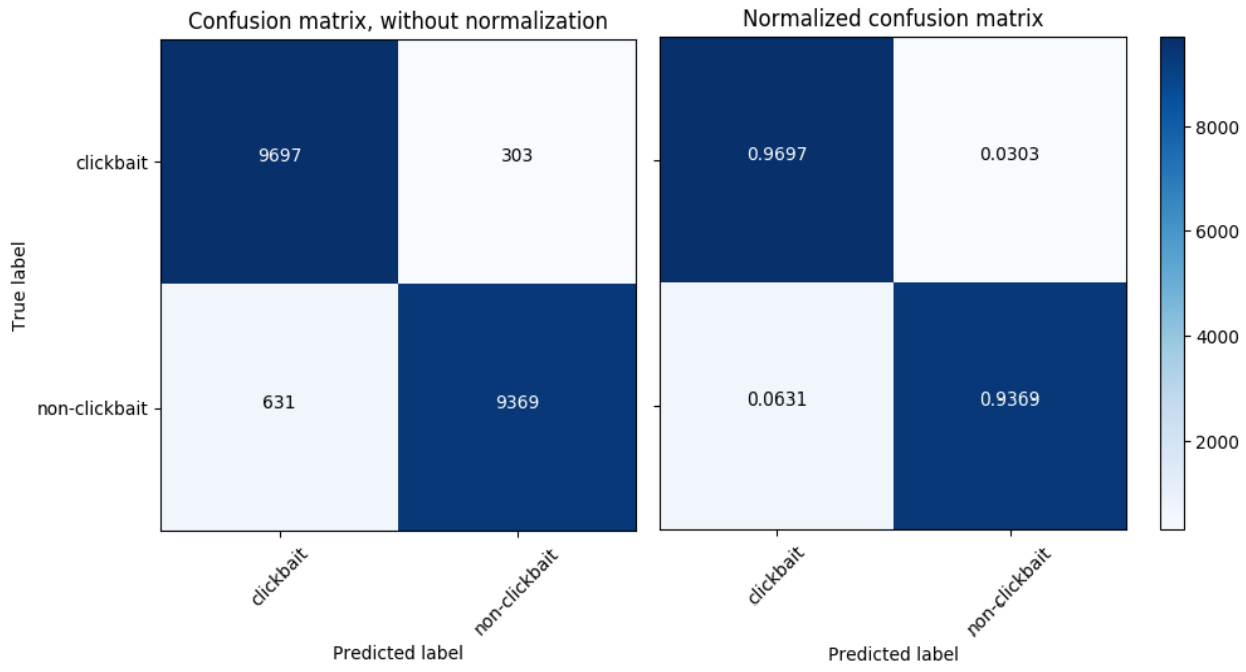


Figure 4.3: Confusion matrix (training)

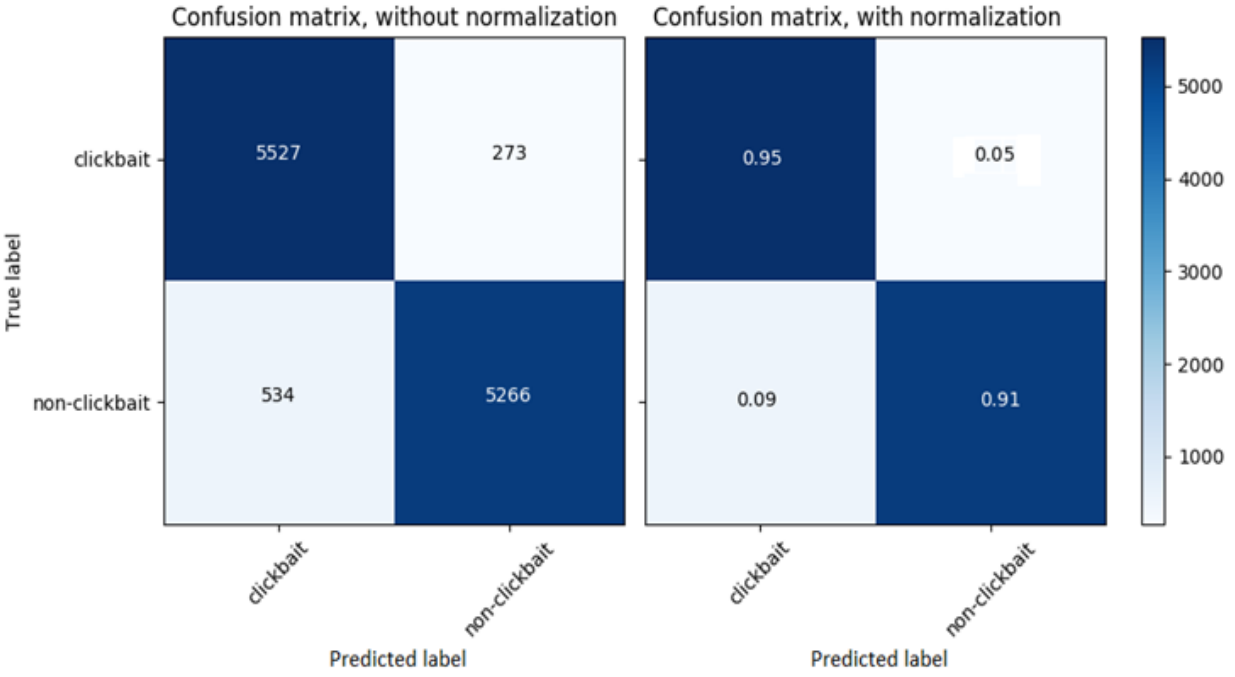


Figure 4.4: Confusion matrix (testing)

With the help of these matrices, it can be observed that the proposed classifier has 0.05% miss-classification rate for the training set and 0.07% miss-classification rate for the training set.

4.3.4 ROC curves and ROC-AUC

The ROC curve is plotted between the true positive rate (TPR) and the false positive rate (FPR). A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

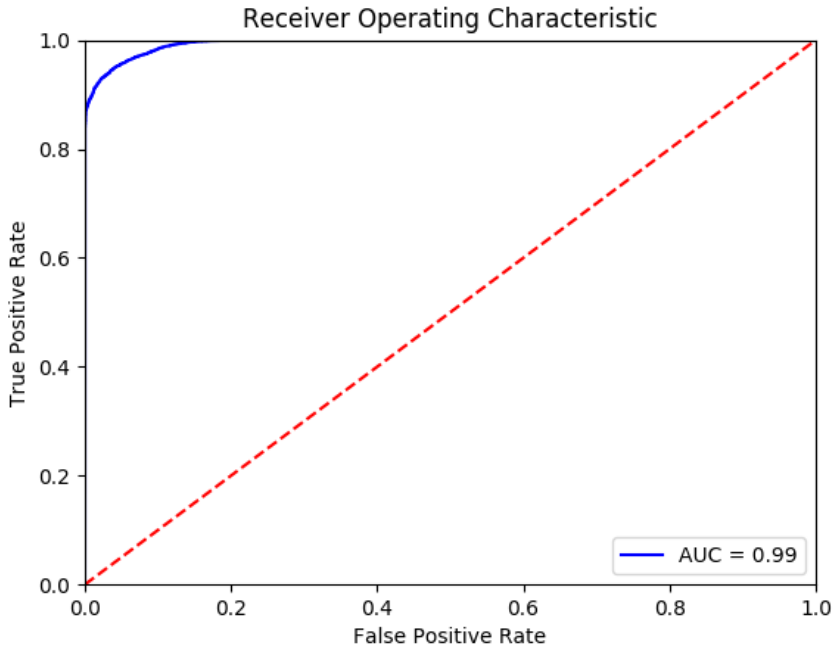


Figure 4.5: ROC curve (training)

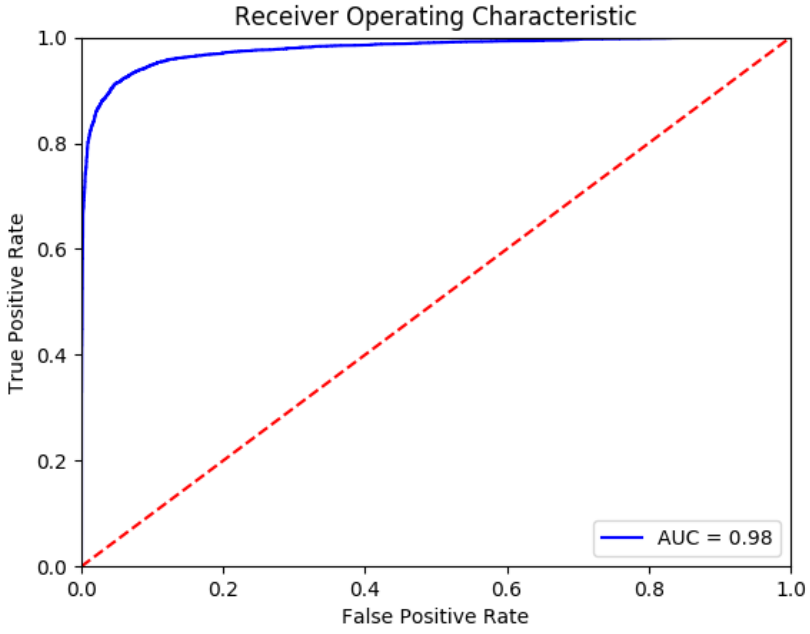


Figure 4.6: ROC curve (training)

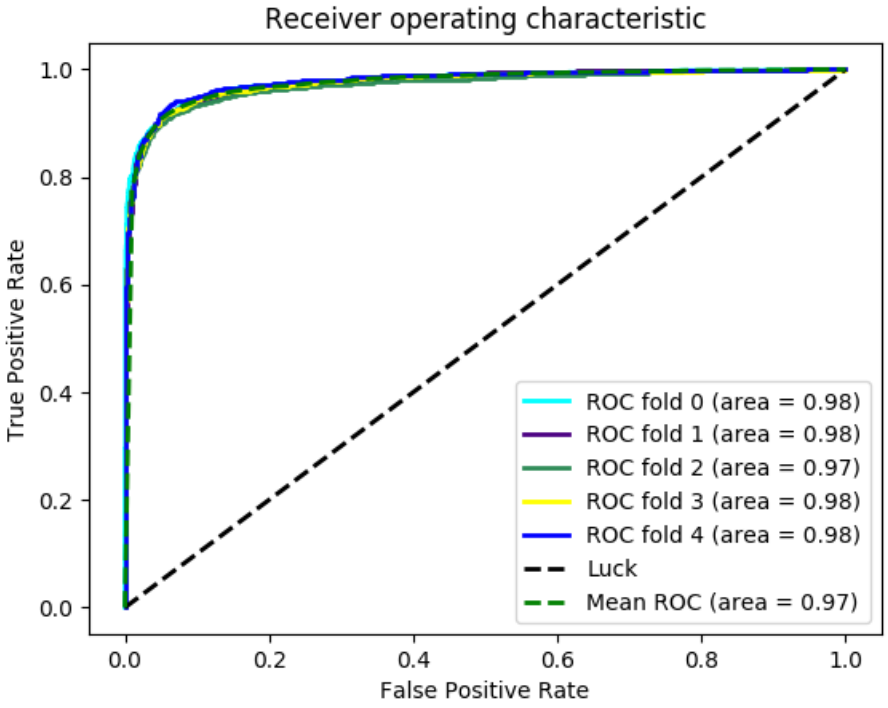


Figure 4.7: ROC curve with 5-fold validation

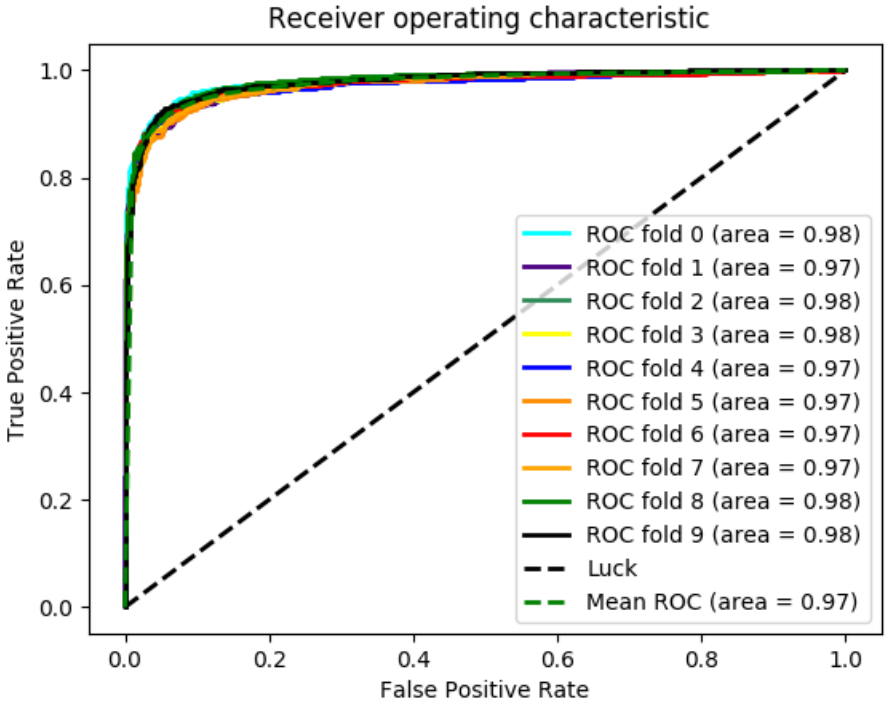


Figure 4.8: ROC curve with 10-fold validation

ROC curves have been plotted for training, testing, 5-fold cross validation, and 10-fold cross validation. ROC curve for training set has 0.99 ROC-AUC. ROC curve for the

testing set has 0.98 ROC-AUC. ROC curves for 5 and 10 fold cross validation contains the ROC-AUC for each fold of the validation process.

Figure 4.6 and 4.7 depict the ROC curves of training and testing of the proposed model. Figure 4.8 and 4.9 show the ROC curves with 5-fold and 10-fold validation respectively. ROC-AUC score for each fold is shown in figures. In each k^{th} fold, out of five subsamples, k subsample is selected for testing and rest 4 samples are used for training.

4.3.5 Comparison among individual classifier and ensemble of classifiers

We have used 18 features for our proposed work and ensemble based classifier. Proposed ensemble classifier is an ensemble of four classifiers which are SVM, DT, RF, and NN. Each of the classifier assigned with weights which are 2, 1, 3, and 4 respectively. The final ensemble classifier used the weighted average for predicating the final result.

The table 4.5 shows that the proposed ensemble classifier successfully outperforms all the other classifiers which are used for designing the ensemble classifier for the proposed work. This is the accuracy and ROC-AUC score which is computed on the testing set containing 11600 headlines.

Table 4.5: Performance of the individual classifiers and proposed classifier

Classifiers	Accuracy	ROC- AUC
SVM	0.9142	0.96
DT	0.9008	0.95
RF	0.9238	0.96
NN	0.9253	0.97
Proposed Model	**0.9313**	**0.98**

Table 4.6 shows the same results as proposed model outperforms all other individual classifiers. Both tables contain the accuracy score as well as ROC-AUC scores for each classifier. And it is clear that ensemble classifier has higher accuracy and ROC-AUC score than all the four classifier which are SVM, DT, RF and NN.

Table 4.6: Performance of the individual classifiers of ensemble and proposed classifier with 5-fold cross validation

Classifiers	Accuracy	ROC- AUC
SVM	0.91	0.967
DT	0.90	0.952
RF	0.92	0.966
NN	0.92	0.973
Proposed Model	**0.93**	**0.975**

4.3.6 Comparison among different ensemble classification methods

There are a number of algorithms which are used for ensemble classification. Some popular algorithms are Bagging, AdaBoost, Random forest, and voted classifier with the majority or soft voting. We have compared all the mentioned algorithm with our proposed model. Table 4.7 contains the accuracy scores of proposed work using different ensemble learning algorithms and the proposed classifier. Our proposed classifier outperforms all the other ensemble classifiers. Base classifier of all of the three algorithms AdaBoost, Bagging, and random classifier is Decision Tree classifier due to its weak learning capability on our dataset. Whereas voted classifier with majority voting is the combination of same four classifier in proposed ensemble approach, the difference in their polling method which is majority in this case unlike proposed one.

Table 4.7: Performance of the classifier using different ensemble classifier

Ensemble Classifiers	Accuracy
AdaBoost classifier	0.9247
Bagging classifier	0.9212
Rndom Forset classifier	0.9238
Voted Classifier with majority voting	0.9225
Proposed Model	**0.9313**

4.4 Discussion

The graphical interface is designed for the user-friendliness and convenience of users. Two options are provided so that they can check the new headline in any way which they find convenient. If a user has the internet access, he can paste the link on GUI, else headline option is also there for the offline case. The user only needs to press a button in order to know the status of the headline.

Proposed work is also evaluated and validated and it is showing better results in comparison to other ensemble techniques. It also outperforms all the other classifiers individually which are used designing the ensemble classifier. Table 4.5, Table 4.6, and Table 4.7 show the testing accuracy, 5-fold cross validation score, and respective ROC-AUC scores. It can be observed that the proposed technique has the highest accuracy of 93.13% and highest 5-fold cross validation score of 93%. Overall performance of proposed approach is better than other approaches.

Summary and Conclusion

Chapter 5 SUMMARY AND CONCLUSIONS

This chapter provides summary a conclusion of the work presented in the thesis. It also discusses the limitations and the future improvements of the proposed work. This thesis is proposed to analyze the credibility of online news and develop a clickbait detection system which automatically classifies the news headlines to clickbait and non-clickbait categories using ensemble based classification.

Online media houses use clickbait headlines in order to gain the readers clicks. Clickbait news compromises with the quality of the news and included exaggerations of news events and sensationalism. Proposed work is designed to detect the clickbait headlines. The problem of clickbait detection can be viewed as a combination task: Text analysis and clickbait classification. For text analysis tasks, Python's NLP tool kit and Stanford CoreNLP open software are used. The first step is to study and analyze the clickbait headlines so that features can be extracted. After feature extraction, feature selection is an important task for classification because low informatics features may decrease the performance. A set of features is chosen which helped to minimize the classification error. Once the features vectors are computed for the headlines, next task is to build a classifier that correctly classify them. For the proposed work, ensemble classifier is designed. It is the ensemble of four conceptually different classifiers which make them diverse: SVM, DT, RF, and NN Classifiers. Diversity is an important factor in ensemble learning for improved results. All four classifiers are assigned a weight with respect to their individual performance over the training set. The final prediction using the weighted average of all four classifiers. Proposed Classifier is trained with 20000 training samples, tested with 11600 test samples, and validated with 5-fold cross validation technique.

5.1 Concluding Remarks

The proposed work used ensemble classifier in order to handle the non-linear decision boundaries between clickbait and non-clickbait samples which will reduced the error rate of classification. The proposed classifier has improved performance over all the four classifiers if trained individually which are used to form proposed ensemble classifier. Experimental

results show that it also showed better results when compared to other algorithms for the ensemble classification; Bagging, AdaBoost, Random Forest, and voted classifier with majority voting.

Graphical user interface is also designed for the proposed work. It is very user-friendly, and readers can directly paste the headline link or headline itself in order to know whether that headline is clickbait or not. This system will aware readers about clickbait news articles so that they can take precautions while dealing with this type of news headline or article.

The current study investigated clickbait headlines as a factor contributing to the credibility of online news. And for detecting these eye-catchy but misleading headlines, ensemble classification method, which is prone to complex decision boundaries, is used so that it can distinguish between effective and clickbait headlines. Feature selection is also used to opt out irrelevant and low informatics features, and by only selecting a subset of relevant features.

5.2 Future Scope

The system can be extended to a multiclass system, which not only classifies into two categories but also gives the value of clickbait-ness over a scale. Many headlines look like clickbait headline but are not actually clickbait. These false negatives can be minimized with the help of multiclass classification. With a multiclass system, it can return the percentage or score value by which a headline is clickbait.

Present system implements a system in which user have to enter news link or headline but this system can be extended to a real time utility, which will automatically extract the news headlines during browsing and indicate the headline's clickbait-ness with help of some indicator for example, star rating. In this era of Facebook, Twitter, users encountered with hundreds of news articles, it will be a very tedious task for a user to copy and paste the news headlines to know whether clickbait or not. A browser extension can be developed which can detect clickbaits automatically from the web page they are visiting, without any extra work by users. It will be more convenient and robust.

Literature Cited

LITERATURE CITED

Agrawal, A., 2016, October. Clickbait detection using deep learning. *In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), IEEE.* pp. 268-272.

Ali, K. M. and Pazzani, M. J. 1996. Error reduction through learning multiple descriptions. *Machine Learning, 24(3).* pp. 173–202.

Anand, A., Chakraborty, T. and Park, N. 2016, April. We used Neural Networks to Detect Clickbaits: You won't believe what happened Next!. *!. In European Conference on Information Retrieval, Springer, Cham.* pp. 541-547.

Banfield, R. E., Hall, L. O., Bawyer, K. W. and Kegelmeyer, W. P. 2007. A Comparison of Decision Tree Ensemble Creation Techniques. *IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 1*. pp. 173-180.

Blom, J. N. and Hansen, K. R. 2015. Clickbait: Forward-reference as lure in online news headlines. *Journal of Pragmatics, vol. 76*: 87-100.

Breiman, L. 1994. Bagging predictors. Technical Report 421, Department of Statistics, University of California, Berkeley.

Breiman, L. 1996. Bagging predictors. *Machine Learning, 24(2).* pp. 123–140.

Castillo, C., Mendoza, M. and Poblete, B. 2011. Information Credibility on Twitter. *International World Wide Web Conference Committee (IW3C2).* pp. 675-684.

Chakraborty, A., Paranjape, B., Kakarla, S. and Ganguly, N. 2016, August. Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media. *In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* pp. 9-16

Chen, Y., Conroy, N.J., Rubin, V. 2015. News in an Online World: The Need for an Automatic Crap Detector. *In The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015).* pp. 6-10.

Chen, Y., Conroy, N.J., Rubin, V. 2015. Misleading Online Content: Recognizing Clickbait as "False News". *ACM WMDD'15.*

Chen, Y., J., Rubin, V. 2017. Perceptions of Clickbait: A Q-Methodology Approach. *In the Proceedings of the 45th Annual Conference of the Canadian Association for Information Science (CAIS/ACSI2017).*

Chung, C. J., Nam, Y. and Stefanone, M. A. 2012. Exploring Online News Credibility: The Relative Influence of Traditional and Technological Factors. *Journal of Computer-Mediated Communication 17*: 171-186.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. and Kuksa, P. 2011. Natural Language Processing (Almost) from Scratch. *In Journal of Machine Learning Research 12*: 2493-2537.

Dietterich, T.G. 2002. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*. pp. 1-22.

Flanagin, A. J. and Metzger, M. J. 2000. Perceptions of Internet Information Credibility. *Journal of Mass communication Quarterly, Vol. 77, No. 3*: 515-540.

Freund, Y. and Schapire, R. E. 1996. Experiments with a new boosting algorithm. *In Proc. 13th International Conference on Machine Learning*. pp. 148-146.

Fritch, J. W., and Cromwell, R. L. 2001. Evaluating Internet resources: Identity, affiliation, and cognitive authority in a networked world. *Journal of the American Society for Information science and Technology, 52(6)*: 499-507.

Fushiki, T. 2009. Estimation of prediction error by using K-fold cross-validation. *Springer Science, Stat Comput (2011) 21*. pp. 137–146.

Garrison, B., Driscoll, P., Salwen M., Abdulla, R. and Casey, D. 2002. The Credibility of Newspapers, Television, and Online News. *Association for Education in Journalism and Mass Communication, annual convention*. pp. 1-30.

Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N. and Perera, A. 2012. Opinion Mining and Sentiment Analysis on a Twitter Data Stream. *The International Conference on Advances in ICT for Emerging Regions – ICTer*. pp. 182-188.

Gupta, A. and Kumaraguru, P. 2012. Credibility Ranking of Tweets during High Impact Events. *ACM PSOSM 12.*

Ha, S. and Ahu, J. 2011. Why are you Sharing Others Tweets?: The Impact of Argument Quality and Source Credibility on Information Sharing Behavior. *Thirty Second International Conference on Information Systems*. pp. 1-8.

Hagen, M., Potthast, M., Buchner, M. and Stein, B. 2015. Twitter Sentiment Detection via Ensemble Classification Using Averaged Confidence Scores.

Hu, X. 2002. Ensembles of Classifiers Based on Rough Sets Theory and Set-oriented Database Operations. *NSF CCF 0514679.*

Kawabe1, T., Namihira, Y., Suzuki, K., Nara, M., Sakurai, Y., Tsuruta, S. and Knauf, R. 2015. Tweet Credibility Analysis Evaluation by Improving Sentiment Dictionary. *IEEE*. pp. 2354-2361.

Kim, H. C., Pang, S., Je, H. M., Kim, D. and Bang S. Y. 2002. Support Vector Machine Ensemble with Bagging. *Springer, Lee and A. Verri (Eds.): SVM 2002, LNCS 2388*: 397–408.

Kiritchenko, S., Zhu, X. and Mohammad, S. M. 2014. Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research 50*: 723-762.

Kohavi, R. 1995. A Study of Cross- Validation and Bootstrap for Accuracy Estimation and Model Selection. *In the International Joint Conference on Artificial Intelligence (IJCAI).*

Krogh, A. and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. *In Advances in Neural Information Processing Systems, vol. 7.*

Loewenstein, G. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin, 116(1).* p.75.

Loper, E. and Bird, S. 2002. NLTK: The Natural Language Toolkit. *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. pp. 62–69.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. and McClosky, D. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. pp. 55-60.

Medhat, W., Hassan, A. and Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal 5*: 1093–1113.

Nasukawa, T. and Yi, J. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *K-CAP'03*. pp. 23–25.

Opitz, D. W. and Shavlik, J. W. 1996. Generating Accurate and Diverse Members of a Neural-Network Ensemble. *In Advances in neural information processing systems, vol. 8*. pp. 535-541.

Opitz, D. and Maclin, R. 1999. Popular Ensemble Methods: An Empirical Study. *Journal of* A*rtificial Intelligence Research 11*: 169-198.

Ordway, D. M. Fake news and the spread of misinformation. journalistsresource.org/studies/society/internet/fake-news-conspiracy-theories-journalism-research.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12.* pp. 2825-2830.

Poliker, R. 2006. Ensemble Bases Systems in Decision Making. *IEEE circuits and systems magazine*: 21-45.

Potthast, M., Kopsel, S., Stein, B., and Hagen, M. 2016. Clickbait Detection. *Springer, ECIR 2016*. pp. 810-817.

Quinlan, J. R. 1996. Bagging, boosting, and C4.5. *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*. pp. 725-730.

Rapoza, K. Can 'Fake News' Impact The Stock Market? forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market.

Rokach, L. 2009. Ensemble-based classifiers. *Springer, Artificial Intelligence Rev 33*. pp. 1–39.

Rubin, V. L., Chen, Y., and Conroy, N. 2015. Deception Detection for News: Three Types of Fakes. *In The Proceedings of the Association for Information Science and Technology Annual Meeting (ASIST2015).*

Russell, N. S. Can Headlines Change the Way You Think About Trust? psychologytoday.com/blog/trust-the-new-workplace-currency/201509/can-headlines-change-the-way-you-think-about-trust.

Scacco, J., Muddiman, A. Investigating the Influence of "Clickbait" News Headlines. engagingnewsproject.org/research/clickbait-headlines/.

Schapire, R.E. 1990. The strength of weak learnability. *Machine learning, 5(2)*. pp.197-227.

Schweiger, W. 2000. Media Credibility - Experience or Image?: A Survey on the Credibility of the World Wide Web in Germany in Comparison to Other Media. *European Journal of Communication 2000, Vol. 15*: 37-59.

Sebastiani, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys, Vol. 34, No. 1*. pp. 1-47.

Viner, K. How technology disrupted the truth. theguardian.com/media/2016/jul/12/how-technology-disrupted-the-truth.

West, E. Forget fake news. The bigger problem is misleading news. blogs.spectator.co.uk/2017/04/forget-fake-news-bigger-problem-misleading-news/.

Yadav, S. and Shukla, S. 2016. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *In 6th International Advanced Computing Conference*. pp. 78-83.

Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. 2003. Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. *In Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03)*.

Yi, J. and Nasukawa, T. 2003. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *ACM K-CAP'03*. pp. 23–25.

APPENDIX A

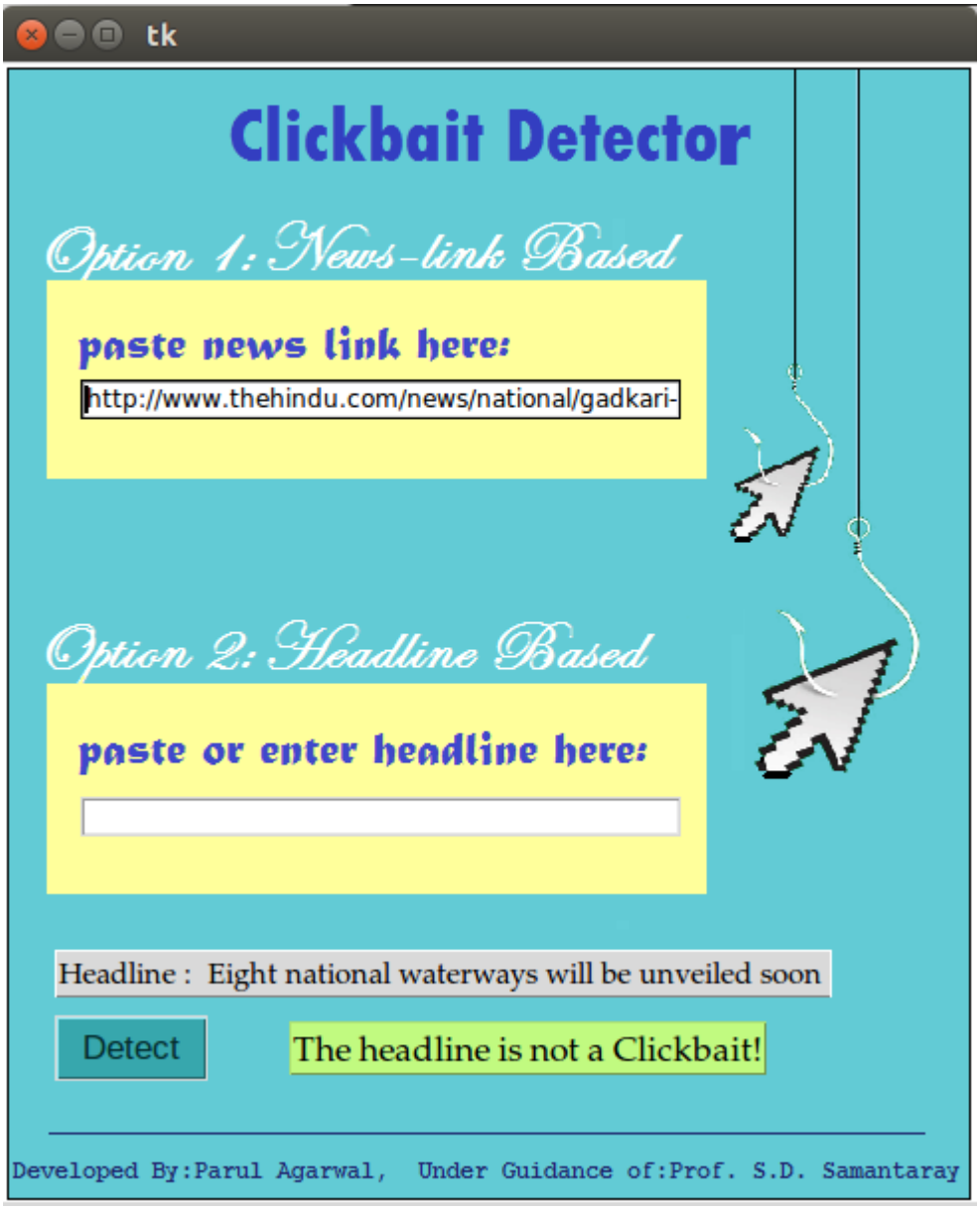
Screenshots

1. **Headline Sample 1:** Eight national waterways will be unveiled soon.

   Source: TheHindu

   Status: Non-Clickbait

   Option 1: News-link based

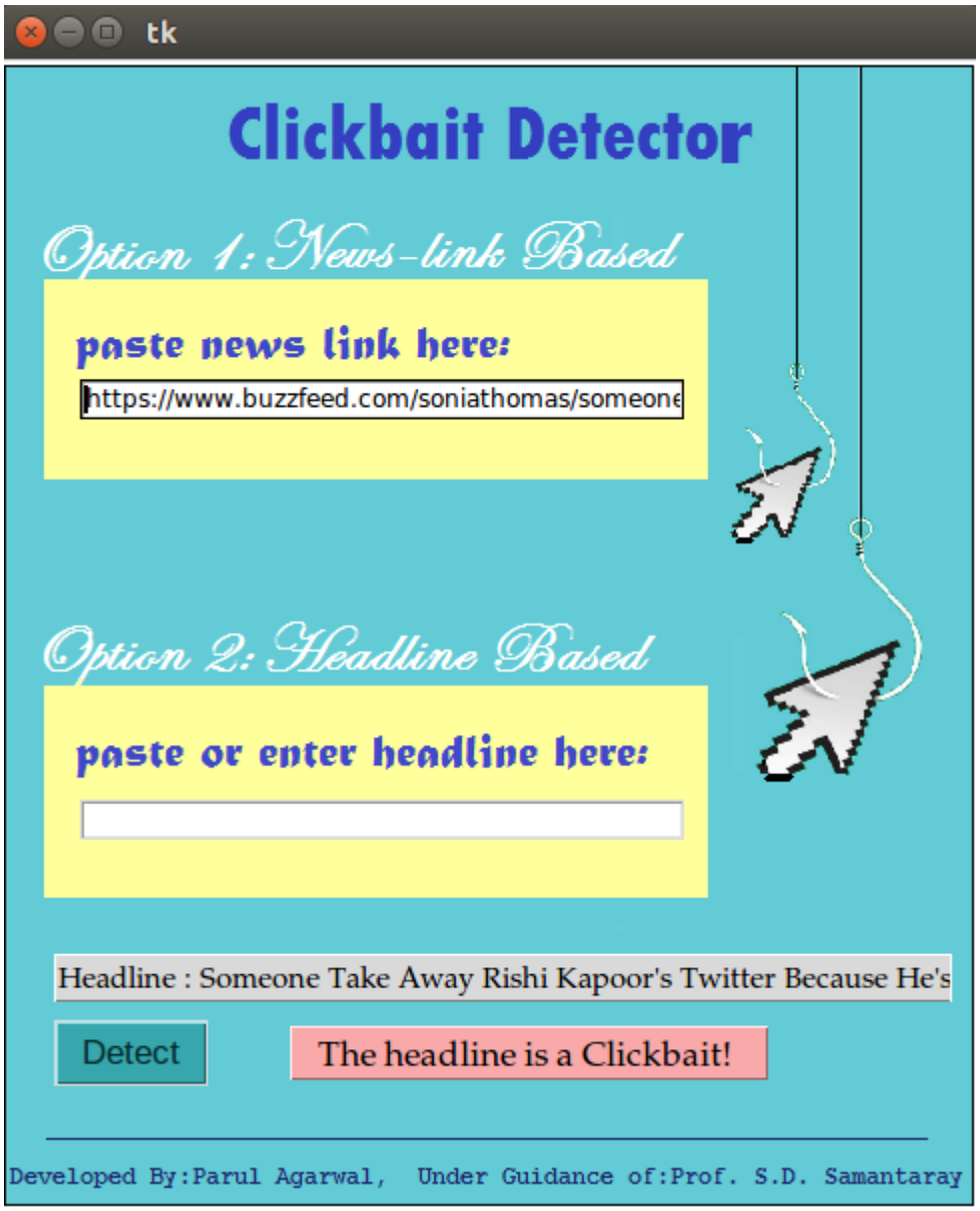


2. **Headline Sample 2:** Someone Take Away Rishi Kapoor's Twitter Because He's Done It Again.

   Source: Buzzfeed

   Status: Clickbait

   Option 1: News-link based

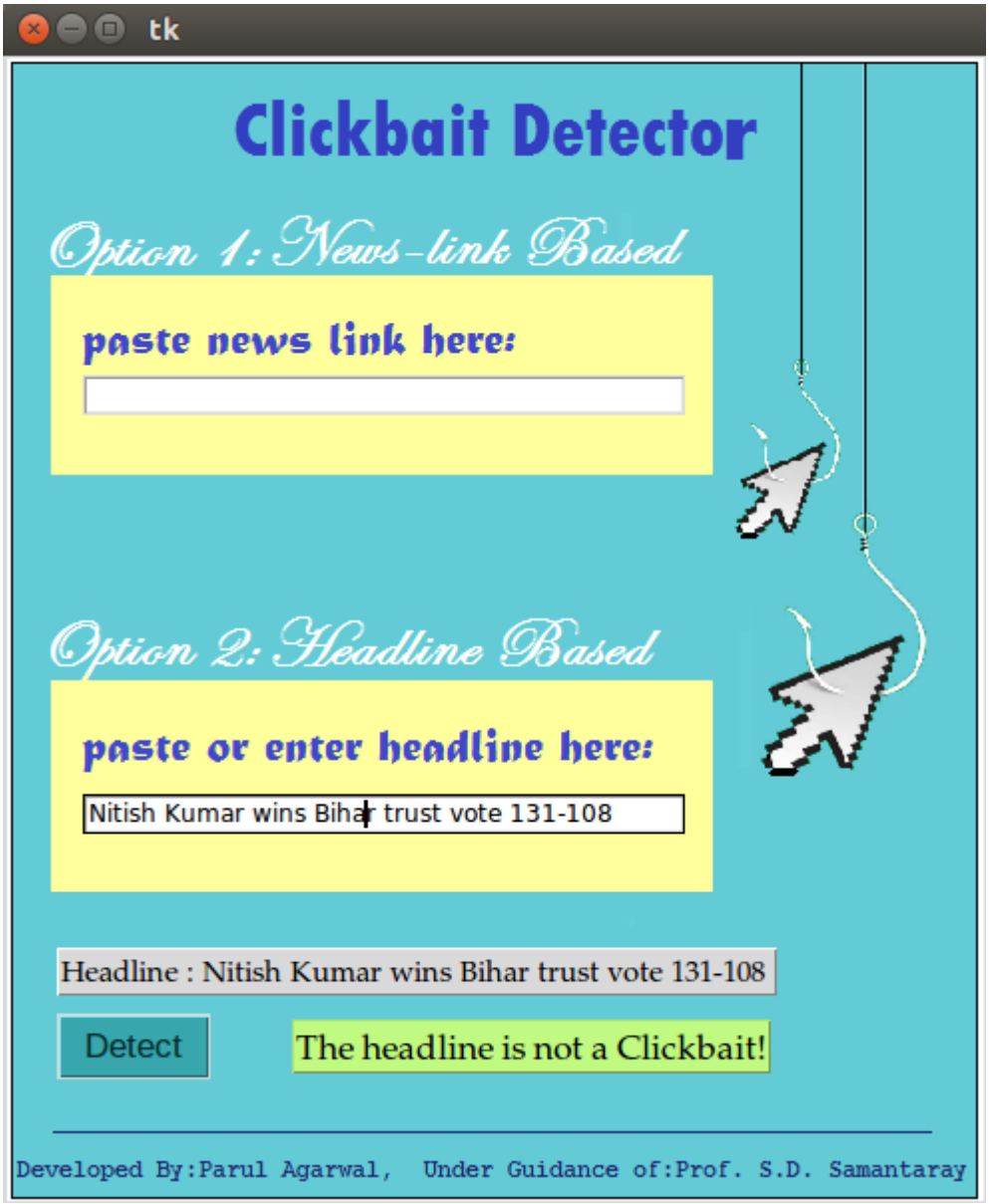


3. **Headline Sample 3:** Nitish Kumar wins Bihar trust vote 131-108.

   Source: NDTV

   Status: Non-Clickbait

   Option 2: Headline based



4. **Headline Sample 3:** She took a day off to focus on mental health. Her CEO's response has gone viral.

   Source: Scoopwhoop

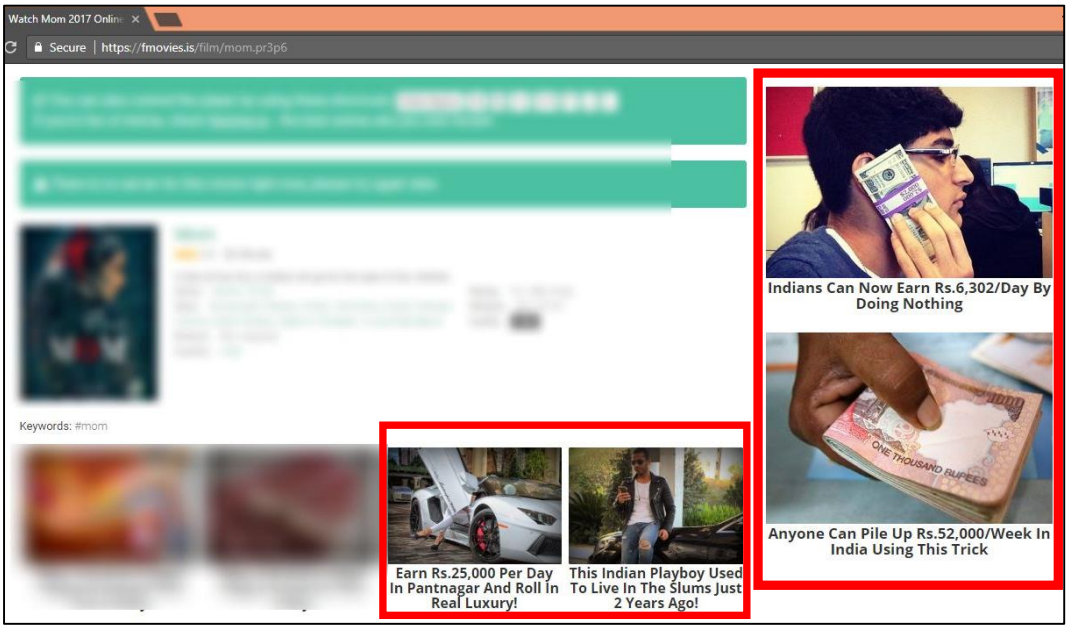
   Status: Clickbait

   Option 2: Headline based

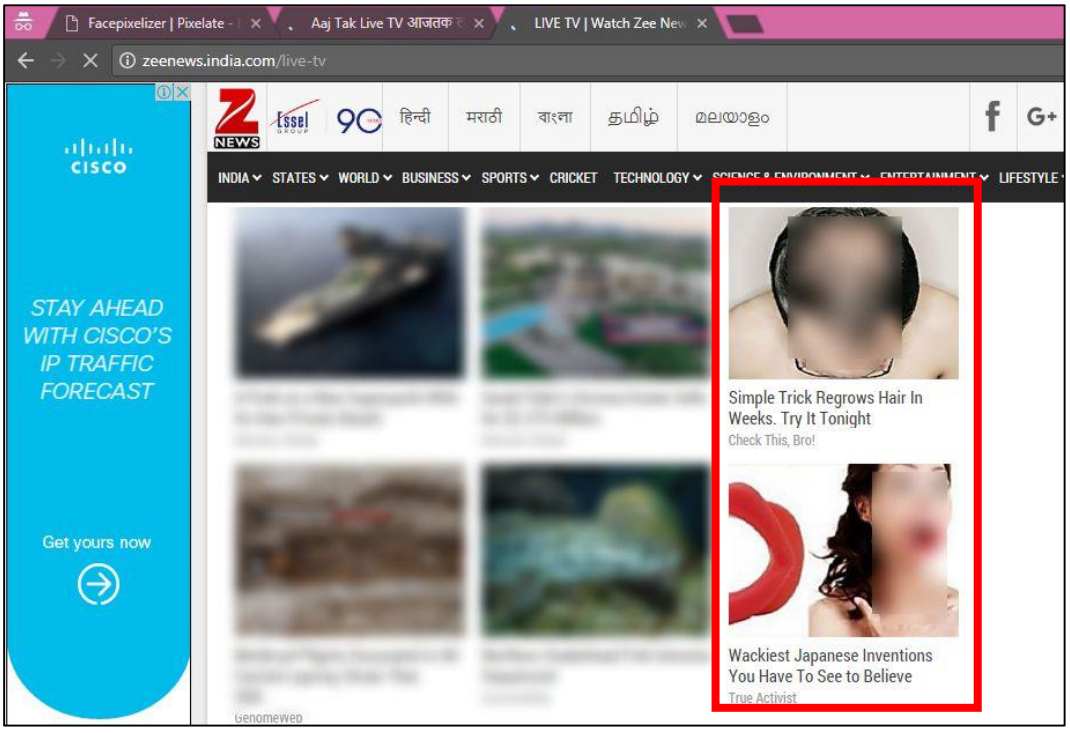
APPENDIX B

Examples of clickbait headlines

1. A movie site with clickbait headlines.



2. A website of a famous Indian media-house with clickbait headlines.



VITA

The authoress, Parul Agarwal, was born on 11th February 1993 in Udham Singh Nagar district of Uttarakhand. She passed her High School and Intermediate Examination from Saraswati Vidhya Mandir, Sitarganj (affiliated to Uttarakhand Board) in 2008 and 2010 respectively. She earned her bachelor's degree in Computer Science and Engineering from SLSET Group of Institutions, Kichha (affiliated to Uttarakhand Technical University) in 2014 with 81.9% marks. She took admission in the College of Post Graduate Studies at Govind Ballabh Pant University of Agriculture & Technology, Pantnagar in July 2015 for Master's degree programme in Computer Engineering.

Address:

Parul Agarwal
D/o Mr. Mahesh Agarwal
Town and P.O. –Sitarganj
Pin- 262405
District – Udham Singh Nagar
Uttarakhand
E-mail ID: er.parul11agarwal@gmail.com
Phone no.: +919045303620

ABSTRACT

Name	**:** Parul Agarwal	**Id. No.**	**:** 49406
Semester & Year of admission	**:** I, 2015-2016	**Degree**	**:** Master of Technology (Computer Engineering)
Major	**:** Computer Engineering	**Department**	**:** Computer Engineering
Thesis Title	**:** **"**An Ensemble Based Classification Approach for Credibility Analysis of Online News by Detecting Clickbait News Headlines"		
Advisor	**:** Prof. S.D. Samantaray		

The present work proposes a methodology for detecting clickbait news headlines in online news media using Ensemble based classification Technique. In this era of Digitization, presenting news now became online. Everyone is accessing online news by one or other medium. When online news is so popular and easily accessible, it also makes online news vulnerable too. Anyone can write anything in the name of news and it becomes viral whether it is informative or not. Due to the high competition and thrust of clicks, clickbait headlines are manufactured just to attract readers to click. These headlines generate enough curiosity by using some tactics so that readers compelled to click on the link to fill the knowledge gap. Clickbait headlines are compromising the meaning of true journalism.

The present work is aimed to develop a clickbait detection system for analyzing the credibility of online news. So that the readers become aware and do not click on these links. News headlines are a piece of text, hence the proposed task is divided into two subtasks; Text analysis and classification. Text analysis is done for the transformation of text into numerical features usable for machine learning. These numerical features are then used for training the ensemble based classifier. The training dataset contains 10000 clickbait and 10000 non-clickbait headlines. Python 2.7 is used for the programming and system is tested for 10600 news headlines which are in an even distribution of 5800 clickbait and non-clickbait headlines and gained 93.13% accuracy. This system is also validated using k-fold cross validation technique.

S.D. Samantaray

(Advisor)

Parul Agarwal

(Authoress)

सारांश

नाम	: पारूल अग्रवाल	**परिचायंक**	: ४९४०६
षठमास एवं प्रवेश वर्ष	: प्रथम, २०१५-२०१६	**उपाधि**	: मास्टर ऑफ़ टेक्नोलॉजी (संगणक अभियांत्रिकी)
मुख्य विषय	: संगणक अभियांत्रिकी	**विभाग**	: संगणक अभियांत्रिकी

शोध ग्रंथ शीर्षक : "एन्सेम्बल आधारित वर्गीकरण पद्धति का प्रयोग करते हुए क्लिकबैट हेडलाइनों का पता लगाकर ऑनलाइन समाचारों का विश्वसनीयता विश्लेषण"

सलाहकार : प्रो० एस.डी. सामंतराय

इस शोध कार्य में ऑनलाइन समाचारों में क्लिकबैट समाचार सुर्खियों का पता लगाने के लिए एन्सेम्बल आधारित वर्गीकरण पद्धति का प्रस्ताव है। डिजिटलीकरण के इस युग में, समाचार पेश करने का तरीका अब ऑनलाइन हो गया है। प्रत्येक व्यक्ति किसी न किसी माध्यम से ऑनलाइन समाचारों तक पहुंच रहा है। एक तरफ जहाँ ऑनलाइन समाचार इतना लोकप्रिय और सुलभ है, यही इसे कमजोर भी बनाता है। उच्च प्रतिस्पर्धा के कारण, क्लिकबेट हेडलाइनें तैयार की जाती हैं, ताकि पाठकों को क्लिक करने के लिए आकर्षित किया जा सके। ये सुर्खियाँ कुछ रणनीति का उपयोग करके पर्याप्त जिज्ञासा पैदा करती है ताकि पाठक उन लिंक पर क्लिक करे। क्लिकबैट हेडलाइनें ईमानदार पत्रकारिता के अर्थ के साथ समझौता कर रहीं हैं।

प्रस्तावित कार्य का उद्देश्य क्लिकबैट हेडलाइनों की पहचान करके ऑनलाइन समाचार की विश्वसनीयता का विश्लेषण करना है। ताकि पाठकों को इन हेडलाइनों के बारे जानकारी हो और वें इन लिंक पर क्लिक न करें। समाचार हेडलाइनें लिखित भाषा का ही एक रूप है, इसलिए प्रस्तावित कार्य को दो उप-कार्यों में बांटा गया है; लिखित भाषा-विश्लेषण और हेडलाइनों का वर्गीकरण। भाषा का विश्लेषण करके एन्सेम्बल आधारित वर्गीकारक को प्रशिक्षित किया गया। प्रशिक्षण डाटासेट में १०००० क्लिकबैट और १०००० गैर-क्लिकबैट हेडलाइनें शामिल हैं। इस शोध कार्य में पायथन २.७ प्रोग्रामिंग भाषा का उपयोग किया गया है एवं १०८०० समाचार हेडलाइनों का उपयोग परीक्षण के लिए किया गया है, जिसमें ५८०० क्लिकबैट और ५८०० गैर-क्लिकबैट हेडलाइनें हैं। प्रस्तावित प्रणाली को मान्य करने के लिए के-फोल्ड क्रॉस वेलीडेशन तकनीक का भी उपयोग किया गया है। प्रस्तावित कार्य की सटीकता ९३.१३% है।

(प्रो० एस. डी सामंतराय)
सलाहकार

(पारूल अग्रवाल)
लेखिका