Project Report

On

Fake news and Sentiment Detector

In partial fulfillment of requirements for the degree

of

BACHELOR OF TECHNOLOGY
IN

INFORMATION TECHNOLOGY

Submitted by:

Bhavesh Somnani [1710DMTIT01489]

Ismith Gehlot [1710DMTIT01494]

Kapil Odiya [1710DMTIT01495]

Manas Joshi [1710DMTIT01497]

Under the guidance of

PROF. Sujit Badodia

PROF. Jayesh Surana

SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE
SHRI VAISHNAV INSTITUTE OF INFORMATIN TECHNOLOGY
DEPARTMENT OF INFORMATION TECHNOLOGY

JULY-DEC 2020

SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE
SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

DECLARATION

We here declare that work which is being presented in the project entitled "**Fake News and Sentiment analysis**" in partial fulfillment of degree of **Bachelor of Technology in Information Technology** is an authentic record of our work carried out under the supervision and guidance of (**Mr. Sujit Badodia)** Asst. Professor of Information Technology. The matter embodied in this project has not been submitted for the award of any other degree.

(Sign)
Bhavesh Somnani
Ismith Gehlot
Kapil Odiya
Manas Joshi

Date:

SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE
SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

PROJECT APPROVAL SHEEET

Following team has done the appropriate work related to the "**Fake News and Sentiment Detector**" in partial fulfillment for the award of **Bachelor of Technology in Information Technology** of "SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY" and is being submitted to SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE.

Team:
1. Bhavesh Somnani
2. Ismith Gehlot
3. Kapil Odiya
4. Manas Joshi

Internal Examiner **External Examiner**

Date

SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE
SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that **Mr. Bhavesh Somnani**, **Mr. Ismith Gehlot, Mr. Kapil Odiya** and **Mr. Manas Joshi** working in a team have satisfactorily completed the project entitled "**Fake News and Sentiment Detector**" under the guidance of Mr. Sujit Badodia and Mr. Jayesh Surana in the partial fulfillment of the degree of **Bachelor of Technology in Information Technology** awarded by SHRI VAISHNAV INSTITUTE OF INFORMATION TECHNOLOGY affiliated to SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE during the academic year **July 2020-Dec 2020**.

Prof. Sujit Badodia
Project Guide

Prof. Manish Kumar
Project Coordinator

Dr. Jigyasu Dubey
Head, Department of Information Technology

ACKNOWLEDGEMENT

We are grateful to a number of persons for their advice and support during the time of complete our project work. First and foremost, our thanks go to **Dr. Jigyasu Dubey** Head of the Department of Information Technology and **Mr. Sujit Badodia** and **Mr. Jayesh Surana** the mentor of our project for providing us valuable support and necessary help whenever required and also helping us explore new technologies by the help of their technical expertise. His direction, supervision and constructive criticism were indeed the source of inspiration for us.

We would also like to express our sincere gratitude towards our Director **Dr. Anand Rajavat** for providing us valuable support.

We are really indebted to **Prof. Manish Kumar,** project coordinator for helping us in each aspect of our academic's activities. We also owe our sincere thanks to all the **faculty members** of Information Technology Department who have always been helpful.

We forward our sincere thanks to all **teaching and non-teaching staff** of Information Technology department, SVVV Indore for providing necessary information and their kind co-operation.

We would like to thanks our parents and family members, our classmates and our friends for their motivation and their valuable suggestion during the project. Last, but not the least, we thank all those people, who have helped us directly or indirectly in accomplishing this work. It has been a privilege to study at SHRI VAISHNAV VIDYAPEETH VISHWAVIDYALAYA, INDORE

ABSTRACT

The project mainly aims for the fake news classification and sentiment analysis which is using NLP (Natural Language Processing) which is the one of the applications of machine learning. Rather than that the application also focuses on other NLP application which are spam classification and summarize the context. The application is a web-based application where user can access the application and uses any of the four application of the application mention above. The Necessity of the such is required because fake news and havoc can produce disturb in the presence of the public and in such advance age fake news can reach out to everyone really fast. To prevent such disturbance any citizen can assure themselves by using this application.

Fake news has been a hot topic in the last few years in the form of Troll Farms and these Hoax News attempt to create public unrest like Lynching, Cyber Mobbing, Subvert and influence the public perceptions using social media platforms. Online life for news utilization is a twofold edged blade. From one viewpoint, its minimal effort, simple access, and fast scattering of data lead individuals to search out and expend news from internet-based life. Then again, it empowers the wide spread of "Fake news", i.e., low quality news with purposefully bogus data. The broad spread of phony news has the potential for very negative effects on people and society. The list of sources from which the fake news is trolling such as social websites, applications like WhatsApp, Facebook, Instagram, fake news channel and other sites many more etc. The main aim of this project is to make a desktop application which is capable to detect the fake news and gives the output to the user.

List of Figures

TABLE OF CONTENT

INTRODUCTION

1.1 INTRODUCTION

In recent times the machine learning applications have gathered all the major attention towards it in IT sector and it has been research further by many big-time researchers till date which means it has quite a bit of importance to it. This project focuses on one of the application machine learning which is NLP (Natural Language Processing) which deals with text rather than numerical data like other machine learning algorithm, but the Machine learning algorithm still works on the numerical expression so the NLP handles the text data and covert them in numerical representation so we can apply algorithm on the data to get the result . This project contains four feature which are using different algorithm to get the result. The features are Fake news detector, Sentiment analysis, Spam Classification and summarizer. We will explain them one by one and their importance as well.

Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is: fictitious articles deliberately fabricated to deceive readers". Social media and news outlets publish fake news to increase readership or as part of psychological warfare. In general, the goal is profiting through clickbaits. Clickbaits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues.

Sentiment analysis is the process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers. Since customers express their thoughts and feelings more openly than ever before, sentiment analysis is becoming an essential tool to monitor and understand that sentiment.

The upsurge in the volume of unwanted emails called spam has created an intense need for the development of more dependable and robust antispam filters. Machine learning methods of recent are being used to successfully detect and filter spam emails. We present a systematic review of some of the popular machine learning based email spam filtering approaches. Discussion on general email spam filtering process, and the various efforts by different researchers in combating spam through the use machine learning techniques was done. Our review compares the strengths and drawbacks of existing machine learning approaches and the open research problems in spam filtering.

1.2 PROBLEM STATEMENT

The Internet has already reached every corner of the world and had deepen its reachability so any false information can be reach out every person in the world within some instance like a wild fire so a false information can create the disturbance in the life of the individual as well as the whole society. The explosive growth in fake news and its erosion to democracy, justice, and public trust has increased the demand for fake news analysis, detection and intervention.

The survey comprehensively and systematically reviews fake news research. The survey identifies and specifies fundamental theories across various disciplines, e.g., psychology and social science, to facilitate and enhance the interdisciplinary research of fake news. A narrow definition of fake news is news articles that are intentionally and verifiably false and could mislead readers. Fake news includes false information and is created with dishonest intention to mislead consumers.

It's estimated that 90% of the world's data is unstructured, in other words it's unorganized. Huge volumes of business data are created every day: emails, support tickets, chats, social media conversations, surveys, articles, documents, etc). But it's hard to analyze for sentiment in a timely and efficient manner. Sentiment analysis helps businesses quickly make sense of all their unstructured text by automatically understanding, processing, and tagging it, in a matter of minutes and with minimal human input.

Spam prevents the user from making full and good use of time, storage capacity and network bandwidth. The huge volume of spam mails flowing through the computer networks have destructive effects on the memory space of email servers, communication bandwidth, CPU power and user time. The menace of spam email is on the increase on yearly basis and is responsible for over 77% of the whole global email traffic. Users who receive spam emails that they did not request find it very irritating.

It is also resulted to untold financial loss to many users who have fallen victim of internet scams and other fraudulent practices of spammers who send emails pretending to be from reputable companies with the intention to persuade individuals to disclose sensitive personal information like passwords, Bank Verification Number (BVN) and credit card numbers.

1.3 NEED FOR THE NEW SYSTEM

The application gather all four feature into one place where user can use this application for many purpose so that they can be sure if the information they have gathered and can use this application and can be sure of the information and apply in their daily life and put their nervousness at ease.

A type of yellow journalism, fake news encapsulates pieces of news that may be hoaxes and is generally spread through social media and other online media. This is often done to further or impose certain ideas and is often achieved with political agendas. Such news items may contain false and/or exaggerated claims, and may end up being virialized by algorithms, and users may end up in a filter bubble. To stop all these things which divert the mind of the people from the goal news and reduce the rate of violence. By this system application helps to spread the real news and diminished the trolling of the inappropriate news around the society. Sentiment analysis helps businesses quickly make sense of all their unstructured text by automatically understanding, processing, and tagging it, in a matter of minutes and with minimal human input.

1.4 OBJECTIVE

The main objective is to display the application of the natural language processing using different algorithm and analysing the user data for themselves. The user can use these features upon their requirement. The objective of the project is also to reduce to the spreading of the false information around the internet and reduce the possibilities to prevent the news to be spread at that level that it causes a big dilemma. The other one is that any user or organization can use the sentiment analysis. Sentiment analysis helps businesses quickly make sense of all their unstructured text by automatically understanding, processing, and tagging it, in a matter of minutes and with minimal human input.

Due to the exponential growth of information online, it is becoming impossible to decipher the true from the false. Thus, this leads to the problem of fake news. To Determine the news that is trolling around is true or not. To Eliminate the false news which had been spread throughout the nation with an irrelevant subject and origin. To Stabilize people, trust in the sharing of the real news around the internet. Fake humans are not the only contributors to the dissemination of false information; real humans are very much active in the domain of fake news. As implied, trolls are real humans who “aim to disrupt online communities” in hopes of provoking social media users into an emotional response. To reduce the trolling and

fake environment we create a desktop-based application were the user can give the news that had been received by them and as an output the user can know whether the news is valid or not

1.5 MODULES OF THE SYSTEM

There are four modules and we will explain it one by one thoroughly. The modules are:

I)Fake News Detector:

The Modules deals with News detection process. The fake news module takes a input from the user and can conclude that the news that is provided is real or fake at some extent. The Fake news module uses machine learning model to predict or classify the news into its category. The Machine leaning uses pre-build data to train a model with a certain extent. The phrase states ‘within a certain extent ‘is being repeatably bring used since till date there is no model in Machine learning is being made with a 100% prediction value. That is why every machine learning model is used for reference model only.

Here the model is being under Multinomial algorithm which is used classification problems and is very effective for data containing text. It is known every data had to be converted into the numerical representation so that the classifier can be applied on the data to train the model and can be used for further usage.

In Linguistic Cue approaches, researchers detect deception through the study of different communicative behaviours. Researchers believe that liars and truthtellers have different ways of speaking. In text-based communication, deceivers tend to have a total word count greater than that of a truth-teller. Also, liars tend to use fewer self-oriented pronouns than other-oriented pronouns, along with using more sensory-based words. Hence, these properties found in the content of a message can serve as linguistic cues that can detect deception (Rubin, 2017). Essentially, Linguistic Cue approaches detect fake news by catching the information manipulators in the writing style of the news content. The main methods that have been implemented under the Linguistic Cue approaches are Data Representation, Deep Syntax, Semantic Analysis, and Sentiment Analysis. When dealing with the Data Representation approach, each word is a single significant unit and the individual words are analyzed to reveal linguistic cues of deception, such as parts of speech or location-based words.

The Data pre-processing is very crucial task and there are many ways to pre-process the data before applying for the model training. It is noted that a well pre-processed data will always produce a better performance. The method is used is TFIDF vectorizer which have two phases which are TF (Term Frequency) and IDF (Inverse Documentation Frequency). TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction. Let's take sample example and explore two different spicy sparse matrices before go into deep explanation.

$$IDF(w) = \log \frac{Total\ number\ of\ messages}{Total\ number\ of\ messages\ containing\ w}$$

Figure 1.1: The formula of IDF

Another Library from python is used named Spacy which is used to perform one of the methods for data pre-processing. The Function used is Stopwords. Stopwords are the words in any language which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on. The last method used is Stemming which reduce the sentence size so that the data can be processed in an efficient way.

II) Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing technique used to interpret and classify emotions in subjective data. Sentiment analysis is often performed on textual data to detect sentiment in emails, survey responses, social media data, and beyond. Sentiment analysis is the process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers.

Since customers express their thoughts and feelings more openly than ever before, sentiment analysis is becoming an essential tool to monitor and understand that sentiment. Automatically analyzing customer feedback, such as opinions in survey responses and social media conversations, allows brands to learn what makes customers happy or frustrated, so that they can tailor products and services to meet their customers' needs.

For example, using sentiment analysis to automatically analyze 4,000+ reviews about your product could help you discover if customers are happy about your pricing plans and customer service. Maybe you want to gauge brand sentiment on social media, in real time and over time, so you can detect disgruntled customers immediately and respond as soon as possible. Sentiment analysis models focus on polarity (positive, negative, neutral) but also on feelings and emotions (angry, happy, sad, etc), urgency (urgent, not urgent) and even intentions (interested v. not interested).

Depending on how you want to interpret customer feedback and queries, you can define and tailor your categories to meet your sentiment analysis needs.

It's estimated that 90% of the world's data is unstructured, in other words it's unorganized. Huge volumes of business data are created every day: emails, support tickets, chats, social media conversations, surveys, articles, documents, etc). But it's hard to analyze for sentiment in a timely and efficient manner.

Sentiment analysis helps businesses quickly make sense of all their unstructured text by automatically understanding, processing, and tagging it, in a matter of minutes and with minimal human input.

The overall benefits of sentiment analysis include:

- Sorting Data at Scale
  Can you imagine manually sorting through thousands of tweets, customer support conversations, or surveys? There's just too much business data to process manually. Sentiment analysis helps businesses process huge amounts of data in an efficient and cost-effective way.
- Real-Time Analysis
  Sentiment analysis can identify critical issues in real-time, for example is a PR crisis on social media escalating? Is an angry customer about to churn? Sentiment analysis models can help you immediately identify these kinds of situations, so you can take action right away.
- Consistent criteria
  It's estimated that people only agree around 60-65% of the time when determining the sentiment of a particular text. Tagging text by sentiment is highly subjective, influenced by personal experiences, thoughts, and beliefs. By using a centralized sentiment analysis system, companies can apply the same criteria to all of their data, helping them improve accuracy and gain better insights.

Natural Language Processing (NLP) and machine learning algorithms (basically rules) are the driving forces behind sentiment analysis.

There are different algorithms you can implement in sentiment analysis models, depending on how much data you need to analyze, and how accurate you need your model to be. We'll go over some of these in more detail, below.

Sentiment analysis algorithms fall into one of three buckets:

- Rule-based: these systems automatically perform sentiment analysis based on a set of manually crafted rules.
- Automatic: systems rely on machine learning techniques to learn from data.
- Hybrid systems combine both rule-based and automatic approaches.

III) Spam Classification

Email spam classification is now becoming a challenging area in the domain of text classification. Precise and robust classifiers are not only judged by classification accuracy but also by sensitivity (correctly classified legitimate emails) and specificity (correctly classified unsolicited emails) towards the accurate classification, captured by both false positive and false negative rates. This Module is used to solve the problem in the following manner

We are going to implement two techniques: Bag of words and TF-IDF. I shall explain them one by one. Let us first start off with Bag of words. Before starting with training, we must pre-process the messages. First of all, we shall make all the character lowercase. This is because 'free' and 'FREE' mean the same and we do not want to treat them as two different words. Then we tokenize each message in the dataset. Tokenization is the task of splitting up a message into pieces and throwing away the punctuation characters. For e.g.:

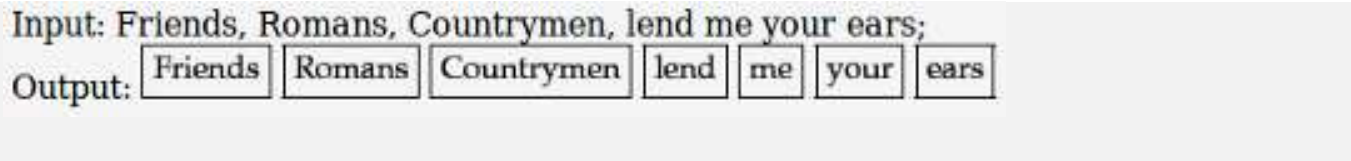


Figure 1.2: Expected Process of Data pre-processing

The words like ‘go’, ‘goes’, ‘going’ indicate the same activity. We can replace all these words by a single word ‘go’. This is called stemming. We are going to use Porter Stemmer, which is a famous stemming algorithm.

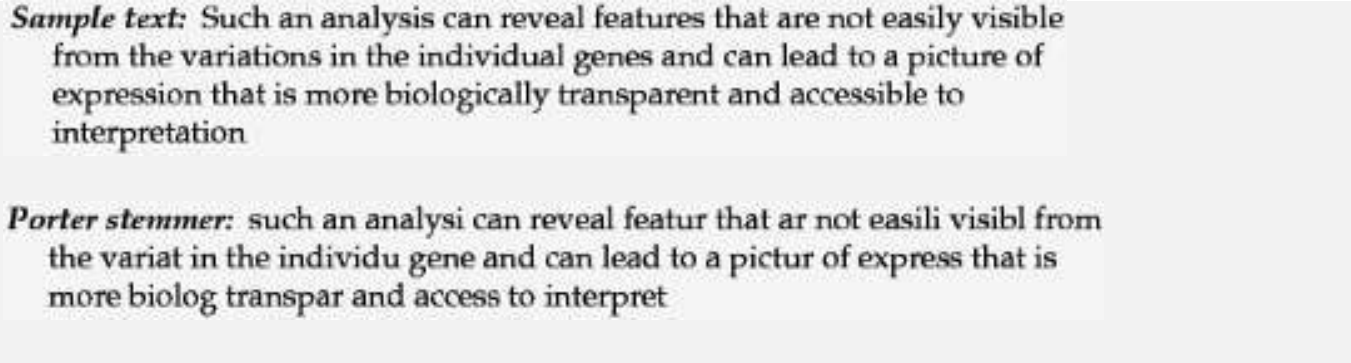


Figure 1.3: The output of Stemming process

We then move on to remove the stop words. Stop words are those words which occur extremely frequently in any text. For example, words like ‘the’, ‘a’, ‘an’, ‘is’, ‘to’ etc. These words do not give us any information about the content of the text. Thus, it should not matter if we remove these words for the text.

For classifying a given message, first we pre-process it. For each word w in the processed messaged we find a product of P(w|spam). If w does not exist in the train dataset, we take TF(w) as 0 and find P(w|spam) using above formula. We multiply this product with P(spam) The resultant product is the P(spam|message). Similarly, we find P(ham|message). Whichever probability among these two is greater, the corresponding tag (spam or ham) is assigned to the input message. Note than we are not dividing by P(w) as given in the formula. This is because both the numbers will be divided by that and it would not affect the comparison between the two.

IV) **Text Summarization**

Text summarization refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document. Automatic text summarization is a common problem in machine learning and natural language processing (NLP). With such a

big amount of data circulating in the digital space, there is need to develop machine learning algorithms that can automatically shorten longer texts and deliver accurate summaries that can fluently pass the intended messages.

Furthermore, applying text summarization reduces reading time, accelerates the process of researching for information, and increases the amount of information that can fit in an area. Usually, text summarization in NLP is treated as a supervised machine learning problem (where future outcomes are predicted based on provided data). Typically, here is how using the extraction-based approach to summarize texts can work:

1. Introduce a method to extract the merited key phrases from the source document. For example, you can use part-of-speech tagging, words sequences, or other linguistic patterns to identify the key phrases.

2. Gather text documents with positively-labelled key phrases. The key phrases should be compatible to the stipulated extraction technique. To increase accuracy, you can also create negatively-labelled key phrases.

3. Train a binary machine learning classifier to make the text summarization. Some of the features you can use include:

- Length of the key phrase
- Frequency of the key phrase
- The most recurring word in the key phrase
- Number of characters in the key phrase

4. Finally, in the test phrase, create all the keyphrase words and sentences and carry out classification for them.

1.6 SCOPE

NLP is one of the growing technologies. With constant innovation and research going on in this field, it is only expected to grow in the future. Since this is such an upcoming field, there is a dire need for skilled professionals. If you are interested in working on making computers learn and understand human language, then this is a good time to upskill yourself. NLP offers good prospects

With the exponential growth of multi-channel data like social or mobile data, businesses need solid technologies in place to assess and evaluate customer sentiments. So far, businesses have been happy analyzing customer actions, but in the current competitive climate, that type of customer analytics is outdated.

Now businesses need to analyze and understand customer attitudes, preferences, and even moods – all of which come under the purview of sentiment analytics. Without NLP, business owners would be seriously handicapped in conducting even the most basic sentiment analytics.

The Scope of the NLP and this project are just boundless as the research is been going at very high rate so that the result will be more accurate. The more accuracy will enhance the efficacy of the software as well. So the Project shows a strong vision for betterment in the future.

LITERATURE SURVEY

2.1 EXISTING SYSTEMS

1.Fake news Detector

There exists a large body of research on the topic of machine learning methods for deception detection, most of it has been focusing on classifying online reviews and publicly available social media posts. Particularly since late 2016 during the American Presidential election, the question of determining 'fake news' has also been the subject of particular attention within the literature.

Conroy, Rubin, and Chen outline several approaches that seem promising towards the aim of perfectly classifying the misleading articles. They note that simple content-related n-grams and shallow parts-of-speech (POS) tagging have proven insufficient for the classification task, often failing to account for important context information. Rather, these methods have been shown useful only in tandem with more complex methods of analysis. Deep Syntax analysis using Probabilistic Context Free Grammars (PCFG) have been shown to be particularly valuable in combination with n-gram methods. Feng, Banerjee, and Choi are able to achieve 85%-91% accuracy in deception related classification tasks using online review corpora.

Feng and Hirst implemented a semantic analysis looking at 'object: descriptor' pairs for contradictions with the text on top of Feng's initial deep syntax model for additional improvement. Rubin, Lukoianova and Tatiana analyze rhetorical structure using a vector space model with similar success. Ciampaglia et al. employ language pattern similarity networks requiring a pre-existing knowledge base.

Top Five Unreliable News Sources		Top Five Reliable News Sources	
Before It's News	2066	Reuters	3898
Zero Hedge	149	BBC	830
Raw Story	90	USA Today	824
Washington Examiner	79	Washington Post	820
Infowars	67	CNN	595

Figure 2.1: List of Unreliable and reliable sources for news

2.Sentiment Detector

The solution is discussed in the article by Mudinas et al. [MZL12]: a concept-level sentiment analysis system called pSenti which combines lexicon based and learning based approaches. It measures and reports the overall sentiment of a review through a score that can be positive, negative or neutral or 1–5 stars classification. The main advantages and main interests of this article are the lexicon/learning symbiosis, the detection and measurement of sentiments at the concept level and the lesser sensitivity to changes in topic domain.

It operates in four parts. First, the pre-processing of the review where the noise (idioms and emoticons) is removed and each word is tagged and stored by the method Part Of Speech (POS). Second, the aspects and views are extracted to generate a list of top 100 aspect groups and top 100 views. The aspects are identified as nouns and noun phrases, and the views as sentiment words, adjectives and known sentiment words which occur near an aspect. Then the lexicon-based approach is used to give a "sentiment value" to any sentiment word and generates features for the supervised machine learning algorithm. Finally, this algorithm generates a "feature vector" for each aspect which is either the sum of the sentiment value for a sentiment word or the number of occurrences of this word in relation with other adjectives.

To evaluate this method, experiments were conducted on two datasets: software reviews (more than 10,000) and movie reviews (7,000). Software reviews were separated into two categories: software editor reviews and customer software reviews. As a result, pSenti's accuracy was proved close to the pure learning-based system and higher than the pure lexicon-based method. It was also shown that the performance was not as good on customer software reviews as on software editor reviews because customer software reviews are usually much "noisier" (with comments that are irrelevant for the subject) than professional software editor reviews. Its accuracy was also affected by a large number of reviews for which it failed to detect any sentiment or assigned neutral score. However, the A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation sentiment separability in movie reviews was much lower than in software reviews. One of the reasons is that many movie reviews contain plot description and many quotes from the movie where words are identified as sentiments by the system.

3.Spam Detector

Various techniques have been explored to relieve the problem of email spams. Used, previous works on spam detection can be generally classified based on features of e-mails. It can be classified into three categories:

1) content-based methods,

2) non content-based methods, and

3) others. Initially, researchers model this problem as a binary text classification task by analyzing email content text.

To relieve the spam problem various techniques have been explored. Spam detection can be classified in previous works based on the characters of email. Naive Bayes and Support Vector Machines are the representatives of this category. Generally Naive Bayes methods train a probability model using classified emails, and each word in emails will be given a probability of being a suspicious spam keyword. As for SVMs, it is a supervised learning method that has outstanding performance on text classification tasks.

Traditional SVMs and improved SVMs have been investigated. The excellent results with static data sets are reported by the conventional machine learning techniques. But it has some disadvantage that it is cost-prohibitive for large-scale applications Thus the latest information to adapt to the rapid evolving nature of spams. The spam detection of these methods on the email corpus with various languages has been less studied yet. Many other classifications are also account for spam detection. They are markov random field model, neural network and logic regression, and certain specific features, such as URLs and images

4.Summarizer

Shah et al says Automatic text summarization of Wikipedia articles is difficult to detect subtopic in documents. There are two new approaches for summarizing the text. The first method is to adjust the frequency of the words based on the root form of the word, and also the frequency of its synonyms presents in the text. The second method is to identify sentences containing citations or references and give them a higher weight. The advantage of this approach is effective sentence ranking in summary. The disadvantage of this approach is use of citations with higher weight to sentence so unimportant information is added in summary.

Jain et al says the limitation of the approach is dealing with problems of information redundancy, sentence ordering and fluency. Graph and Cluster Based approaches in Multi Document summarization and gives the idea to improve summary in less effort or even to construct new or hybrid procedure for next generation. The advantage of this approach is to generate smooth summaries as compared to ranking algorithms. The disadvantage of this approach is information loss during summarization.

Atefeh Ferdosipour says the effectiveness of sentence scoring method depends upon length of document and the type of language used in document. Cohesion approach is grammatical and lexical linking within a text or sentence that holds a text together and gives it meaning for increasing effectiveness of summary. Cohesion devices that create coherence in text are as Reference, Substitution, Elipse, Lexical cohesion and Conjunction. Grammatical Cohesion is referring to structural content. Lexical Cohesion is referring to the language content of a piece. The advantage of this approach is to use a hybrid approach. The disadvantage of this approach is varying the document length

2.2 PROPOSED SOLUTION

The proposed solution involves the use of a tool that is designed with the specific aim of detecting and eliminating web pages that contain misinformation intended to mislead readers. For purposes of attaining this goal, the approach will utilize some factors as a guide to making the decision as to whether to categorize a web page as fake news. The user will, however, need to have the tool downloaded and installed on a personal computer before making use of its services. It is expected that the proposed method will be compatible with the browsers that are commonly used by users all over the world. The syntactical structure of the links used to lead users to such sites will be considered a starting point. For instance, when a user keys in a group of search terms with the aim of finding web pages that contain information related to the same terms, the tool will come into operation and run through the sites that have been retrieved by the search engine before they are delivered to the user. In doing so, the extension will identify sites whose links contain words that may have a misleading effect on the reader, including those that are characterized by a lot of hyperbole and slang phrases. Such web pages will be flagged as being potential sources of fake news, and the user will be notified before electing to click on either one of them. A visualization of the links and their syntactical structure will help the user understand the decision

- We create a Web-based application where the user can give the article that had been received by them and as an output the user can know whether the article is valid or not, or detecting sentiment of the news, or can check whether the article is spam or not, and at last summarizing the article. The Web based application will be using a machine learning model to carry out all above operations.

- For Fake news Detection we first create the article in the form of data frames the next part is applying TFIDF vectorization which will count the frequency of each word and then creating the accuracy matrix to predict the accuracy score, next step is applying Naive Bayes Classifier to predict news belong to which class, the last step is applying Passive Aggressive Classifier, Passive Aggressive algorithms in this model gives us a 93.33% accurate result. It is an online learning algorithm. Such an algorithm remains passive for a correct classification outcome, and turns aggressive in the event of a miscalculation, updating and adjusting. Unlike most other algorithms, it does not converge. Its purpose is to make updates that correct the loss, causing very little change in the norm of the weight vector.

- For Sentiment Analysis first we create a bag of words, it is a function in which a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. The next part is applying TFIDF VECTORIZATION here the frequency of each word is counted and stored. The next part is word embedding. Word embedding is any of a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers. Last part is applying Natural Language Processing to Extract the sentiment of the string.

- For Spam Classifier we first create the article in the form of data frames the next part is applying TFIDF vectorization which will count the frequency of each word and then creating the accuracy matrix to predict the accuracy score, next step is applying Naive Bayes Classifier to predict news belong to which class, Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. The last is applying multinomial classification to predict whether the article is spam or not, the model gives the accuracy of 83-86%.

- For Summarizer we create a bag of words, it is a function in which a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity. The next step is applying the summarize function to get the summary of the article.

2.3 FEASIBILITY STUDY

- It includes consideration of all possible ways to provide solution to a given problem.
- The proposed solution should satisfy all the user requirements and should be flexible enough so that future changes can be easily done based on future upcoming requirements

Economical Feasibility:

1.This is a very important aspect to be considered while developing a project, we decided the technology based on minimum cost factory.

2.All hardware and software cost have to be borne by the organization.

Technical Feasibility:

1.This includes study of function, performance and constrains that may affect the ability to achieve the acceptance system.

2.For this feasibility study, we studied complete functionality to be provided in system and check if eve-rything as possible using different types of front-end and back-end

Operational Feasibility:

1.No doubt the proposed system is fully GUI based that is very user friendly and all inputs to be taken all self-explanatory.

2.Besides, a proper training has been conducted to let know the essence of the system to the user so that they can feel comfortable with the new systems.

3.As far as our study is concerned the clients are comfortable and happy as the system has cut down their loads and doing.

REQUIREMENTS ANALYSIS

3.1 METHOD USED FOR REQUIREMENT ANALYSIS

Requirements Analysis is the process of defining the expectations of the users for an application that is to be built or modified. It involves all the tasks that are conducted to identify the needs of different stake-holders. Therefore, requirements analysis means to analyze, document, validate and manage software or system requirements.

The software requirements analysis process involves the following steps/phases:

1. Eliciting requirements
2. Analyzing requirements
3. Requirements modeling
4. Review and retrospective

1- Eliciting requirements

The process of gathering requirements by communicating with the customers is known as eliciting requirements. In this process of requirement analysis to understand what the customer want in software and communicating with the customer to gather information and understand the requirements that are required by the customer.

In our software we collect all the information and understand the requirement of the user by gathering the important requirements or needs.

2- Analyzing requirements

This step helps to determine the quality of the requirements. It involves identifying whether the requirements are unclear, incomplete, ambiguous, and contradictory. These issues resolved before moving to the next step. After the first process requirement gathering, the next step is to understand all the requirements that they are complete, clear and easily understandable or anything that find unclear, ambiguous or contradictory so that can be find or solve in these steps before moving to the next step.

Analyzing requirement is a essential part of any requirement analysis, the analysis of all the requirement is needed for understanding the actual requirements.

3- Requirements modeling

In Requirements modeling, the requirements are usually documented in different formats such as use cases, user stories, natural-language documents, or process specification. Requirement modeling plays an important role in requirement analysis. In this process requirements are documented in various forms like use cases, diagrams, user stories to understand the actual view or goal or functions of any software system.

4- Review and retrospective

This step is conducted to reflect on the previous iterations of requirements gathering in a bid to make improvements in the process going forward.
This is the last step in this step, team members reflect on what happened in the iteration and identifies actions for improvement going forward.

Requirements analysis is a team effort that demands a combination of hardware, software and human factors engineering expertise as well as skills in dealing with people. Here are the main activities involve in requirement analysis:

- Identify customer's needs.
- Evaluate system for feasibility.
- Perform economic and technical analysis.
- Allocate functions to system elements.
- Establish schedule and constraints.
- Create system definitions.

For our software system to gather information or requirement analysis we completely follow the above process to understand the requirements of the user this can play a important role in any software development process.

3.2 DATA REQUIREMENTS

Data requirements definition establishes the process used to identify, prioritize, precisely formulate, and validate the data needed to achieve business objectives. When documenting data requirements, data should be referenced in business language, reusing approved standard business terms if available. Data requirements means to collectively gather all the requirements that are useful and try to analyze how to achieve that all the requirements to make it possible in our software system. In our software system or application there is collected data requirements to know or understand how to achieve or approach all the requirements gathered from the user to make it implemented to solve the real-world problems. Data requirements also

helps us in understanding the required business goals that are essentially useful for proceeding to start further processes.

3.3 FUNCTIONAL REQUIREMENTS

Functional requirements define the basic system behavior. Functional requirements usually define if/then behaviors and include calculations, data input, and business processes. Functional requirements are features that allow the system to function as it was intended. Functional requirements are product features and focus on user requirements. A Functional Requirement (FR) is a description of the service that the software must offer. It describes a software system or its component. A function is nothing but inputs to the software system, its behavior, and outputs. It can be a calculation, data manipulation, business process, user interaction, or any other specific functionality which defines what function a system is likely to perform. Functional Requirements are also called Functional Specification**.**

User module –

- User need to provide information which he/she wants to know that this information or news is fake or genuine, or if it is genuine than know the sentiment of that news by knowing the nature that it is positive, negative or neutral.
- And the user retrieves all the required information.

Admin –

- Admin do operation to know that the information or news given by user is genuine or fake.
- If it is genuine than admin finds that the sentiments of news or information is positive, negative or zero.

Benefits of Functional Requirement-

Here, are the pros/advantages of creating a typical functional requirement document-

- Helps you to check whether the application is providing all the functionalities that were mentioned in the functional requirement of that application.
- A functional requirement document helps you to define the functionality of a system or one of its subsystems.
- Functional requirements along with requirement analysis help identify missing requirements. They help clearly define the expected system service and behavior.

- Errors caught in the Functional requirement gathering stage are the cheapest to fix.
- Support user goals, tasks, or activities

3.4 NON-FUNCTIONAL REQUIREMENTS

Non-functional requirements specify the quality attribute of a software system. They judge the software system based on responsiveness, usability, security, portability and other non-functional standards that are critical to the success of the software system. Usability – These web-based applications has appropriate and adequate information to guide the user in order to use the application.

Non-functional Requirements allows you to impose constraints or restrictions on the design of the system across the various agile backlogs. Example, the site should load in 3 seconds when the number of simultaneous users is > 10000. Description of non-functional requirements is just as critical as a functional requirement.

- Portability – This is portable as it is online running web-based application across the internet.
- Flexibility – It is very flexible in nature.
- Security – This is secured in various aspects to use the application.
- Maintainability – Maintenance is done in an efficient way that the data should secure and easily retrieve.
- Scalability – These applications can be further modified in future.

Advantages of Non-Functional Requirement

- The non-functional requirements ensure the software system follow legal and compliance rules.
- They ensure the reliability, availability, and performance of the software system.
- They ensure good user experience and ease of operating the software.
- They help in formulating security policy of the software system.

3.5 SYSTEM SPECIFICATION

Technology can be most broadly defined as the entities in the form of software and hardware. It created by the application of mental and physical effort in order to achieve some value. In this

usage, technology refers to tools and machines that may be used to solve real-world problems. For this purpose, some software and hardware are required. The required ones are as follows.

3.5.1 Hardware Specification

- Ram: 2GB (minimum)
- Storage: 150GB
- Processor: Intel Core i3(minimum) or equivalent

3.5.2 Software Specification

Software is a set of instructions, data or programs used to operate computers and execute specific tasks. To performing this task following some specific software and technology are used for developing this application.

- Web browser: - A browser is software that is used to access the internet. A browser lets you visit websites and do activities within them such as post your selected news which you want to check.
- Internet access: - It is must require to perform this task.
- Python: -Python is a programming language. It is used on a server to create web applications.
- Flask: -Flask is a web framework. This means flask provides you with tools, libraries, and technologies that allow you to build a web application. This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.
- Machine Learning:-A subset of artificial intelligence (AI), machine learning (ML) is the area of computational science that focuses on analyzing and interpreting patterns and structures in data to enable learning, reasoning, and decision making outside of human interaction. Simply put, machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations and decisions based on only the input data.
- Data Science: -Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.
- HTML, CSS and Java Script: - HTML (Hyper Text Markup Language) is the most basic building block of the Web. It defines the meaning and structure of web content. Other technologies besides

DESIGN

4.1 SOFTWARE REQUIREMENT SPECIFICATION:

A software requirements specification (SRS) is a comprehensive description of the intended purpose and environment for software under development. The SRS fully describes what the software will do and how it will be expected to perform.

An SRS minimizes the time and effort required by developers to achieve desired goals and also minimizes the development cost. A good SRS defines how an application will interact with system hardware, other programs and human users in a wide variety of real-world situations. Parameters such as operating speed, response time, availability, portability, maintainability, footprint, security and speed of recovery from adverse events are evaluated. Methods of defining an SRS are described by the IEEE (Institute of Electrical and Electronics Engineers) specification 830-1998.

Using the SRS helps an enterprise confirm that the requirements are fulfilled and helps business leaders make decisions about the lifecycle of their product, such as when to retire a feature.

4.1.1 Glossary

- **Pandas:**

  Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

- **SKlearn:**

  scikit-learn is a Python module for machine learning built on top of SciPy and is distributed under the 3-Clause BSD license.

- **Matplotlib:**

  Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits.

- **Passive Aggressive Classifier:**

  Passive Aggressive Algorithms are a family of online learning algorithms (for both classification and regression) proposed by Crammer at al. The idea is very simple and their performance has been

proofed to be superior to many other alternative methods like Online Perceptron and MIRA (see the original paper in the reference section).

- **Confusion Matrix**

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

- **Count Vectorization:**

Count Vectorization involves counting the number of occurrences each word appears in a document (i.e. distinct text such as an article, book, even a paragraph!).

- **TFID vectorization:**

In information retrieval, tf–idf or TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

- **Hash Vectorization:**

In machine learning, feature hashing, also known as the hashing trick (by analogy to the kernel trick), is a fast and space-efficient way of vectorizing features, i.e. turning arbitrary features into indices in a vector or matrix.

- **Numpy:**

NumPy is the fundamental package for scientific computing with Python. It contains among other things: useful linear algebra, Fourier transform, and random number capabilities. Besides its obvious scientific uses,

- **Flask:**

Flask is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. It began as a simple wrapper around Werkzeug and Jinja and has become one of the most popular Python web application frameworks.

- **HTML:**

It stands for HyperText Markup Language. Hypertext means that the document contains links that allow the reader to jump to other places in the document or to another document altogether.

- **CSS:**

Cascading Style Sheets (CSS) is a stylesheet language used to describe the presentation of a document written in HTML or XML (including XML dialects such as SVG, MathML or XHTML). CSS describes how elements should be rendered on screen, on paper, in speech, or on other media.

- **JavaScript:**

JavaScript attempts to convert the string numeric literal to a Number type value. First, a mathematical value is derived from the string numeric literal. Next, this value is rounded to the nearest Number type value.

- **Machine Learning:**

Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model.

- **MutinomailNB**

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts.

- **Spam**

irrelevant or unsolicited messages sent over the internet, typically to a large number of users, for the purposes of advertising, phishing, spreading malware, etc.

- **Bag of words**

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its **words**, disregarding grammar and even word order but keeping multiplicity.

- **Stemming**

Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers.

- **Lemmatization**

Lemmatization is the process of grouping together the different inflected forms of a word so they can be analysed as a single item. Lemmatization is similar to stemming but it brings context to the words. So, it links words with similar meaning to one word.

- **Sparse matrix**

parse matrix is a **matrix** which contains very few non-zero elements. When a sparse matrix is represented with a 2-dimensional array, we waste a lot of space to represent that matrix.

- **Confusion matrix**

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

- **Spacy**

Spacy is a free, open-source library for advanced Natural Language Processing (NLP) in Python.

- **Pickle**

Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickl**e** does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.)

- **Genism**

Gensim is an open-source library for unsupervised topic modeling and natural language processing, using modern statistical machine learning. Gensim is implemented in Python and Cython.

- **Textblob**

TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving

into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

- **nltk**

The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

4.1.2 Supplementary Specification:

- It is a web-based application, portable and can work on various operating systems.
- The system is reliable and atomic as it provides true information.
- We are required to tie up with top news sources to get the correct news information.
- In addition, the system requires various servers to store news information.
- The result predicted is based on a machine learning model which does provide accuracy but may be not true in some cases.
- The system predicts the percentage by which the news is true and, on that percentage, it is decided that news is true or false.
- The system also gives the sentiment analysis of the information, by the showing the nature of the information i.e. positive, negative or neutral.
- The system also gives the result for checking the spam activity and for getting the overall summary of the information or news with the help of spam classifier and summarizer.

4.1.3 Use Case Model

Description:

Enter website, enter news, press button to validate news result

Text Description:

I.U1-Enter Website

Using this the user get access to website

1.Scenario-Main line sequence

I. User-enter the right domain name or URL

ii. system-Click to enter the website

II.U2-Enter News

Using this the user can enter the news

1.Scenario-Main line Sequence

i.System-Enter the news in given Area

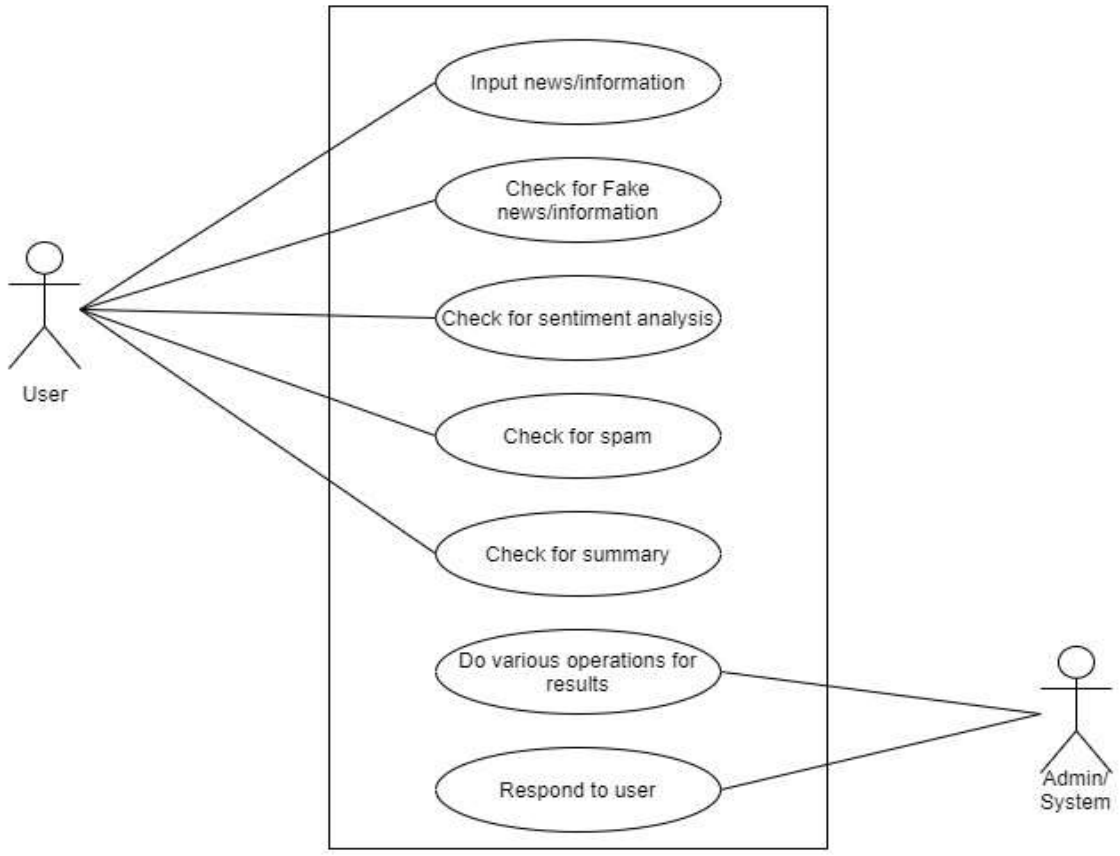


Figure 4.1: Use Case Diagram

III.U3-Press button to check news is fake/genuine or to check sentiment or to check for spam or to check for summary.

1.Scenario-Main line sequence

i.user -Enters the news

ii.System-Click on button to proceed

IV.U4: -Result

1.Main line sequence

i.System-The result is ready for Fake/Genuine news or positive/negative/neutral sentiments or spam or for the summary of the news.

4.2CONCEPTUAL LEVEL CLASS DIAGRAM

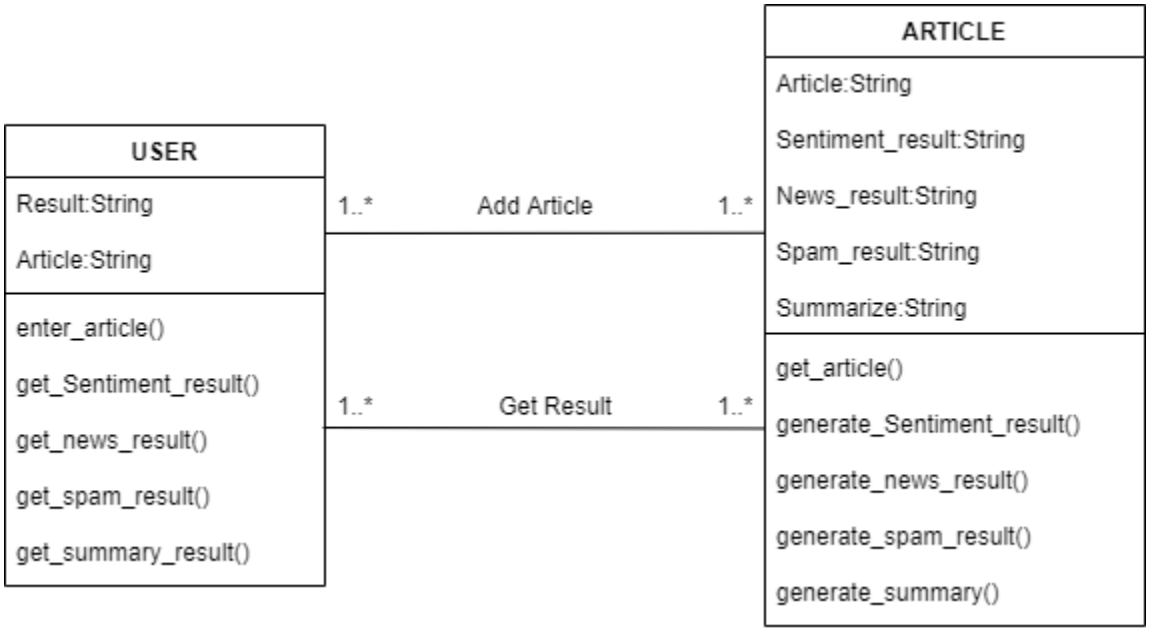


Figure 4.2: Conceptual Level Diagram

1.Class user

There are following methods and attributes in class user

I Attributes-

a.Article-Contain the article to be validated, spam classified, summarized or sentiment identified

b.result-Stores any of the one result as only on result is generated at the time

II Methods-

a.enter_article() - Method allowing system to get article from user

b.get_sentiment_result() - Method allowing user to get result of sentiment analysis

c.get_news_result() - Method allowing user to get result of news analysis

e.get_spam_result() - Method allowing user to get result of spam analysis

f.get_summary_result() - Method allowing user to get result of summary of article

2 Article

I Attributes-

a.Article-Contain the article to be validated, spam classified, summarized or sentiment identified

b.Sentiment_result-String Containing sentiment analysis result.

c.News_result-String Containing Fake news result.

d.Spam_result-String Containing Spam analysis result

e.Summarize-String Containing Summary of Article

II Methods-

a.get_article() - Method allowing system to fetch article entered by user.

b.generate_sentiment_result() - Method allowing user to generate result using sentiment analysis

c.generate_news_result() - Method allowing user to generate result using Fake news analysis

e.generate_spam_result() - Method allowing user to generate result using spam analysis

f.generate_summary_result() - Method allowing user to generate summary of article

4.3 CONCEPTUAL LEVEL ACTIVITY DIAGRAM-

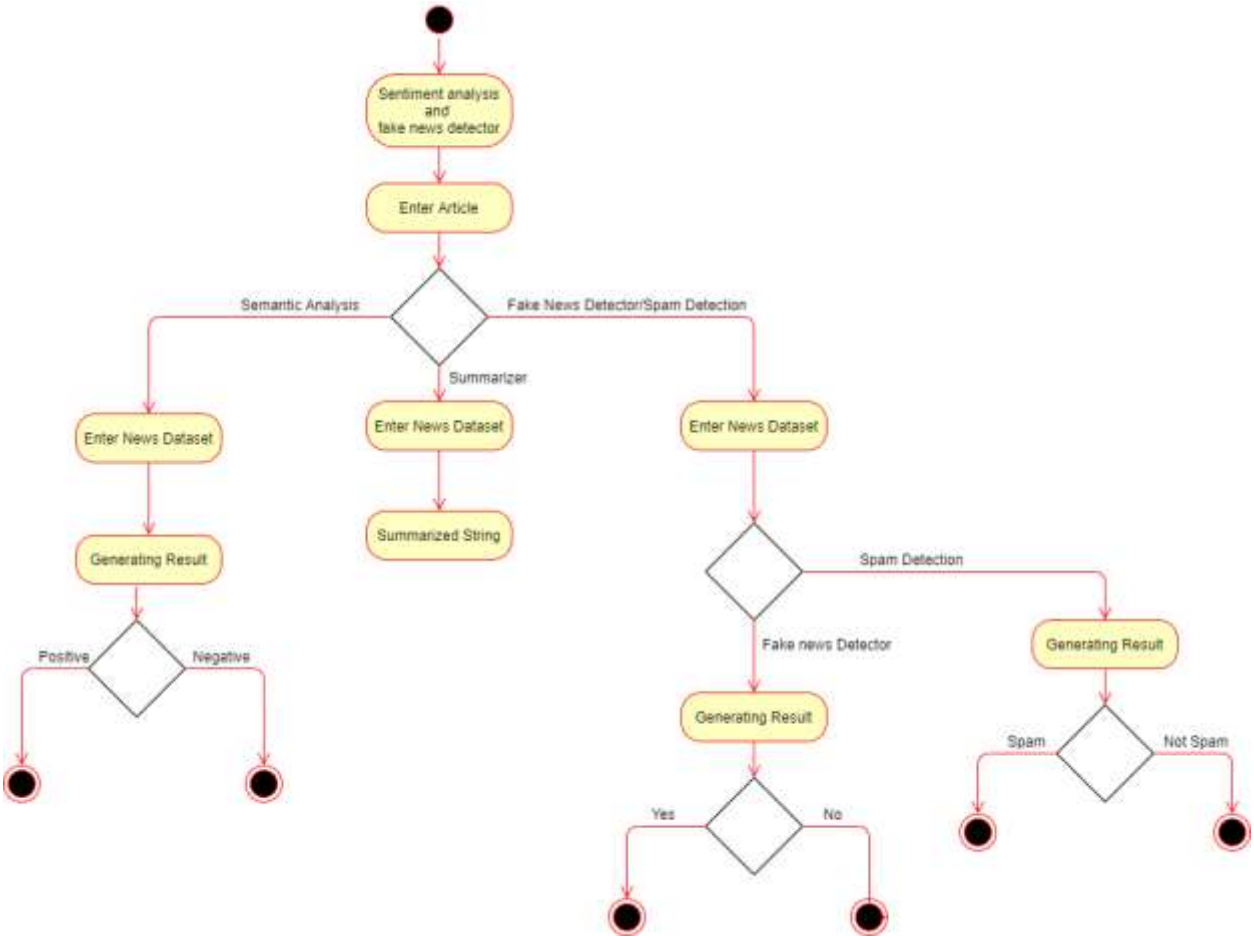


Figure 4.2: Conceptual Level Activity Diagram

Description

1.The first activity is visiting the website.

2.The second activity is Enter Article Here users enter the news article.

3.In the third activity the user has to select the activity he wants to perform out of four activities listed as Sentiment Detection, Summarizer, Fake News Detector and Spam Detector.

4.If the User chooses the activity Sentiment Detection Then following are the next activities performed-

4.1. First activity is converting the article into the form of a dataset.

4.2. The fifth activity is generating results out of the analysis.

4.3. If last step is positive means that sentiment is positive else sentiment is negative.

5.If the user chooses the activity summarizer the following activities are performed-

5.1. First activity is converting the article into the form of a dataset.

5.2. The last activity is generating the summary out of analysis.

6.If the person chooses activities Fake news detector or Spam detector following are the common activities involved in both-

6.1. The first activity in both activities is converting the article in the form of a data frame that is a dataset.

6.2. If the person has chosen Fake new detector here are some other activities involved-

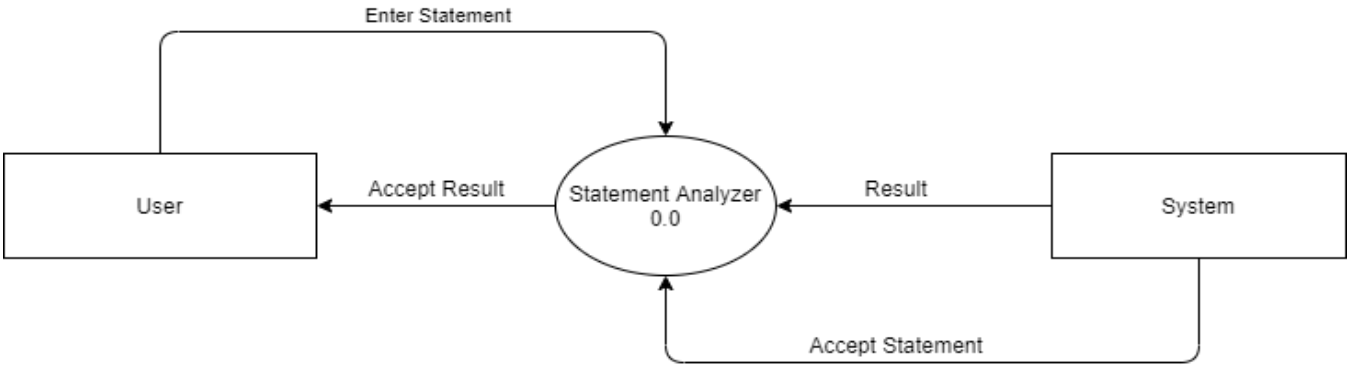
6.2.1. The next activity is generating the result from the analysis, if the condition is yes it means the news is true, else if condition is no it means the news is fake.

6.3. If the person has chosen Spam Detector here are some other activities involved-

6.3.1. The next activity is generating the result from the analysis, if the condition is yes it means the article is Spam, else if condition is no it means the article is not Spam.

4.4 DATA FLOW DIAGRAM (LEVEL 0,1,2)

Level 0



Level 0

Figure 4.3: DFD level 0

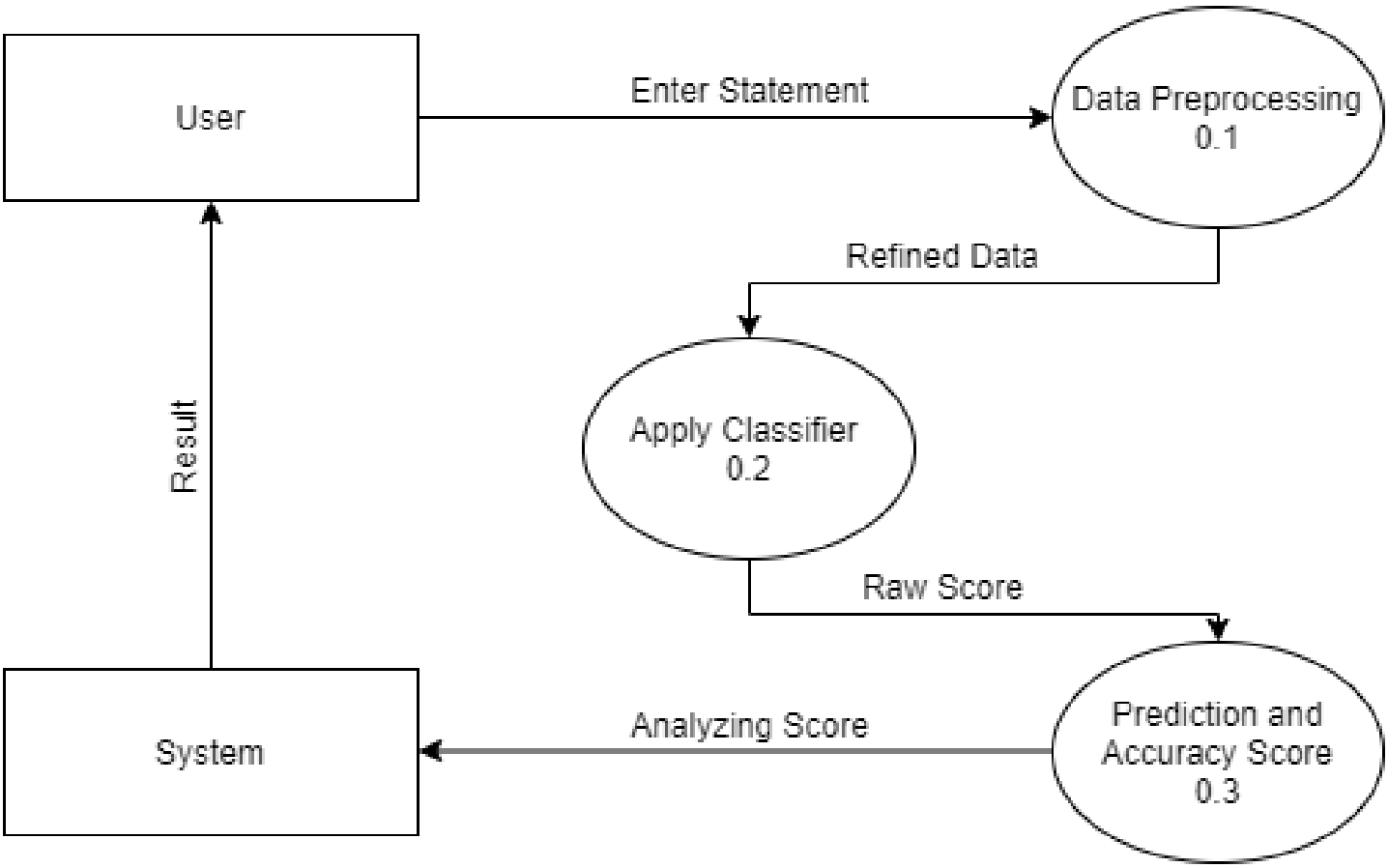
This is a most simple representation of working of the software. The Level 0 Data flow diagram represent the simplest working of the system. The center oval represents the Process which is indicating the Statement analyzer. As it for the rectangular block it is represented as external entities who will give input and will receive output to. The arrow represents the flow of the data where the arrow represents the giving of the data in the direction the arrow is pointing.

So, the after the basic terminology the user first send the input to the process as a text format and statement analyzer pass it to the system and the system return back the result to the process and also accept the statement and give the statement result to the process. Then the process passes the result to the user in the GUI (Graphical User Interface).

Level 1

The Level 1 Data flow diagram is the further representation of the software. Here the single process is further breakdown to easily understood by another non-technical person. So, the statement analyzer is divided into three oval which are in oval that are also process. The User is sending the text data to the data pre- processing area which will pre-process the data and will give a refined data to the next process that is apply classifier. The Apply classifier is going to apply the machine learning algorithm to train the model and after that it will give raw score to next process which is prediction and accuracy score. This process will have the value of the result in the numerical form and convert into required result and finally the result

will be provided to the user which asked for it. The analyzed score will go the system where it will generate the result and will provide it to the User.



Level-1

Figure 4.4: DFD level 1 Diagram

Level 2

Level 2 Data Flow Diagram is further breakdown of the level 1 Data flow diagram. The data preprocessing process is further breakdown so the process under data pre-processing are as follows, the first is the process of stemming and lemmatization is for reducing the frequency of similar words then after removing the similar words the data is sent to the Stop keywords where the process stop or eliminates the unnecessary words which are of no use in training the model. After that we will calculate the term frequency and after that IDF is calculated after that at last the TFIDF formula is used to calculated then is forwarded for the further process.

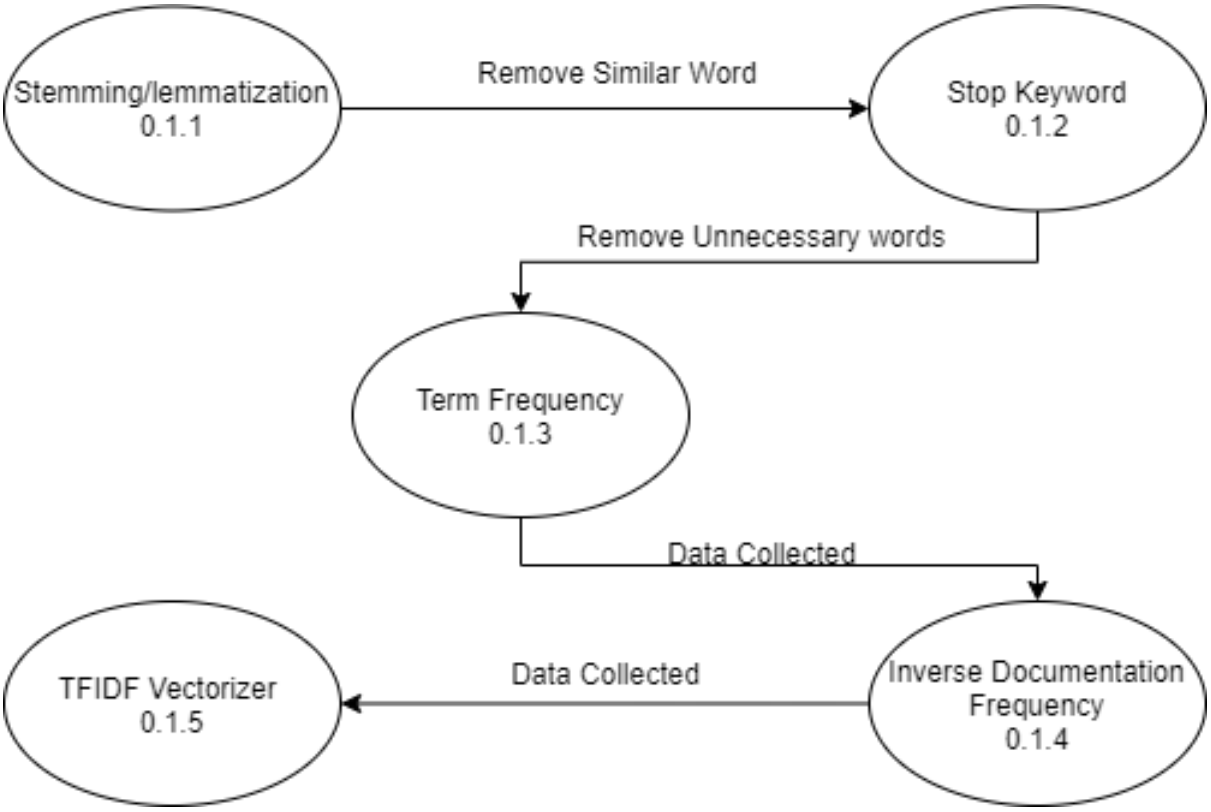


Figure 4.5: DFD level 2 Diagram

4.5 DATABASE DESIGN ER DIAGRAM

DESCRIPTION

1.The first entity is User which contains the attributes result and article

2.The Result attributes contain Sub attribute accuracy percentage

3.The Article attribute contains Sub attributes such as URL, language, text and time.

4.The relation between User and System is one to many. As there can be many users at one time

5.The System Contains the entities News Analysis and Sentiment analysis

6.The News analysis entity contains attributes sports, educational, entertainment and news results which contain the attribute accuracy percentage.

7.The sentiment analysis contain attributes article and sentiment result

8.The sentiment result contains the attribute positive, negative and neutral.

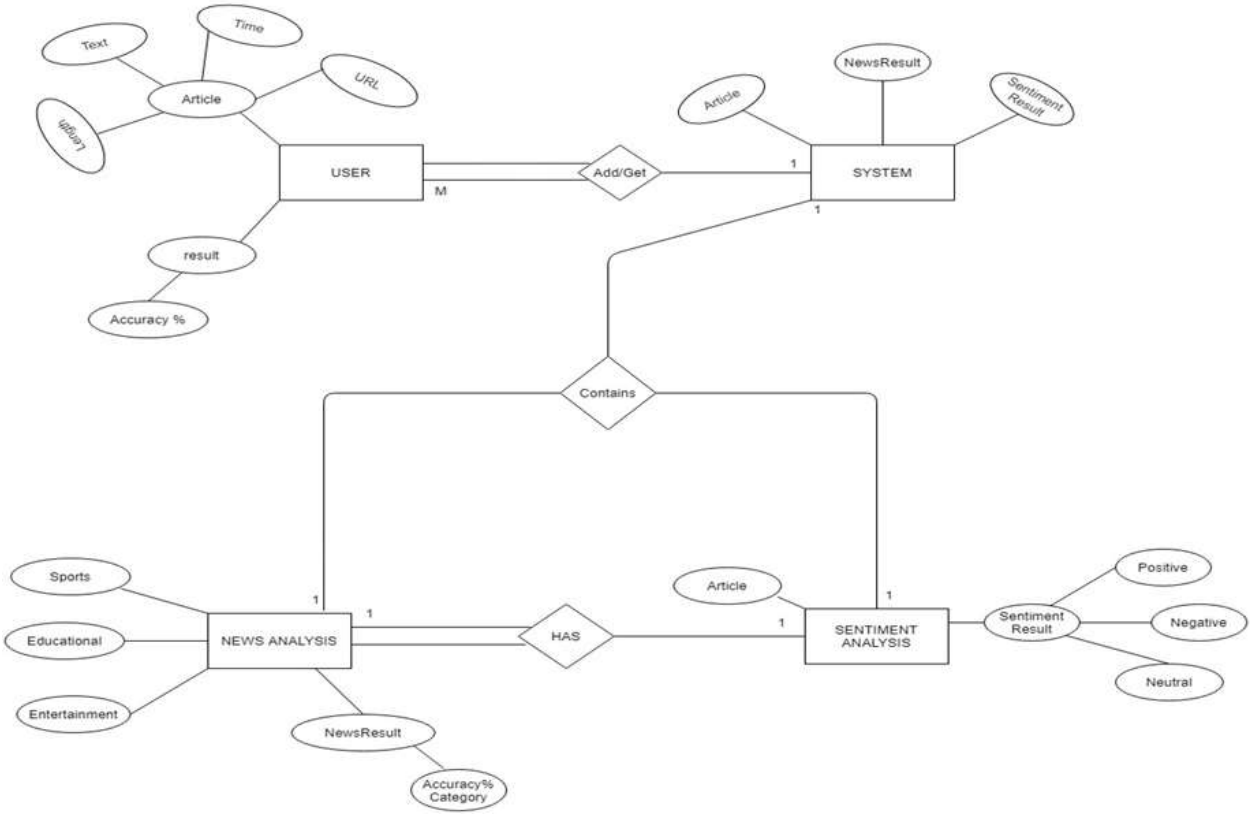


Figure 4.6: ER Diagram

9.The relation between News analysis and Sentiment analysis is that News analysis has Sentiment analysis

10.There is one to one connectivity between news analysis and Sentiment analysis.

11.News analysis has total participation whereas the Sentiment analysis has partial participation.

SYSTEM MODELING

5.1 DETAILED CLASS DIAGRAM

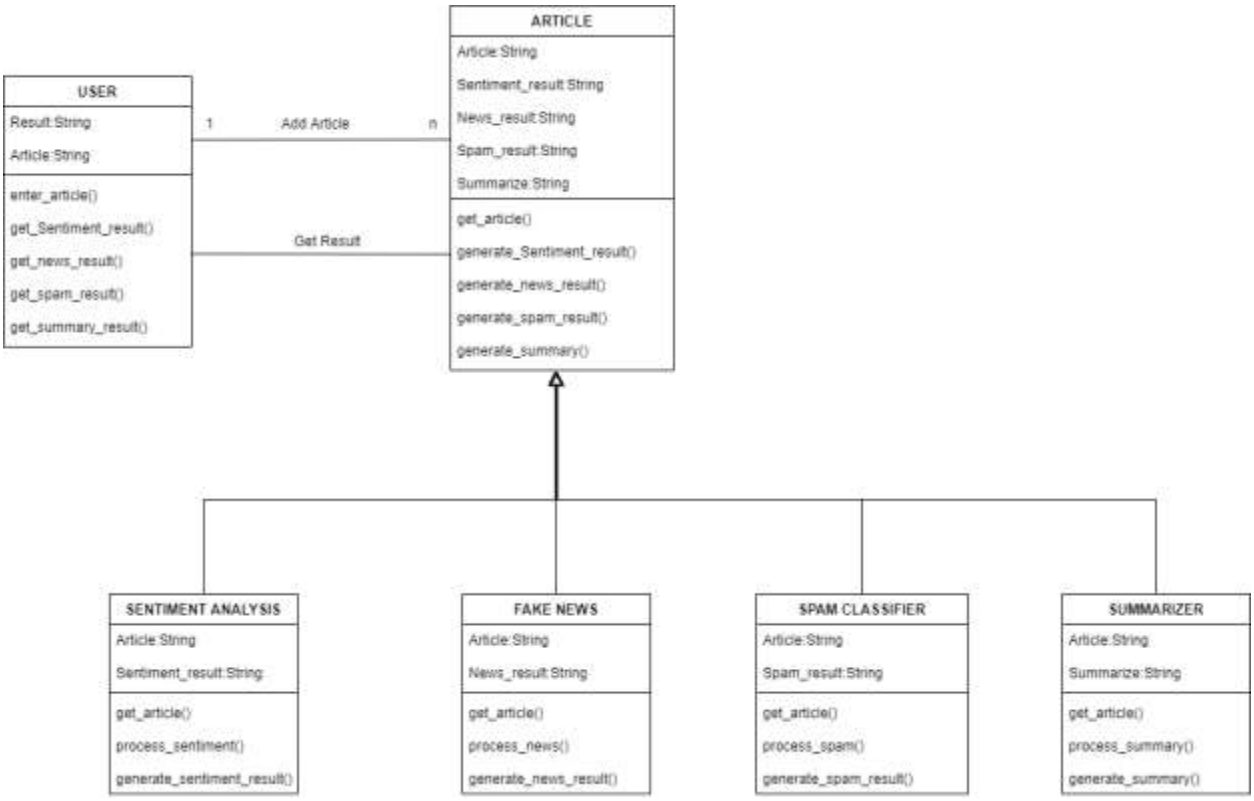


Figure 5.1: Detailed Class Diagram

Description for detailed class diagram

1.Class user

There are following methods and attributes in class user

I Attributes-

a.Article-Contain the article to be validated, spam classified, summarized or sentiment identified

b.result-Stores any of the one result as only on result is generated at the time

II Methods-

a.enter_article() - Method allowing system to get article from user

b.get_sentiment_result() - Method allowing user to get result of sentiment analysis

c.get_news_result() - Method allowing user to get result of news analysis

e.get_spam_result() - Method allowing user to get result of spam analysis

f.get_summary_result() - Method allowing user to get result of summary of article

2 Article

I Attributes-

a.Article-Contain the article to be validated, spam classified, summarized or sentiment identified

b.Sentiment_result-String Containing sentiment analysis result.

c.News_result-String Containing Fake news result.

d.Spam_result-String Containing Spam analysis result

e.Summarize-String Containing Summary of Article

II Methods-

a.get_article() - Method allowing system to fetch article entered by user.

b.generate_sentiment_result() - Method allowing system to generate result using sentiment analysis

c.generate_news_result() - Method allowing system to generate result using Fake news analysis

e.generate_spam_result() - Method allowing system to generate result using spam analysis

f.generate_summary_result() - Method allowing system to generate summary of article

2.1 Sentiment Analysis

I Attributes-

a.Article-Contain the article to be validated, spam classified, summarized or sentiment identified

b.Sentiment_result-String Containing sentiment analysis result.

II Methods-

a.get_article() - Method allowing system to fetch articles entered by user.

b.process_sentiment()-Method allowing the system to extract type of sentiment from analysis of articles.

c.generate_sentiment_result() - Method allowing system to generate results using sentiment analysis.

2.2 News Analysis

I Attributes-

a.Article-Contain the article to be validated, spam classified, summarized or sentiment identified

b.News_result-String Containing Fake news analysis result.

II Methods-

a.get_article() - Method allowing system to fetch articles entered by user.

b.process_news()-Method allowing the system to extract type of news from analysis of articles.

c.generate_news_result() - Method allowing system to generate results using Fake news analysis.

2.3 Spam Analysis

I Attributes-

a.Article-Contain the article to be validated, spam classified, summarized or sentiment identified

b.Spam_result-String Containing Spam analysis result .

II Methods-

a.get_article() - Method allowing system to fetch articles entered by user.

b.process_article()-Method allowing the system to extract whether the article is spam or not from analysis of articles.

c.generate_spam_result() - Method allowing system to generate result using spam analysis

2.4 Summarizer

I Attributes-

a.Article-Contain the article to be validated, spam classified, summarized or sentiment identified

b.Summarize-String Containing Summary of Article

II Methods-

a.get_article() - Method allowing system to fetch articles entered by user.

b.process_article()-Method allowing the system to extract summary of article.

f.generate_summary_result() - Method allowing system to generate summary of article

5.2 INTERACTION DIAGRAM

5.2.1 Sequence diagram

Description-

1.The initialization of process takes place.

2.The system object is called which in turn initiates other objects.

3.In case of Sentiment analysis the system takes the article and passes it to another object called Sentiment analysis which generates an analysis report from which the system generates a result and passes it to the user. The result specifies the sentiment (positive or negative) of the article.

4.In case of Fake News analysis the system takes the article and passes it to another object called Fake News analysis which generates an analysis report from which the system generates a result and passes it to the user. The result specifies the validity(true or false) of the article.

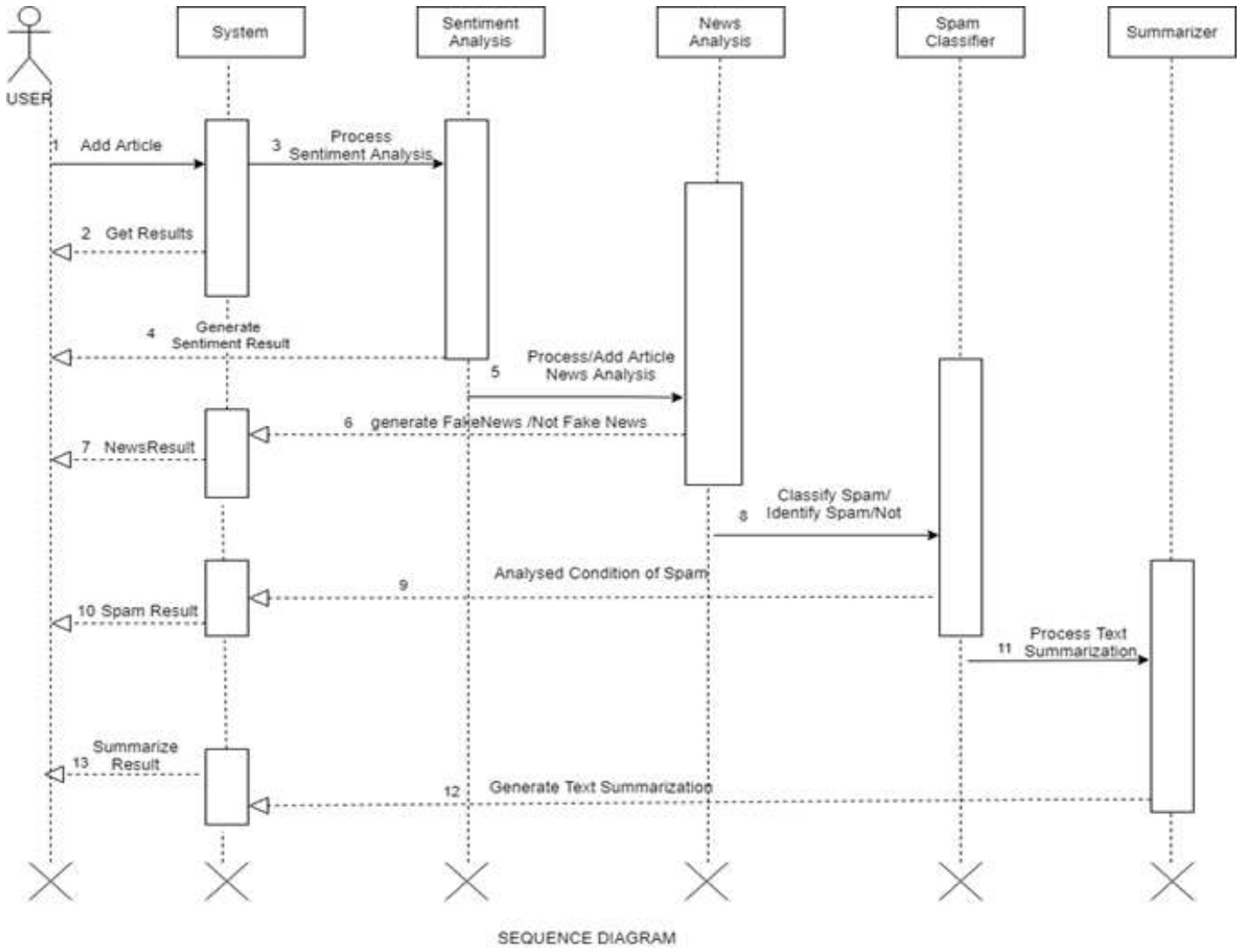


Figure 5.2 Sequence Diagram

5.In case of Spam analysis the system takes the article and passes it to another object called Spam analysis which generates an analysis report from which the system generates a result and passes it to the user. The result specifies the authenticity (spam or not spam) of the article.

6.In case of Summarizer the system takes the article and passes it to another object called Summarizer which generates an analysis report from which the system generates a result and passes it to the user. The end result is the summary of the article.

7.At last there is End of the process which is the termination of project.

5.2.2 Collaboration Diagram

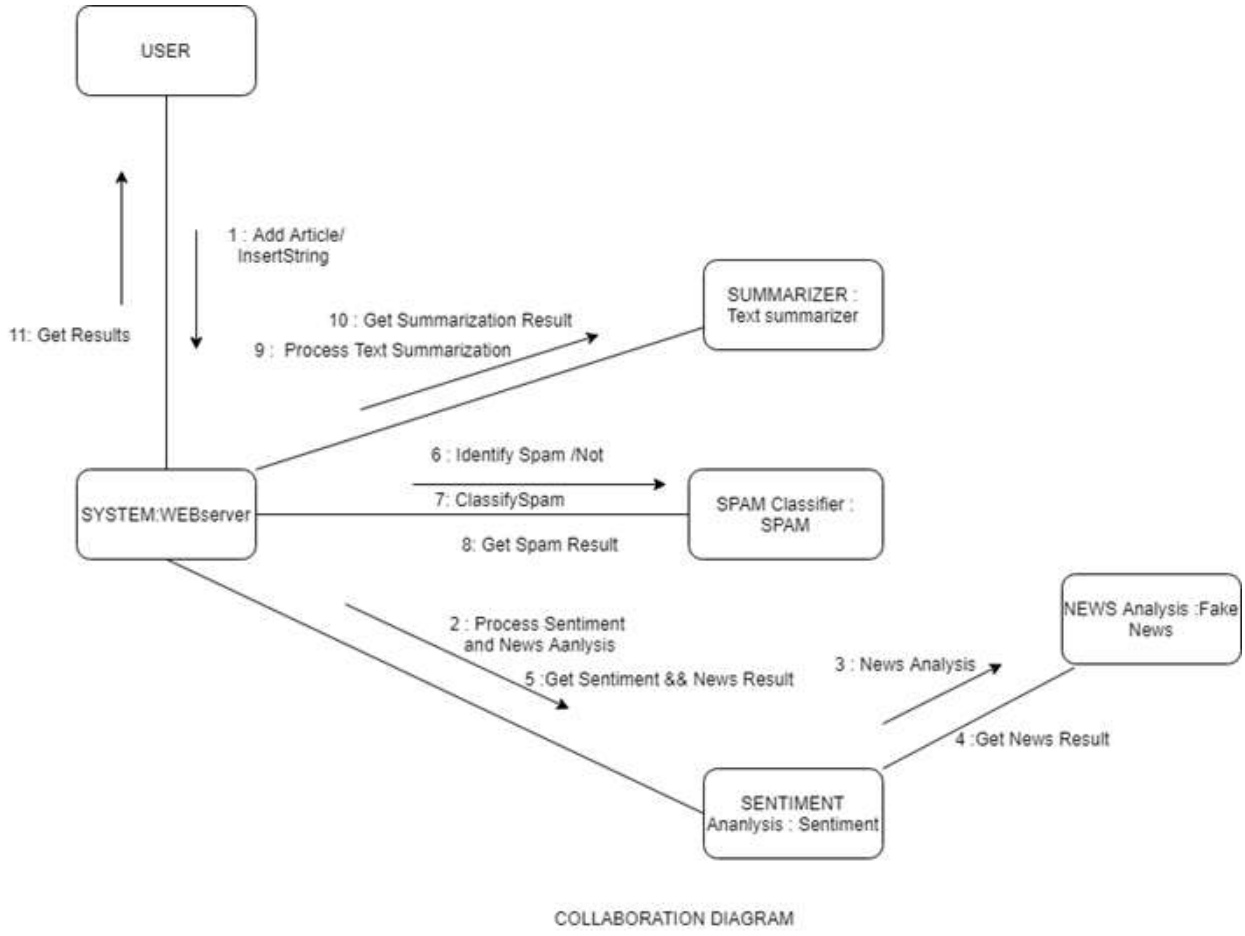


Figure 5.3: Collaboration Diagram

Description-

1.The initialization of process takes place.

2.The System WEBserver object is called for getting articles from the user for performing different operations.

3.After System WEBserver is called there are three different objects that can be called followed as Sentiment Analysis, Spam Classifier, Summarizer.

4.There is one more object that is called after selecting the Sentiment Analysis that is the Fake News Analysis object.

5.The Summarization object is called to get the process text summarization of the article and the object generates a summary as the result.

6.After Summarization the system generates the summary which is transferred to the user.

7.The Spam Analysis object is called for identifying whether the given article is spam or not.

8.After Spam Analysis the system generates the result which is transferred to the user.

9.The sentiment Analysis object is called with the objective of abstracting sentiment (positive or negative) from the article.

10.After Sentiment Analysis the system generates the result which is transferred to the user or the Fake News Analysis object is called.

11.After sentiment analysis if Fake News Analysis object is called then it is called with the aim of identifying whether the article is fake or not.

12.After Fake News Analysis the system generates the result which is transferred to the user.

13.At last there is End of the process which is the termination of the project.

5.3 STATE DIAGRAM

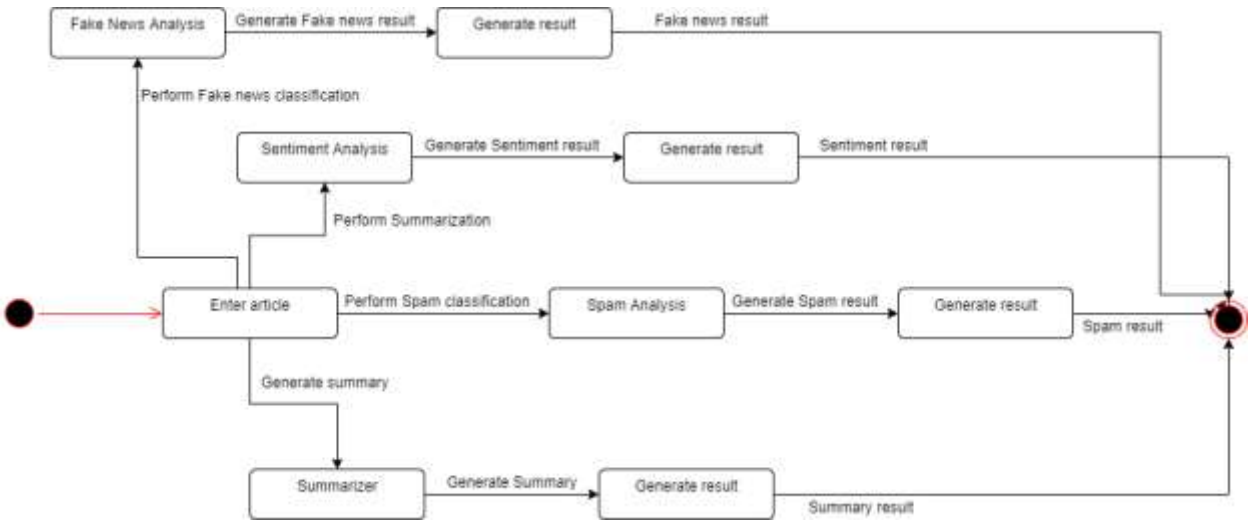


Figure 5.4: State Diagram

Description-

1.The first state of the user is to enter an article.

2.There are multiple states after the first state.

2.1. One of the following states is fake news analysis.

2.1.1. The next state is generating result of the analysis

2.2. Another of the following states is sentiment analysis.

2.2.1. The next state is generating result of the analysis

2.3. Another of the following states is spam detection.

2.3.1. The next state is generating result of the analysis

2.4. Last of the following states is summarization of articles.

2.4.1. The next state is generating result of the analysis

5.4 DETAILED OBJECT DIAGRAM

Description for activity diagram

1.The first activity is visiting the website.

2.The second activity is Enter Article Here users enter the news article.

3.In the third activity the user has to select the activity he wants to perform out of four activities listed as Sentiment Detection, Summarizer, Fake News Detector and Spam Detector.

4.If the User chooses the activity Sentiment Detection Then following are the next activities performed-

4.1. The first activity is creating a bag of words, it is a function in which a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

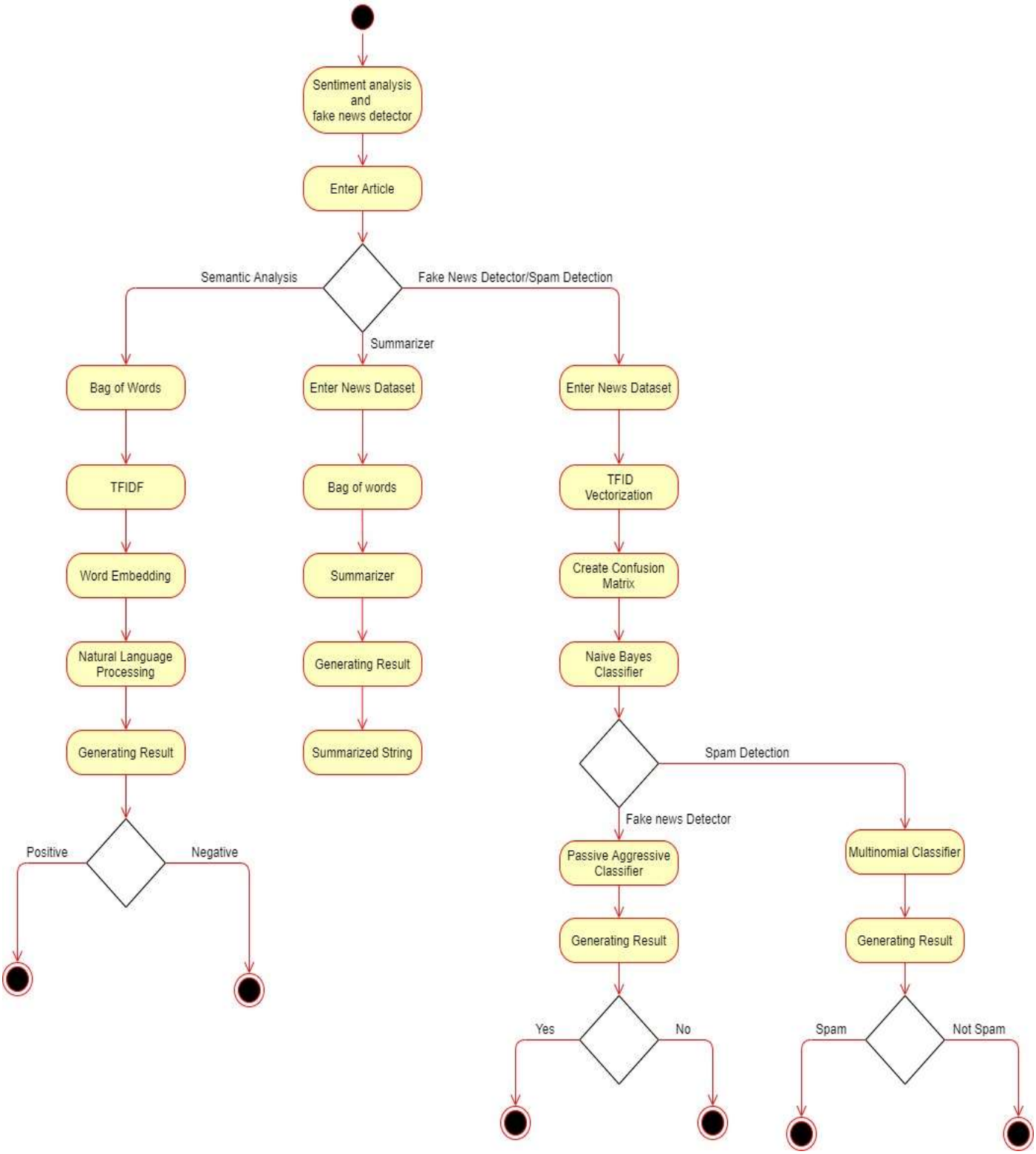


Figure 5.5: Detailed Object Diagram

4.2. The Second activity is TFIDF VECTORIZATION here the frequency of each word is counted and stored.

4.3. The third activity id word embedding. Word embedding is any of a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers.

4.4. Next step is applying Natural Language Processing to Extract the sentiment of the string.

4.5. The fifth activity is generating results out of the analysis.

4.6. If last step is positive means that sentiment is positive else sentiment is negative.

5.If the user chooses the activity summarizer the following activities are performed-

5.1. First activity is converting the article into the form of a dataset.

5.2. The Second activity is creating the bag of words. It is a function in which a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

5.3. The third activity is applying the summarize function.

5.4. The last activity is generating the results out of analysis.

6.If the person chooses activities Fake news detector or Spam detector following are the common activities involved in both-

6.1. The first activity in both activities is converting the article in the form of a data frame that is a dataset.

6.2. The second activity is applying the TFIDF VECTORIZATION that is counting the frequency of each term in the article

6.3. The next activity is creating the confusion matrix for getting the accuracy matrix.

6.4. The next activity is applying the Naive Bayes Classifier.

6.5. If the person has chosen Fake new detector here are some other activities involved-

6.5.1. The next activity is applying the passive aggressive classifier which remains passive the condition is true and becomes aggressive when the condition is false.

6.5.2. The next activity is generating the result from the analysis, if the condition is yes it means the news is true, else if condition is no it means the news is fake.

6.6. If the person has chosen Spam Detector here are some other activities involved-

6.6.1. The next activity is applying the multinomial classifier to get the desired result

6.6.2. The next activity is generating the result from the analysis, if the condition is yes it means the article is Spam, else if condition is no it means the article is not Spam.

5.5 OBJECT DIAGRAM

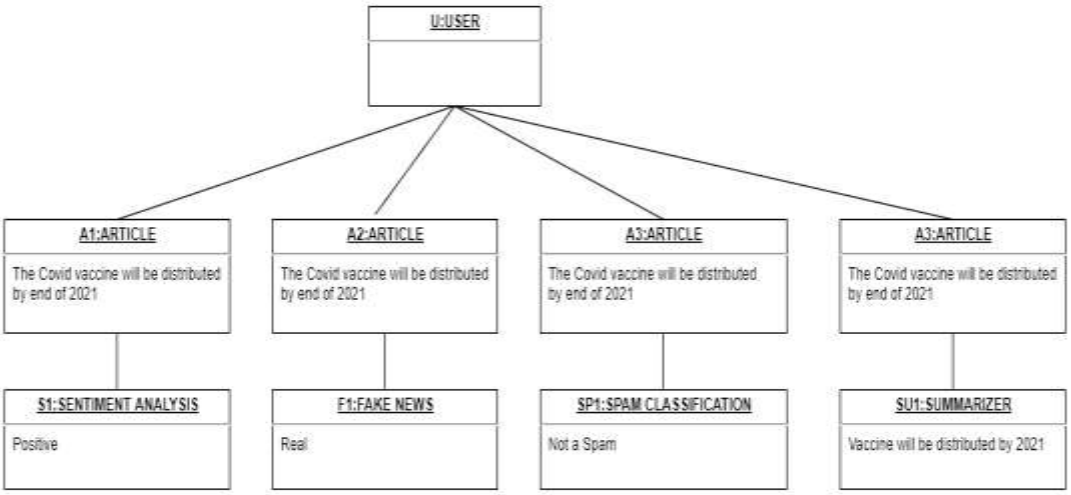


Figure 5.6: Object Diagram

Description

In object diagram is a graph of instances, including objects and data values. A static object diagram is an instance of class diagram that shows the snapshot of the detailed state of the system at a point in time. The use of object diagrams is fairly limited, namely to show examples of data structure.

In the object diagram there are two objects- user and Article. These objects are connected through a single association. User and system object have one same attribute article-Which is the article to be validated and the system also contains one more attribute output.

The Article class is generalized into four sub classes namely which are used one at a time:

- Fake News Detector-It Detects Whether the article is true or false.
- Sentiment Detector-It shows the sentiment of the particular article (Positive/Negative).
- Spam Detector-Detects Whether the article is spam or not.
- Summarizer-Provides the summary of the article.

5.6 COMPONENT DIAGRAM

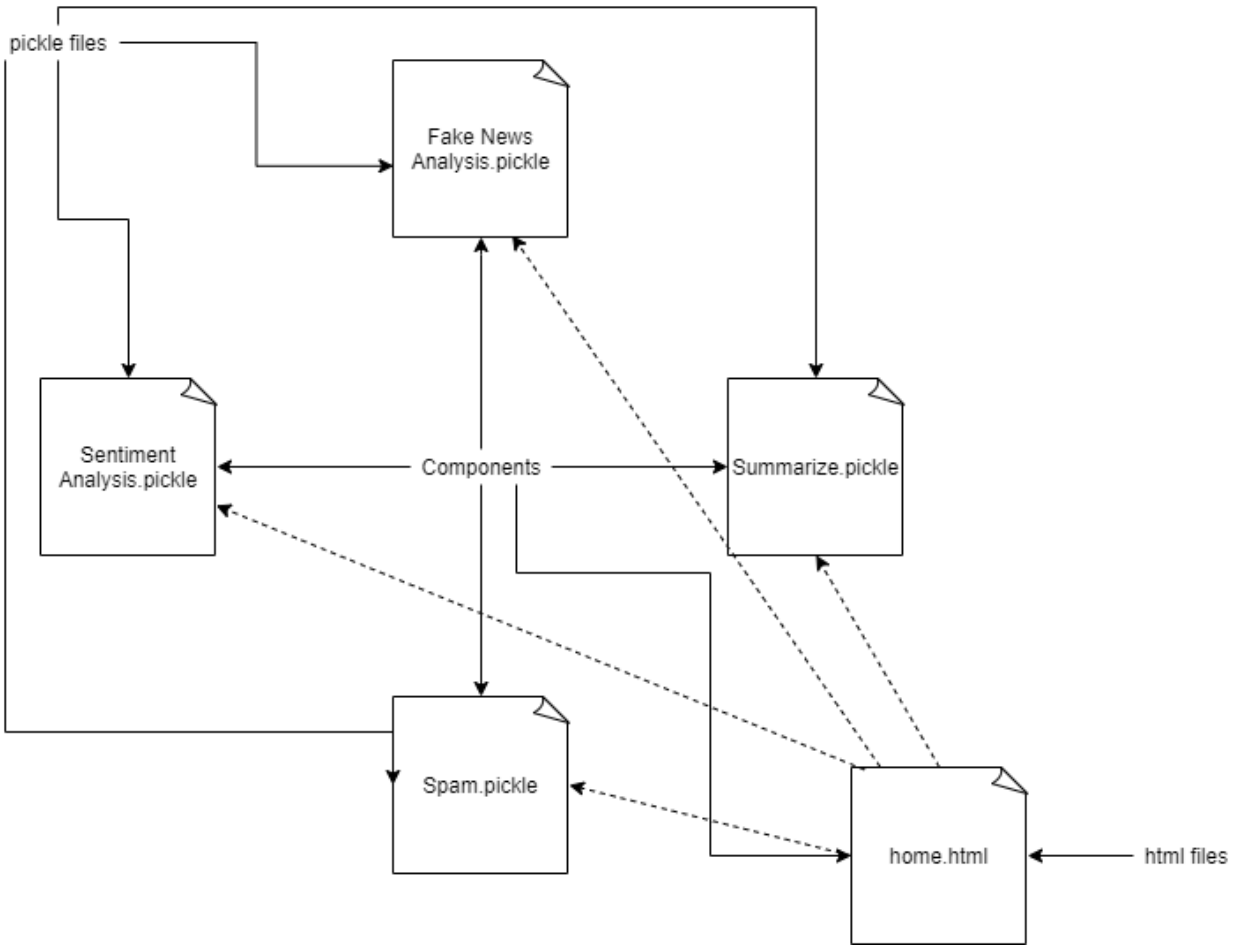


Figure 5.7 Component Diagram

Description-

1.There are two types of files first is pickle and another is html

2.There are total 4 pickle files-

2.1. The first is fake news analysis which is used to predict whether the result is true or not.

2.2. Next is Sentiment Analysis which is used to get sentiment of the result

2.3. Next is spam analysis which is used to predict whether the news is spam or not.

2.4. The last file is summarizing which is used to get the summary of the article.

3.The next type of file is html file which provide the home page and base pages form all pickle files.

DEPLOYMENT DIAGRAM

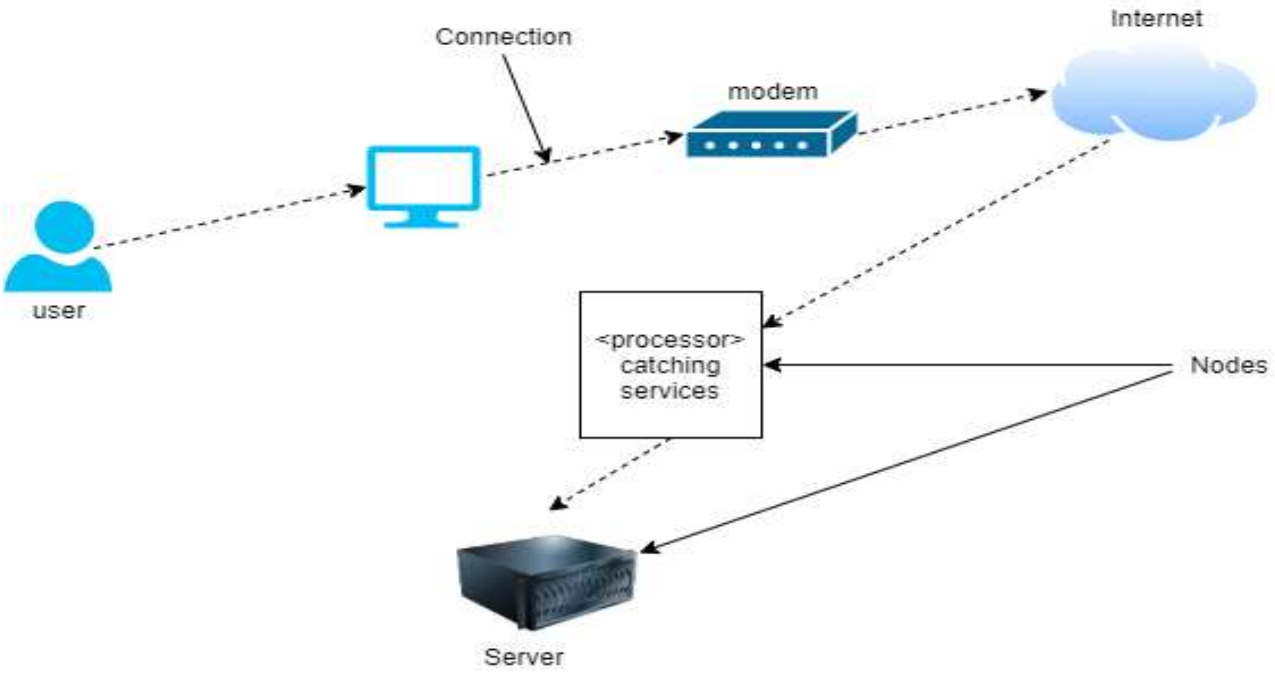


Figure 5.8: Deployment Diagram

Description-

1.The user uses a device such as computers to get connected to the internet.

2.The internet is provided using modem by service provider.

3.The internet sends user requests to the server.

4.The server processes the user request to provide the correct solution.

5.The server then responds back with the solution to the request.

5.7 TEST PLANS AND IMPLEMENTATION IMAGES

Test Plan

A Test Plan is a detailed document that describes the test strategy, objectives, schedule, estimation, deliverables, and resources required to perform testing for a software product. Test Plan helps us determine the effort needed to validate the quality of the application under test. The test plan serves as a blueprint to conduct software testing activities as a defined process, which is minutely monitored and controlled by the test manager.

As per ISTQB definition: "Test Plan is A document describing the scope, approach, resources, and schedule of intended test

Making Test Plan document has multiple benefits -

- Help people outside the test team such as developers, business managers, customers understand the details of testing.
- Test Plan guides our thinking. It is like a rule book, which needs to be followed.
- Important aspects like test estimation, test scope, Test Strategy are documented in Test Plan, so it can be reviewed by Management Team and re-used for other projects.

S. No.	Parameter	Description
1.	Introduction	Fake news detection and sentiment analysis is a web-based application that predicts the information or news is fake or genuine and also analyze the sentiment of information is positive, negative or neutral and also help in finding spam classification and summary of the information.
2.	Features to be tested	The feature that needs to be tested are fake news detection and sentiment Analysis of information or news.
3.	Test schedule	It includes some variety of the phases. For ex. Requirement understanding, test plan creation, test cases, test execution in different environments. • Firstly, team understands the requirements for implementation of the projects.

		• Then create the schedule for every phase or functionality. • Then test every functionality of system that means buttons in application, output result or recommendation and then make test cases for the test results. • After the test cases system will check on every platform or device for the environment testing. • If all testing will complete successfully then system will ready for run and then we stop testing.
4.	Environmental testing	We need some environmental requirements such as hardware, software, OS, network configurations, tools required that are system should have at least 4gb RAM, 500gb hard disk, windows 7,8,10 and 2mbps network connection.
5.	Open risk/issue	In implementation we face some issues that are data accuracy, calculations. Some functionality which are left to implement and testing. System have also some bugs and error which we will resolve soon.
6.	Exit criteria	When system has no bug/error and all functionality work properly and also system run on every platform then we will stop testing and system will ready for run.

Test Cases

A TEST CASE is a set of actions executed to verify a particular feature or functionality of your software application. A Test Case contains test steps, test data, precondition, postcondition developed for specific test scenario to verify any requirement. The test case includes specific variables or conditions, using which a testing engineer can compare expected and actual results to

determine whether a software product is functioning as per the requirements of the customer.

S.No.	Title	Input	Action	Expected Output	Actual Output	Status	Remark
1.	Verify whether application launched on local system or not.	Enter run commands	Open application on the local system.	Application should run on local system browser.	Application successfully run on local system	Pass	

2.	Verify that the application's display is adapted to the screen and all the buttons and menus work properly.		Open application in browser and check screen size and buttons.	Application display should adaptable in screen size and all buttons and menu work properly.	Application display is adaptable in screen size and all buttons and menu work properly.	Pass	
3.	Verify user able to enter information or news.	Click on text area.	Enter some information or news link.	Application allows to enter the information or news.	Application allow to enter the information or news successfully.	Pass	
4.	Check application able to process or show result on the screen.	Click on proceed.	Enter some information or news link and proceed further.	Application shows the result on the screen.	Application shows the result on the screen successfully.	Pass	
4.	Check application able to process or show result on the screen.	Click on proceed.	Enter some information or news link and proceed further.	Application shows the result on the screen.	Application shows the result on the screen successfully.	Pass	
5.	Check application provides result that news or information is fake or genuine.	Click on proceed.	Enter some information or news link and proceed further.	Application shows the result that information or news is fake or genuine.	Application provides the result that information or news is fake or genuine successfully.	Pass	
6.	Check application provides result which shows sentiment of the news or information i.e. positive, negative or neutral.	Click on proceed.	Enter some information or news link and proceed further.	Application shows the result that information or news is either positive, negative and neutral.	Application provides the result that information or news is either positive, negative and neutral successfully.	Pass	

IMPLEMENTATON IMAGES

Figure 5.9: Home Page -1

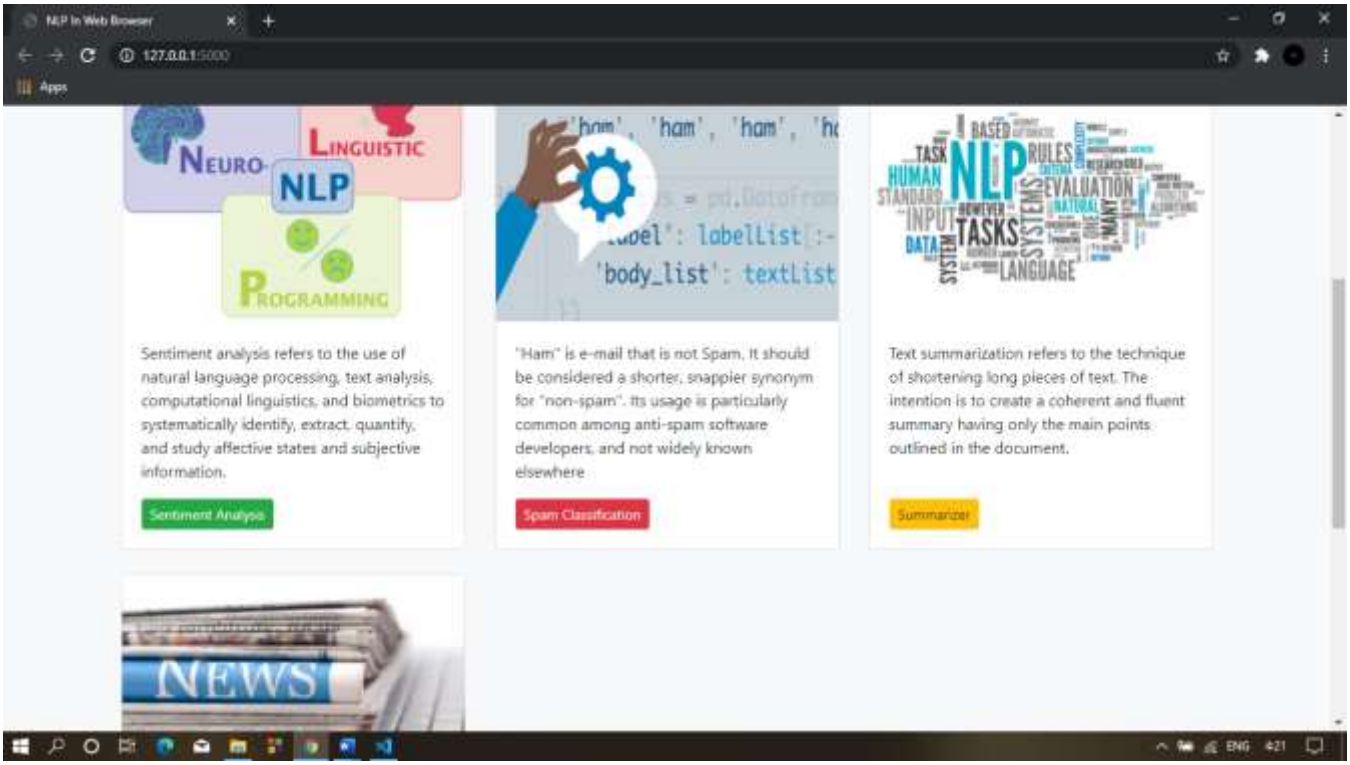


Figure 5.10: Home Page -2

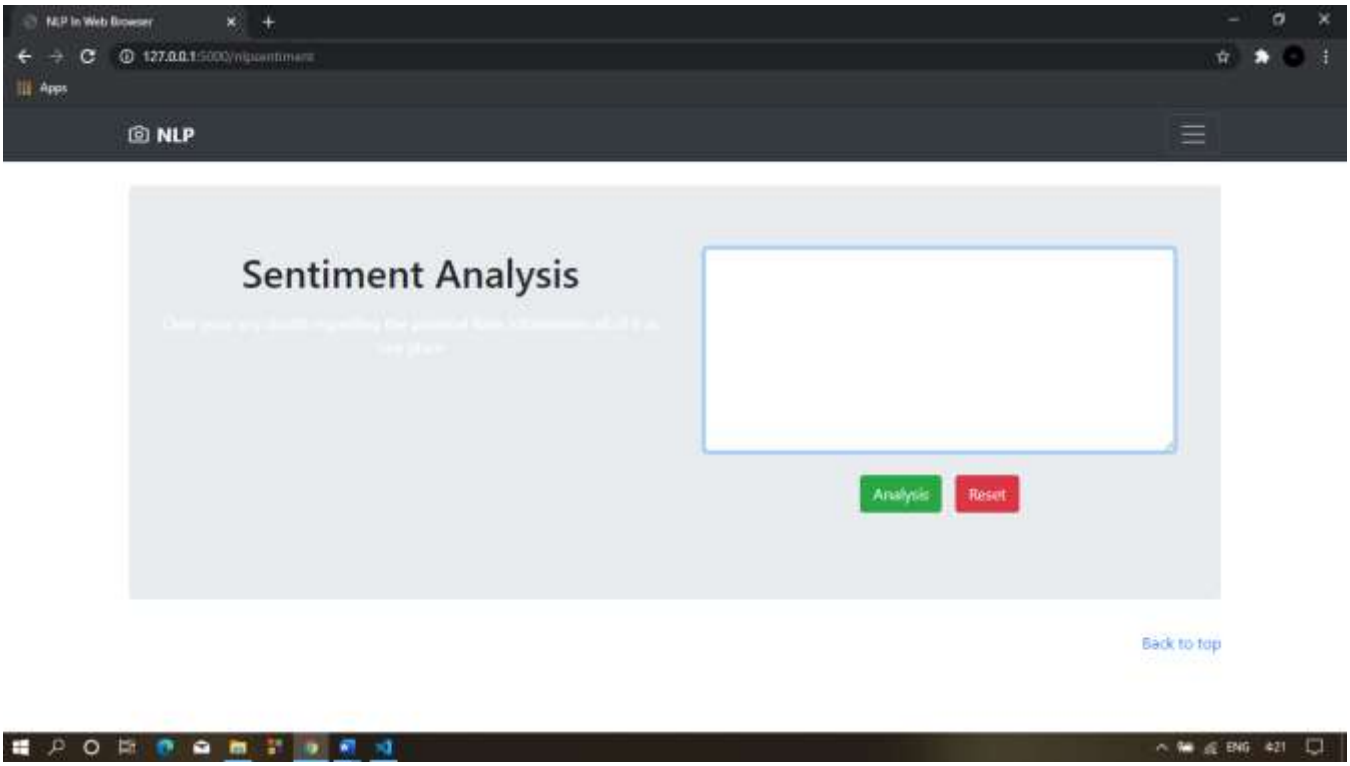


Figure 5.11: Sentiment analysis

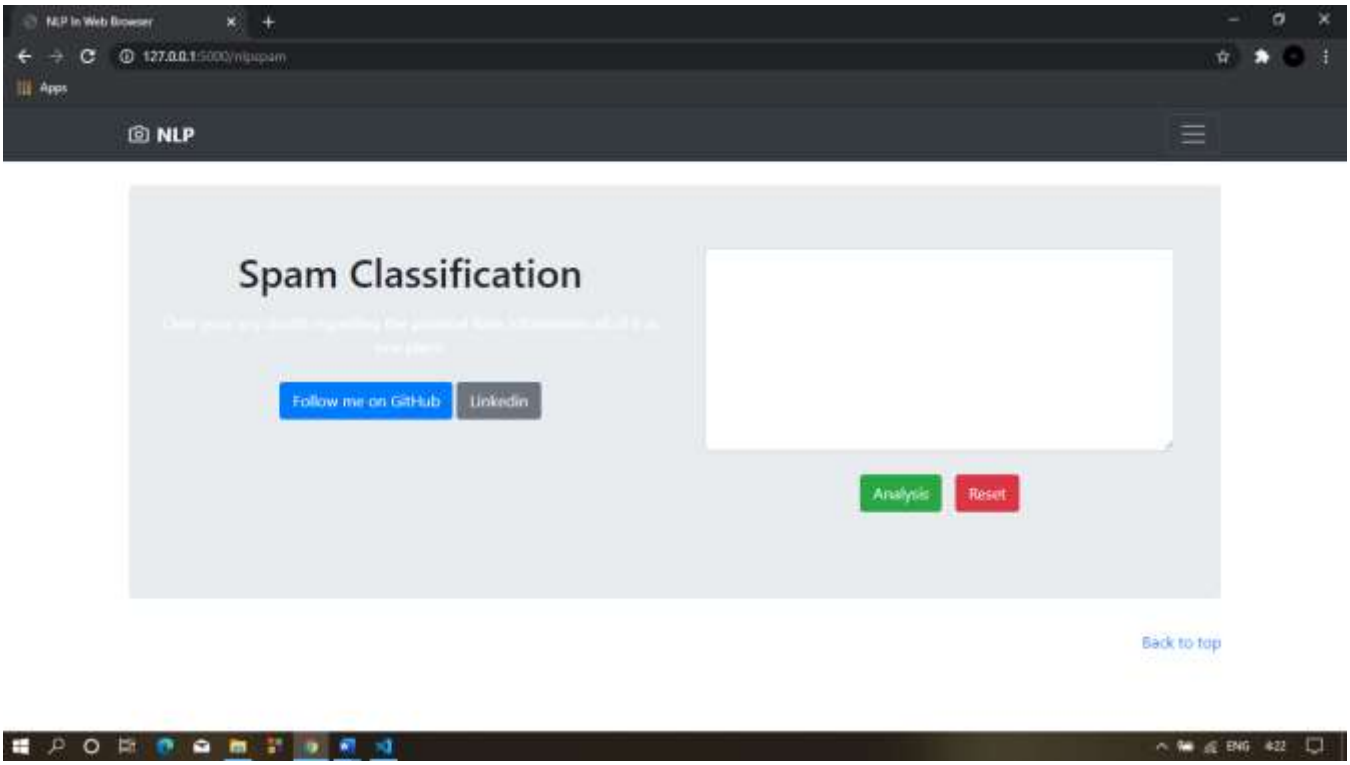


Figure 5.12: Spam classification

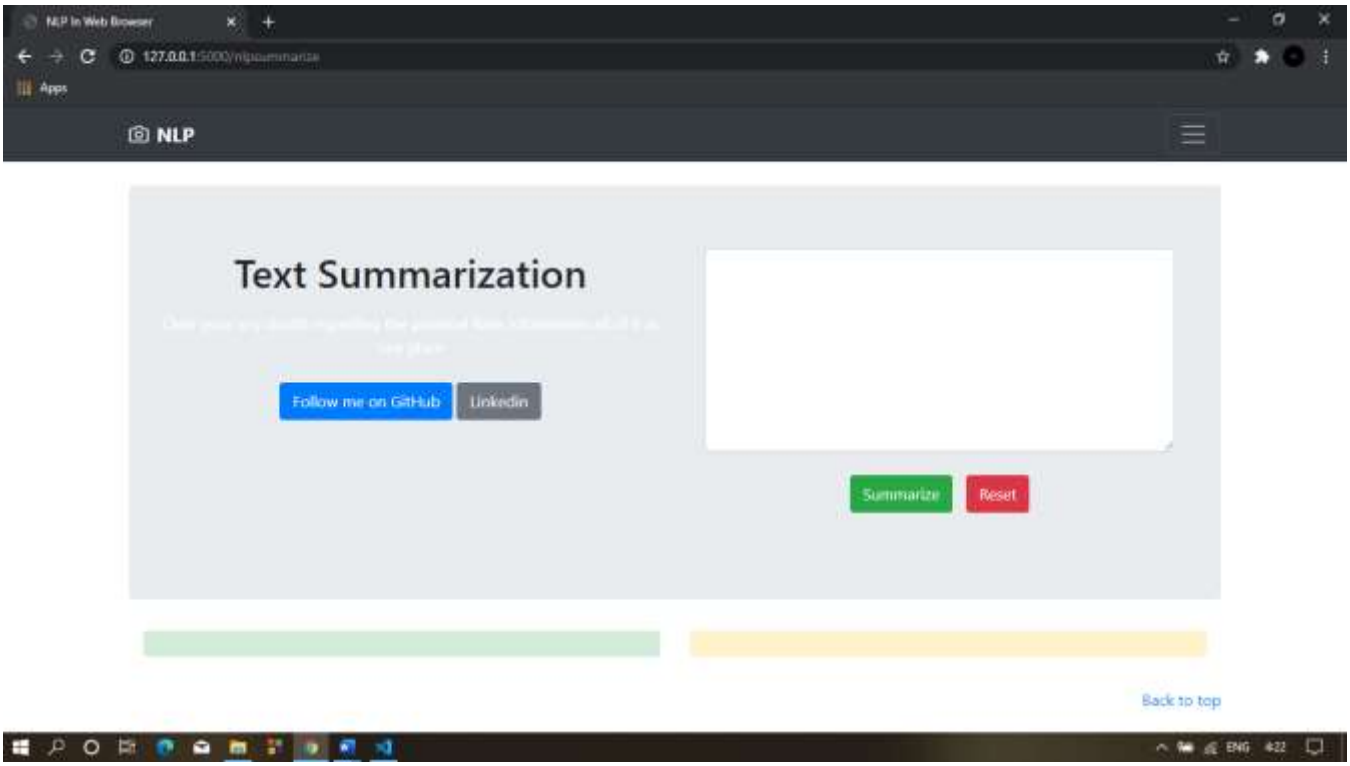


Figure 5.13: Text Summarization

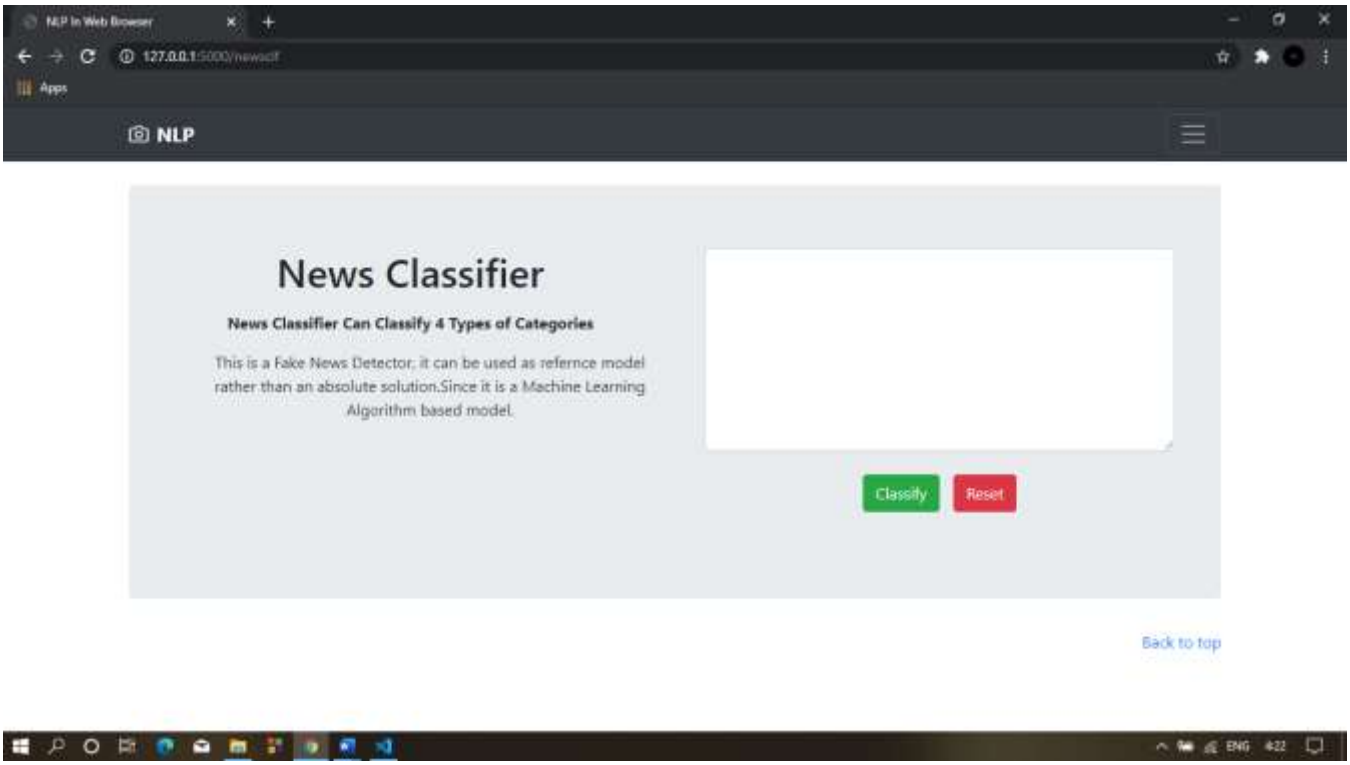


Figure 5.14: News Classifier

CONCLUSION & FUTURE WORK

6.1 Limitation of Project

One of the biggest limitations now you may apparently notice is machine translation (MT). Even google translate cannot guarantee a good translation without any modifications. The alignment and language modeling have been a challenging issue for researchers to improve MT. As the arrival of deep learning, something like word2vec comes out. There are many revolutions of NLP are going on. We still have a long way to go to understand natural language. But the future is bright.

Like all opinions, sentiment is inherently subjective from person to person, and can even be outright irrational. It's critical to mine a large — and relevant — sample of data when attempting to measure sentiment. No particular data point is necessarily relevant. It's the aggregate that matters. An individual's sentiment toward a brand or product may be influenced by one or more indirect causes; someone might have a bad day and tweet a negative remark about something they otherwise had a pretty neutral opinion about. With a large enough sample, outliers are diluted in the aggregate. Also, since sentiment very likely changes over time according to a person's mood, world events, and so forth, it's usually important to look at data from the standpoint of time

As to sarcasm, like any other type of natural language processing (NLP) analysis, *context matters*. Analyzing natural language data is, in our opinion, the problem of the next 2-3 decades. It's an incredibly difficult issue, and sarcasm and other types of ironic language are inherently problematic for machines to detect when looked at in isolation. It's imperative to have a sufficiently sophisticated and rigorous enough approach that relevant context can be taken into account. For example, that would require knowing that a particular user is generally sarcastic, ironic, or hyperbolic, or having a larger sample of the natural language data that provides clues to determine whether or not a phrase is ironic.

6.2 Future Enhancement

As technology continues to grow, future applications of Natural language Processing will be more user-oriented. Data scientists dealing with natural language processing and other aspects of AI, rely on established NLP library platforms to build and test their applications. Today, the platform pool is made up of trusted mainstays such as OpenNMT, Stanford's CoreNLP, SpaCy and Tensor Flow.

As NLP progresses in the future, bigger and better platforms are going to get built as alternatives to existing ones and their flaw ones. Some new platforms, such as Spark NLP, have already been released and cited as the future of NLP. Spark NLP is known for its speed, scalability and its massive library of pipelines, pre-trained neural network models and embeddings.

The ultimate test for NLP in the future will be whether it can go beyond mere natural language processing and be able to shift to understanding the human language better. So far, the technology has only had to derive meaningful responses from raw text data and it has proven its worth. In the future, experts predict that natural language processing will have to evolve in its function to become natural language understanding.

The latter is a more sophisticated level of processing that would allow processors to understand language as it naturally occurs rather than just processing words and text to derive meaning. That would ideally involve understanding the accents, slang and other nuances that make up natural language. It would also, to a certain degree, mean that the machines would then be able to generate text for themselves. This is the ultimate future of natural language processing, but how it plays out in the coming years remains to be seen.

As the Technology is advancing the NLP will sure to be advance as well, so in the coming year there will be better classifier and regression algorithm which will perform even better than today. So as the better is available in the market we will upgrading the software with the new algorithm which will enhance the performance and accuracy as well.

And last but not the least as of the data is coming on daily is increasing as per day, so the that amount is can also be used for training the model as we know the more, we the model with different varieties of data the better the model will perform. So, in future the better model will be deployed which is trained not only by later data but also with the new variety of data.

BIBLIOGRAPHY & REFERENCES

7.1Reference Books

- Natural Language Processing with Python *by Steven Bird, Ewan Klein and Edward Loper.*

- Foundations of Statistical Natural Language Processing *by Christopher Manning and Hinrich Schütze.*

- Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition *by Dan Jurafsky and James H. Martin*

- Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 1st Edition by *Aurélien Géron*

- NLP: The Essential Guide to Neuro-Linguistic Programming

7.2Other Documentations & Resources

- Fake News Detection on Social Media: A Data Mining Perspective *by Kai Shu, Amy Sliva Suhang Wang , Jiliang Tang , Huan Liu*

- Fake News Detection as Natural Language Inference1, *Kai-Chou Yang, Timothy Niven, Hung-Yu Kao*

- Fake News Detection Using Machine Learning approaches: A systematic Review *by Syed Ishfaq Manzoor, Jimmy Singla, Nikita*

- Analysis of Classifiers for Fake News Detection *by VasuAgarwala, H.ParveenSultanaa ,SrijanMalhotraa , AmitrajitSarkarb*

- Deep Learning for Hate Speech Detection in Tweets *by Pinkesh Badjatiya (IIIT-H) , Shashank*

Gupta (IIIT-H), Manish Gupta (Microsoft), Vasudeva Varma (IIIT-H)

- Multilingual Twitter Sentiment Classification: The Role of Human Annotators *by Igor Mozetič, Miha Grčar, and Jasmina Smailovič, from the Department of Knowledge Technologies at the Jožef Stefan Institute*

Article

- *https://towardsdatascience.com/spam-classifier-in-python-from-scratch-27a98ddd8e73*
- *https://www.kaggle.com/benvozza/spam-classification*
- *https://www.kdnuggets.com/2017/03/email-spam-filtering-an-implementation-with-python-and-scikit-learn.html*
- *https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/*
- *https://towardsdatascience.com/detecting-fake-news-with-and-without-code-dd330ed449d9*
- *https://github.com/nishitpatel01/Fake_News_Detection*
- *https://realpython.com/sentiment-analysis-python/*
- *https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6*
- *https://monkeylearn.com/blog/sentiment-analysis-with-python/*