**Mini Project Report on**



**REAL-TIME SENTIMENTAL ANALYSIS USING NATURAL LANGUAGE PROCESSING**



**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF COMPUTER APPLICATION**

**Submitted by:**

**PALLAVI** **2102258**

**Under the Mentorship of**

**Mr. Anuj Rawat**



**Department of Computer Application**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**July 2023**

**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled “**REAL-TIME SENTIMENTAL ANALYSIS USING NATURAL LANGUAGE PROCESSING**” in partial fulfillment of the requirements for the award of the Degree of Bachelor of Computer Application of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of Mr. Anuj Rawat, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name of Candidate - Pallavi

Signature of Candidate……………………

This is Certify that the above mentioned statement in the candidate declaration is correct to the best of my knowledge.

Date………………

Signature of Guide……………………..

**ACKNOWLEDGEMENT**

I would like to show my gratitude for the assistance co-operation guidance and clarification provided by “Graphic Era Deemed to Be University, Dehradun” during the development of REAL -TIME SENTIMENTAL ANALYSIS USING NATURAL PROCESSING LANGUAGE. My extreme thanks to Mr. Anuj Rawat who guided me throughout the project and helped me out to solve the problems I conquered during the project development. Without him willing deposition and above all faith in me, this project count not have been completed in due time. His readiness to discuss all important matters at works deserves special attention.

I would like to thanks all the faculty members of Computer Science Department who have helped me with their valuable suggestions and guidance throughout the various phases of the completion of the project.

PALLAVI

(2102258)

**ABSTRACT**

With the advancement of web technology and its growth, there is a huge volume of data present in the web for internet users and a lot of data is generated too. Internet has become a platform for online learning, exchanging ideas and sharing opinions. Social networking sites like Twitter, Facebook, Google+ are rapidly gaining popularity as they allow people to share and express their views about topics, have discussion with different communities, or post messages across the world. There has been lot of work in the field of sentiment analysis of twitter data. This survey focuses mainly on sentiment analysis of twitter data which is helpful to analyze the information in the tweets where opinions are highly unstructured, heterogeneous and are either positive or negative, or neutral in some cases. In this paper, we provide a survey and a comparative analysis of existing techniques for opinion mining like machine learning and lexicon-based approaches, together with evaluation metrics. Using various machine learning algorithms like Naive Bayes, Max Entropy, and Support Vector Machine, we provide research on twitter data streams. We have also discussed general challenges and applications of Sentiment Analysis on Twitter.

**Table of Contents**

| **Chapter No.** | **Description** | **Page No.** |
| --- | --- | --- |
| Chapter 1 | Introduction | **1-13** |
| Chapter 2 | Literature Survey | **14-16** |
| Chapter 3 | Methodology | **17-19** |
| Chapter 4 | Result and Discussion | **20-39** |
| Chapter 5 | Conclusion | **40-43** |
|  | References | **44** |

**Chapter1**

**INTRODUCTION**

**Introduction to Sentiment Analysis**

Sentimental Analysis is a process of collecting and analyzing data based upon the person feelings, reviews and thoughts. Sentimental Analysis often called as opinion mining as it mines the important feature from people opinions. Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data.

Sentiment Analysis can be done at document, phrase and sentence level. In document level, summary of the entire document is taken first and then analyze whether the sentiment is positive, negative or neutral. In phrase level, analysis of phrases in a sentence is taken in account to check the polarity. In Sentence level, each sentence is classified in a particular class to provide the sentiment.

Sentiment Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelings or thoughts which they provide in form of text. Sentiment Analysis is domain centered, i.e. results of one domain cannot be applied to other domain. Sentimental Analysis is used in many real life scenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing.

Twitter is a micro blogging platform where anyone can read or write short form of message which is called tweets. The amount of data accumulated on twitter is very huge. This data is unstructured and written in natural language. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negative, or neutral with the help of different machine learning algorithm.

### Introduction to Python

Python is a high level, dynamic programming language. Python3.4 version was used as it is a mature, versatile and robust programming language. It is an interpreted language which makes the testing and debugging extremely quickly as there is no compilation step. There are extensive open source libraries available for this version of python and a large community of users.

Python is simple yet powerful, interpreted and dynamic programming language, which is well known for its functionality of processing natural language data, i.e. spoken English using NLTK. Other high level programming languages such as ‘R’ and ‘Mat lab’ were considered because they have many benefits such as ease of use but they do not offer the same flexibility and freedom that Python can deliver.

As the internet is growing larger, its reach to the masses is becoming wider. Social Media and Micro blogging platforms like Twitter, Facebook, Tumbler dominate in spreading encapsulated news and trending topics across the globe at a rapid pace. A topic or news becomes trending if many users are contributing their opinion and judgments, thereby making it a valuable source of online perception on that particular topic. These topics generally intended to spread awareness or to promote political campaigns, public figures during elections, product endorsements, and entertainment like award shows, movies. Large organizations and firms take advantage of people's feedback on these platforms to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of the upcoming IPhone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analyzing interesting patterns from the infinite social media data for business-driven applications. Sentiment analysis is the prediction of emotions in a word, sentence, or corpus of documents. It is intended to serve as an application to understand the opinion, attitudes, and emotions expressed within an online mention. The intention is to gain or access an overview of the wider public opinion behind certain topics. Precisely, it is a paradigm of categorizing conversations into positive, negative, or neutral labels. Many people use social media sites for networking with other people and to stay up- to-date with news and current events. These sites (Twitter, Facebook, Instagram, Google+) offer a platform for people to voice their opinions. For example, people quickly post their reviews online as soon as they watch a movie and then start a series of comments to discuss the acting skills depicted in the movie. This kind of information forms a basis for people to evaluate, a rate about the performance of not only any movie but about other products and to know about whether it will be a success or not. This type of vast information on these sites can be used for marketing and social studies. Therefore, sentiment analysis has wide applications and includes emotion mining, polarity, classification, and influence analysis. Twitter is an online networking site driven by tweets which are 280 characters limited messages. Thus, the character limit enforces the use of hash tags for text classification. Currently, around 6500 tweets are published per second, which results in approximately 561.6 million tweets per day. These streams of tweets are generally noisy reflecting multi-topic, changing attitudes information is an unfiltered and unstructured format. Twitter sentiment analysis involves the use of natural language processing to extract, identify to characterize the sentiment content.

Datasets Through this study, a myriad of different techniques have been explored in order to learn and identify the techniques that best cater to the domain of sentiment and semantics analysis. In this pursuit, the authors developed certain unique own models that yielded remarkable accuracies based on multimedia inputs. To train these models, certain datasets were used that could ensure a plethora of different options for the holistic coverage of most bases in the model. For textual sentiment analysis that takes the input from a user in the form of text and takes the speech to text input, the Amazon Reviews dataset was employed. Meanwhile for the chat bot model, a custom dataset was created in the .json format consisting of several relevant tags and responses.



### Need of Sentimental Analysis

#### Industry Evolution

Only the useful amount of data is required in the industry as compared to the set of complete unstructured form of the data. However the sentiment analysis done is useful for extracting the important feature from the data that will be needed solely for the purpose of industry. Sentimental Analysis will provide a great opportunity to the industries for providing value to their gain value and audience for themselves. Any of the industries with the business to consumer will get benefit from this whether it is restaurants, entertainment, hospitality, mobile customer, retail or being travel.

#### Research Demand

Another important reason that stands behind the growth of SA deals with the demand of research in evaluation, appraisals, opinion and their classification. Present solution for the purpose of sentiment analysis and opinion mining are rapidly evolving, specifically by decreasing the amount of human effort that will be required to classify the comments. Also the research theme that will be based in the long established disciplines of computer science like as text mining, machine learning, natural language processing and artificial intelligence, voting advise applications, automated content analysis.

#### Decision Making

Every person who stores information on the blogs, various web applications and the web social media, social websites for getting the relevant information you need a particular method that can be used to analyze data and consequently return some of the useful results. It is going to be very difficult for company to conduct the survey that will be on the regular basis so there comes the need to analyze the data and locate the best of the products that will be based on user’s opinions, reviews and advices. The reviews and the opinions also help the people to take important decisions helping them in research and business areas.

#### Understanding Contextual

As human language is getting very complex day by day so it has become difficult for the machine to be able to understand human language that can be expressed in the slangs, misspelling, nuances, and the cultural variation. Thus, there will be a need of system that will make better understanding between the human and the machine language.

#### Internet Marketing

Another important reason behind the increase in the demand of sentimental analysis is the marketing done via internet by the business and companies organization. Now they regularly monitor the opinion of the user about their brand, product, or event on blog or the social post. Thus, we see that the sentimental Analysis could also work as a tool for marketing tool.

## Applications of Sentiment Analysis in Business

#### Research and Analysis of Market Trends:

Sentiment Analysis helps in gauging market trends by analyzing the online presence of any brand/product/features. An upcoming brand can also use it to educate itself about what is happening in the industry and what is expected of them in its niche. The brand can use this data to make critical business decisions regarding product features, launches, etc. Deep-dive analysis algorithms have made it feasible to comprehend aspects, traits, and attributes in addition to client sentiment toward a product. These insights help organizations tailor their offerings and make them appealing to their target market. For instance, when analyzing customer feedback and menu preferences, food giants like Domino’s, KFC, Pizza Hut, and McDonald’s use sentiment analysis. This aids them in enhancing consumer satisfaction and raising sales.



#### Catering to Customer Satisfaction:

The sheer volume of requests, the variety of topics, and the variety of departments inside a firm, not to mention the urgency of any given request, make customer service management a tedious task. Customer support interactions create a vast quantity of customer data, including chat transcripts, voice recordings, product reviews, and emails projected to the [Natural Language Processing](https://www.analyticsvidhya.com/blog/2022/09/sentiment-analysis-on-flipkart-dataset/) (NLP) learning model, which performs Sentiment Analysis on the given data and generates a response. This response can help to determine the priority levels for customers and can also help decide the best-suited resource for a variety of queries. For instance, a smart phone company may decide to work on improvements in the app after studying the sentiment in product evaluations and finding that the majority of unfavorable reviews after product launch mention “bug in application”. It can help them set agendas for the following quarters.



#### Decision-Making for Investments:

The financial market’s high volatility and psychological elements, such as user perceptions of policy changes, new investments, or natural calamities, significantly impact how stock prices fluctuate. Sentiment Analysis of such data and financial news can aid in predicting profitable options in an otherwise unpredictable scenario. It is imperative that traders have lightning-fast reflexes to execute deals in nanosecond increments. A variety of causes influence the market’s sentiments. It has been observed that there is a correlation between changes in stock price and the polarity of the most popular comments mentioning a company’s stock symbol.



#### Cyber Bullying on Social Media & Cyber Security:

With easy accessibility to the Internet, online presence is increasing, hence, online threats. There are many instances of hate comments and discriminatory mentions that are reported. Sentiment Analysis can be quite helpful in keeping a close check on practices like cyber bullying. Also, we must investigate more complex additional data that can boost prediction accuracy and offer an understanding of the behavioral elements involved in developing and carrying out a cyber attack. This similar principle can provide early warning signs of cyber attacks because evidence reveals that public dialogue in online sources, such as social media, is substantially connected with the likelihood of real-world activities. For instance, highly unfavorable opinions of a company can point to a high probability that the company will be the target of a cyber attack.



#### Business Intelligence:

Using sentiment analysis tools, you may measure how potential consumers perceive you. Analyzing social media and survey data, you can gain essential insights into how your company is performing well or poorly for your customers. Businesses gauge how their service offerings are being received by their target market. Sentiment analysis uses AI-driven technologies to decipher the text’s undertone by utilizing vast amounts of digital data. Aspect-based Sentiment Analysis in the business allows one to find gaps in the marketing strategy, manage one’s brand reputation, and focus on key areas where customer sentiments are positive or negative. While other businesses examine social media, Intel utilizes software from Kanjoya Inc. that uses language processing and machine-learning algorithms to identify emotions in writing.



#### Employee Satisfaction:

By analyzing the surveys, peer reviews, and feedback from managers, information about employee behavior could be obtained, and their grievances could be handled well. It can also be used to track de-motivation, dissatisfied employees, and initiatives that could be taken to make them feel productive by launching events that cater to them. You can enhance productivity, lower turnover, and better engage your workforce by examining the tone of employee feedback. Use sentiment analysis to assess employee surveys, as well as emails, Slack messages, online reviews, tweets on professional platforms, and more. For instance, IBM analyses and assesses the sentiment of employee posts on its internal social networking site. The topics which are most popular among employees are given the most weight.



#### Marketing Campaigns:

In psychology, sociology, and political science, sentiment analysis finds application in examining trends, viewpoints, inherent bias, measure response, etc. The functionality in Sentiment Analysis can be particularly helpful in creating campaigns targeted toward marketing a product or a feature in a company or even when launching a new product. Companies can work on audience engagement and contextualize and granulate key performance indicators. They can build better messaging for their marketing and advertising campaigns that can aid in smooth transitions by keeping the customer’s feedback in mind.



#### Crisis Management for Brands:

Brand Management is gaining traction these days. Brands are taking giant leaps in resorting to practices conducive to their growth. Many of those practices respond to what information or perception is derived from the target audience. The online presence, reviews, and vocal expectations are accessed to make judgments about the new campaigns aimed at improvement and promotion. If a change is detected in the public’s perception of any component of your business, sentiment analysis can reveal it to you. Peaks or dips in sentiment scores provide a starting point for developing new marketing campaigns, sales rep or customer service agent training, or product upgrades.

**Example:** Luxury fashion house Balenciaga was criticized for unveiling their holiday Ad campaign collection, which showed kids holding teddy bears in bondage harnesses and costumes. This Ad was taken down within a few hours of its launching, and the brand issued a public apology. Hence, Semantic Analysis can prove to be a valuable practice for the reputation management of brands because the longer a negative sentiment lingers on a social platform, the more damage it causes to a brand’s reputation.



#### Politics:

It enables you to forecast outcomes based on the views of social media users. We are unable to foresee the future. However, sentiment Analysis can predict how people will vote in the future election if you have enough historical information about previous elections and how they played out. Politics have seen a great application of this technique in analyzing how the audience has taken a policy announcement. The feedback from this has also found application in determining the sentiments associated with potential candidates, which can very well be used in a competitive environment.

Twitter has been utilized in studies in the past to examine conversations and communication during the 2020 U.S. Presidential elections. Twitter data gathered ten days before and after election was subjected to sentiment analysis. The goal was to correlate Twitter user characteristics, such as their number of followers, amount of activity, and quantity of tweets, with the tone of the debates. The main focus was the relationship between the tweet’s sentiment and the hash tag topic that was included in the tweet. The tweet’s status and its author’s status are related to the sentiment analysis of the tweets. By doing this, it can reveal details about user demographics and how they act during a tumultuous election season.



#### Banking:

Artificial Intelligence (AI) powered sentiment research is essential for financial organizations experiencing a digital shift to promote financial products and services effectively. Over the last decade, the financial industry has experienced unprecedented transformation. With innovative digital financial platforms and solutions, new challenger start-ups have entered the market, posing a danger to an industry that has grown too complacent due to outdated leadership, obsolete thinking, and legacy systems. Due to this, banks and insurance businesses have been obliged to adopt a more agile digital transformation strategy by integrating Machine Learning algorithms into many elements of their operations. This can be targeted toward customer retention. Financial firms are discovering a lot about how customers respond to their services thanks to social media monitoring. For instance, BBVA Compass studied feedback on social media to enhance its rewards program. With the help of analytics, BBVA was able to see trends, understand how customers on social media feel about the bank, and take advantage of competitor products’ advantages. BBVA increased the cash back benefits on its credit cards as a result.

**PROBLEM STATEMENT**

In this project, we try to implement an NLP **Twitter Sentiment Analysis Model** that helps to overcome the challenges of sentiment classification of tweets. We will be classifying the tweets into positive, negative or neutral sentiments. With the rapid growth of the World Wide Web, people are using social media such as Twitter which generates big volume of opinion texts in the form of tweets which is available for the sentiment analysis. This translates to a huge volume of information from a human point of view which makes it difficult to extract a sentence, read them, analyze tweet by tweet, summarize them and organize them into an understandable format in a timely manner.

The dataset provided is the **Sentimental Dataset**which consists of **1,600,000 tweets** that have been extracted using the Twitter API.

**Chapter 2**

**Literature Survey**

Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification to learning the polarity of words and phrases. Given the character limitations on tweets, classifying the sentiment of Twitter messages is most similar to sentence level sentiment analysis however, the informal and specialized language used in tweets, as well as the very nature of the micro blogging domain make Twitter sentiment analysis a very different task. It is an open question how well the features and techniques used on more well-formed data will transfer to the micro blogging domain. Just in the past year there have been a number of papers looking at Twitter sentiment and buzz other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to micro blogging (e.g., emoticons) are also common, but there has been little investigation into the usefulness of existing sentiment resources developed on non-micro blogging data. Researchers have also begun to investigate various ways of automatically collecting training data. Several researchers rely on emoticons for defining their training data. Others also use hash tags for creating training data, but they limit their experiments to sentiment/non-sentiment classification, rather than 2-way polarity classification, as we do. We use data mining methods and apply the following Machine Learning algorithm for this second classification to arrive at the best result:

**NLTK**

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as Word Net, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguist, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more.

NLTK is the most famous Python Natural Language Processing Tool Kit.

NLTK is intended to support research and teaching in NLP or closely related areas, including linguistics, cognitive science, artificial information retrieval, and machine learning.

**Test Data and Trained Data**

In many areas of information science, finding predictive relationships from data is a very important task. Initial discovery of relationships is usually done with a training set while a test set and validation set are used for evaluating whether the discovered relationships hold. More formally, a training set is a set of data used to discover potentially predictive relationships. A test set is a set of data used to assess the strength and utility of a predictive relationship.

Test and training sets are used in intelligent systems, machine learning, genetic programming and statistics.

**Naive Bayes:**

Naive Bayes Classification: Many language processing tasks are tasks of classification, although luckily our classes are much easier to define than those of Borges. In this classification the naive Bayes algorithms classification, demonstrated on an important classification problem: text categorization, the task of classifying an entire text by assigning it a text categorization label drawn from some set of labels. We focus on one common text categorization task, sentiment analysis, the ex-sentiment analysis traction of sentiment, the positive, negative or neutral orientation that a writer expresses toward some object.

Naive Bayes classifier is the simplest and the fastest classifier. Many researchers claim to have gotten best results using this classifier. For a given tweet, if we need to find the label for it, we find the probabilities of all the labels, given that feature and then select the label with maximum probability. The accuracy of Uni rams is the lowest at 79.67%. The accuracy increases if we also use Negation detection (81.66%) or higher order n-grams (86.68%). We see that if we use both Negation detection and higher order n-grams, the accuracy is marginally less than just using higher order n-grams (85.92%). We can also note that accuracies for double step classifier are lesser than those for corresponding single step.

**Natural Language Processing**

Natural language processing (NLP) is a field of artificial intelligence in which computers analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform tasks such as automatic summarization, translation, named entity recognition, relationship extraction, sentiment analysis, speech recognition, and topic segmentation.

Apart from common word processor operations that treat text like a sequence of symbols, NLP considers the hierarchical structure of language: several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas, John Rehling, an NLP expert at Meltwater Group, said in How Natural Language Processing Helps Uncover Social Media Sentiment. By analyzing language for its meaning, NLP systems have long filled useful roles, such as correcting grammar, converting speech to text and automatically translating between languages. NLP is used to analyze text, allowing machines to understand how humans speak. This human-computer interaction enables real- world applications like automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction, stemming, and more.

NLP is commonly used for text mining, machine translation, and automated question answering. NLP is characterized as a difficult problem in computer science. Human language is rarely precise, or plainly spoken. To understand human language is to understand not only the words, but the concepts and how they’re linked together to create meaning. Despite language being one of the easiest things for the human mind to learn, the ambiguity of language is what makes natural language processing a difficult problem for computers to master.

**Chapter 3**

**Methodology**

**Data collection:**

-The data which is a set of tweets.

**Lower case converter:**

-In this set of review first convert from upper case to lower case reason behind this is that one word count once not again.

-For example – ‘SMALL’, ‘small’ it count both at once mean SMALL=small.

**Reduce the short word:**

-In this we reduce the word which having the length less than 3 because these words not make a sense of sentiment.

-For example –the, is etc.

**Feature vector formation:**

- Sentiment tokens and sentiment scores are information extracted from the original data set. They are also known as features, which will be used for sentiment categorization. In order to train the classifiers, each entry of training data needs to be transformed to a vector that contains those features, namely a feature vector. For the sentence-level (review-level) categorization, a feature vector is formed based on a sentence (review). One challenge is to control each vector’s dimension.

**Pre-processing of the datasets**

A tweet contains a lot of opinions about the data which are expressed in different ways by different users. The twitter dataset used in this survey work is already labeled into two categories viz. negative and positive polarity and thus the sentiment analysis of the data becomes easy to observe the effect of various features. The raw data having polarity is highly susceptible to inconsistency and redundancy.

Preprocessing of tweet include following points,

* Remove all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username).
* Correct the spelling, sequence of repeated characters is to be handled.
* Replace all the emoticons with their sentiment.
* Remove all punctuation, symbols, number.
* Remove Stop Words.
* Expand Acronyms (we can use an acronym dictionary).
* Remove Non-English Tweets.

A General model for sentiment analysis is as follows:

Word Features

Negative Tweets

Positive Tweets

Features Extractor

Training Set

Features Extractor

Tweets

Classifiers

Positive

Negative

Figure. Sentiment Analysis Architecture

**Chapter 4**

**Result and Decissions**

**Aim of the project**

In this project we are going to consider any sports event, movie trending on a given time and analyze the opinion of the public using real time data present on social media platform twitter. By using the tweets, we try to predict or come to a decision based on the mass opinions which are expressed in the tweets. With more than 321 million active users, sending a daily average of 500 million Tweets, Twitter allows businesses to reach a broad audience and connect with customers without intermediaries. On the downside, it is harder for brands to quickly detect negative content, and if it goes viral you might end up with an unexpected PR crisis on your hands. This is one of the reasons why social listening monitoring conversation and feedback in social media has become a crucial process in social media marketing. Monitoring Twitter allows companies to understand their audience, keep on top of what is being said about their brand and their competitors, and discover new trends in the industry. Are users talking positively or negatively about a product? Well, that is exactly what sentiment analysis determines.

**Online Commerce**

The most general use of sentiment analysis is in ecommerce activities. Websites allows their users to submit their experience about shopping and product qualities. They provide summary for the product and different features of the product by assigning ratings or scores. Customers can easily view opinions and recommendation information on whole product as well as specific product features. Graphical summary of the overall product and its features is presented to users. Popular merchant websites like amazon.com provides review from editors and also from customers with rating information. http://tripadvisor.in is a popular website that provides reviews on hotels, travel destinations. They contain 75 millions opinions and reviews worldwide. Sentiment analysis helps such websites by converting dissatisfied customers into promoters by analyzing this huge volume of opinions.

**Voice of the Market (VOM)**

Voice of the Market is about determining what customers are feeling about products or services of competitors. Accurate and timely information from the Voice of the Market helps in gaining competitive advantage and new product development.

Detection of such information as early as possible helps in direct and target key marketing campaigns. Sentiment Analysis helps corporate to get customer opinion in real time. This real-time information helps them to design new marketing strategies, improve product features and can predict chances of product failure. Zhang et al. proposed weakness finder system which can help manufacturers find their product weakness from Chinese reviews by using aspects based sentiment analysis.

**Voice of the Customer (VOC)**

Voice of the Customer is concern about what individual customer is saying about products or services. It means analyzing the reviews and feedback of the customers. VOC is a key element of Customer Experience Management. VOC helps in identifying new opportunities for product inventions. Extracting customer opinions also helps in identifying functional requirements of the products and some non- functional requirements like performance and cost.

**Brand Reputation Management**

Brand Reputation Management is concern about managing your reputation in market. Opinions from customers or any other parties can damage or enhance your reputation. Brand Reputation Management (BRM) is a product and company focused rather than customer. Now, one-to-many conversations are taking place online at a high rate. That creates opportunities for organizations to manage and strengthen brand reputation. Now Brand perception is determined not only by advertising, public relations and corporate messaging. Brands are now a sum of the conversations about them. Sentiment analysis helps in determining how companies brand, product or service is being perceived by community online.

**Government**

Sentiment analysis helps government in assessing their strength and weaknesses by analyzing opinions from public. For example, our PM enforced complete nationwide lockdown even when there was no outbreak in our country, kudos to our PM. this example clearly shows positive sentiment about government. Whether it is tracking citizen opinions on a new 108 systems, identifying strengths and weaknesses in a recruitment campaign in government job, assessing success of electronic submission of tax returns, or many other areas, we can see the potential for sentiment analysis.

### 

### Design and Implementations

**Proposed System**

1. An idea which can overcome the disadvantages being faced by traditional survey method to get people opinions, to develop a Machine Learning Model by training the model to categorize the tweets based on sentiment of the tweet and make the model as accurate as possible, first the user will give input i.e. the keyword for extracting the tweets and then the extracted tweets will be categorized by the Machine Learning Model which will be either positive or negative tweet and then the output will be displayed in graphical manner for better understanding of the results.
2. Another representation of the software is activity diagram which is a graphical representation of workflows of stepwise activities and actions with support for choice, iteration and concurrency. It visually presents a series of actions or flow of control in a system similar to flow chart or a data flow diagram. Figure 1 shows the step by step procedure of how the Machine Learning Model is built and trained for prediction
3. Figure 1: Activity diagram for our ML model
4. The Figure 2 shows step by step process of the end user side program
5. Figure 2: Activity diagram for the user side program

### Advantages of Proposed System:

There is no need to manually start a survey because in twitter there are already available tweets which are opinions of the people

There is no need to manually take tweets one by one.

The user just has to download the application. There is no external hardware components required.

**Aim of the proposed system:**

The application should be functionally competent as such that it must loaded with features that serve the purpose for which it is created. The features in the application should map to the needs of the users which the app is designed to meet. The application UI should be simple enough for a user to understand how the application works. The application should work successfully without crashing, and people all over the world should be able to use the application without any ambiguity.

**General Working of the System**

The application mainly consists of the following tasks

[1] Building and Training the Machine Learning Model

In this step the end user has nothing to do but this step would be done in the background the end user has no knowledge about this process

Fig.1. Giving the keyword as input

In this step the user has to give the keyword which should be present in all the tweets which we are going to extract from twitter

TABLE I. Preprocess the tweets extracted

In this step also the user has nothing to do but the extracted tweets has to be preprocessed before sending these tweets to the ML model

a. Getting the prediction of sentiment of extracted tweets using our Machine Learning Model

In this step our ML model predicts the sentiment of the tweets but the results will be stored in an array.

--Displaying the results in graphical representation

In this step the results will be displayed to the end user in graphical manner like bar graph or pie chart for better understanding of the results.

### Activity Diagrams

Another representation of the software is activity diagram which is a graphical representation of workflows of stepwise activities and actions with support for choice, iteration and concurrency. It visually presents a series of actions or flow of control in a system similar to flow chart or a data flow diagram.

**Figure 1** shows the step by step procedure of how the Machine Learning Model is built and trained for prediction

**Figure2** Activity diagram for the user side program

**Training Data**

**Preprocessing**

**Feature Extractor**

**Figure 1**

**Machine Learning Algorithm**

**Classifier Model**

**Label**

**Taking Input from User to** **extract tweets**

**Preprocessing the Extracted Tweets**

**Figure 2**

**Processing the tweets using NLP model**

**Output Format**

**If bar graph selected If pie Chart selected**

**Output in Pie Chart**

**Output in Bar Graph**

**Results**

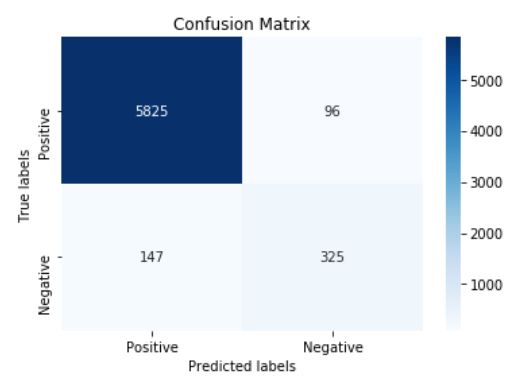
****

Fig 1 Confusion Matrix of NLP Model (82.9% Accuracy)

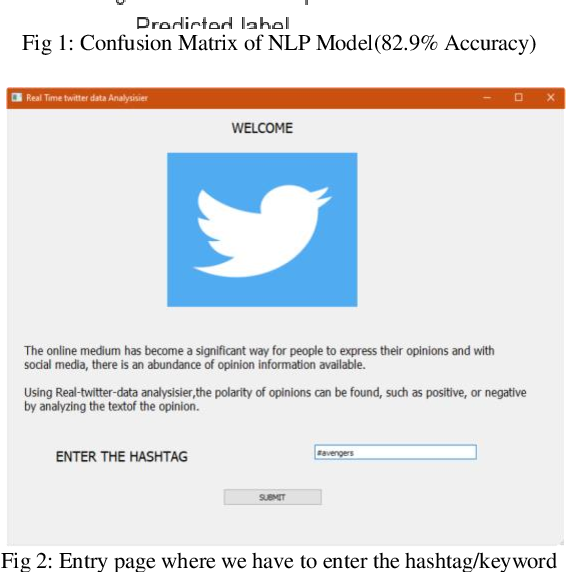
****

Fig 2 Entry page where we have to enter the hash tag/keyword

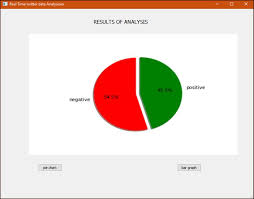
****

Fig2.1 Results are display in pie-chart

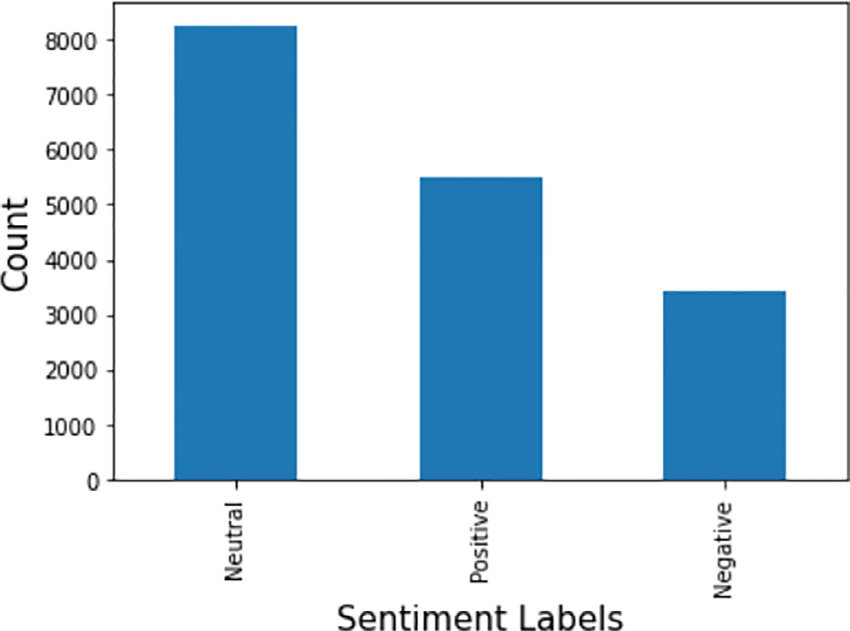
****

Fig2.2 Results are display in bar graph

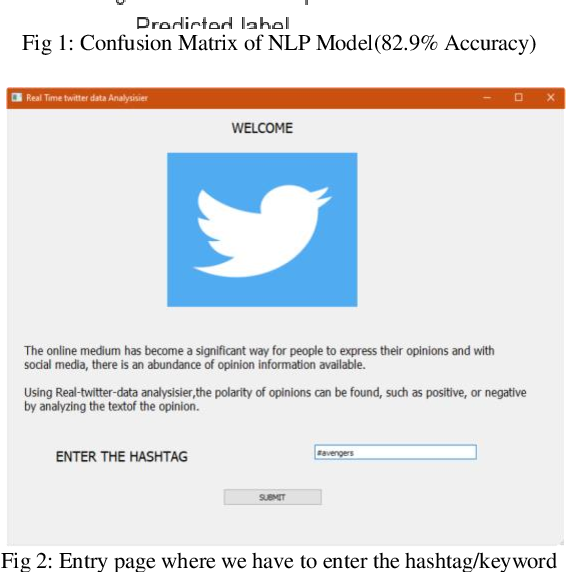
****

Fig3 Entry page where we have to enter the hash tag/keyword

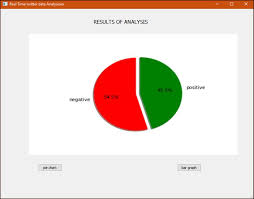
****

Fig3.1 Results are displayed in pie-chart

**##Description: This is a sentiment analysis program that parses the dataset of tweets fetched from Twitter using Natural Language Processing.**

import pandas as pd

import numpy as np

from string import punctuation

import re

import nltk

from nltk.corpus import twitter\_samples

import random

nltk.download('stopwords')

import string

from tensorflow.keras.preprocessing.text import Tokenizer

from nltk.corpus import stopwords

from nltk.stem import PorterStemmer

from nltk.tokenize import TweetTokenizer

from sklearn.preprocessing import LabelEncoder

import tensorflow as tf

from tensorflow import keras

from tensorflow.keras.models import Model

from tensorflow.keras.layers import Input, Embedding, LSTM, Dense,Dropout

from tensorflow.keras.preprocessing.sequence import pad\_sequences

from sklearn.feature\_extraction.text import TfidfVectorizer

import matplotlib.pyplot as plt

import seaborn as sns

**##Loading the data**

from google.colab import files

uploaded=files.upload()

tweets=pd.read\_csv("Twitter\_Data.csv")

tweets.head(10)

tweets['clean\_text']=tweets['clean\_text'].astype('str')

**##Dropping Nan Values**

tweets=tweets.dropna()

**##Data Visualization**

all\_positive\_tweets=tweets[tweets['category']==1]['clean\_text']

all\_neutral\_tweets=tweets[tweets['category']==0]['clean\_text']

all\_negative\_tweets=tweets[tweets['category']==-1]['clean\_text']

**##Interpreting nature of tweet with length of the tweet**

total\_positive\_words = []

for sentence in all\_positive\_tweets:

total\_positive\_words.append(sentence.count(' '))

total\_negative\_words = []

for sentence in all\_negative\_tweets:

total\_negative\_words.append(sentence.count(' '))

total\_neutral\_words = []

for sentence in all\_neutral\_tweets:

total\_neutral\_words.append(sentence.count(' '))

import plotly.graph\_objects as go

import numpy as np

x0 = np.array(total\_positive\_words)

x1 = np.array(total\_negative\_words)

x2 = np.array(total\_neutral\_words)

fig = go.Figure()

fig.add\_trace(go.Histogram(x=x0, name = 'Positive'))

fig.add\_trace(go.Histogram(x=x2, name = 'Neutral'))

**## Overlay both histograms**

fig.update\_layout(barmode='overlay')

# Reduce opacity to see both histograms

fig.update\_traces(opacity=0.75)

fig.show()

total\_negative\_words = []

for sentence in all\_negative\_tweets:

total\_negative\_words.append(sentence.count(' '))

total\_neutral\_words = []

for sentence in all\_neutral\_tweets:

total\_neutral\_words.append(sentence.count(' '))

import plotly.graph\_objects as go

import numpy as np

x0 = np.array(total\_positive\_words)

x1 = np.array(total\_negative\_words)

x2 = np.array(total\_neutral\_words)

fig = go.Figure()

fig.add\_trace(go.Histogram(x=x1, name = 'Negative'))

fig.add\_trace(go.Histogram(x=x2, name = 'Neutral'))

**## Overlay both histograms**

fig.update\_layout(barmode='overlay')

# Reduce opacity to see both histograms

fig.update\_traces(opacity=0.75)

fig.show()

total\_positive\_words = []

for sentence in all\_positive\_tweets:

total\_positive\_words.append(sentence.count(' '))

total\_negative\_words = []

for sentence in all\_negative\_tweets:

total\_negative\_words.append(sentence.count(' '))

total\_neutral\_words = []

for sentence in all\_neutral\_tweets:

total\_neutral\_words.append(sentence.count(' '))

import plotly.graph\_objects as go

import numpy as np

x0 = np.array(total\_positive\_words)

x1 = np.array(total\_negative\_words)

x2 = np.array(total\_neutral\_words)

fig = go.Figure()

fig.add\_trace(go.Histogram(x=x0, name = 'Positive'))

fig.add\_trace(go.Histogram(x=x1, name = 'Negative'))

**## Overlay both histograms**

fig.update\_layout(barmode='overlay')

# Reduce opacity to see both histograms

fig.update\_traces(opacity=0.75)

fig.show()

plt.figure(figsize=(8,10))

sns.countplot(x=tweets['category'])

from wordcloud import WordCloud

wc = WordCloud(width = 500, height = 500, min\_font\_size = 10, background\_color = 'white')

positive\_wc = wc.generate(tweets[tweets['category'] == 1.0]['clean\_text'].str.cat(sep = " "))

neutral\_wc = wc.generate(tweets[tweets['category'] == 0.0]['clean\_text'].str.cat(sep = " "))

negative\_wc = wc.generate(tweets[tweets['category'] == -1.0]['clean\_text'].str.cat(sep = " "))

plt.figure(figsize = (12, 12))

plt.imshow(positive\_wc)

plt.figure(figsize = (12, 12))

plt.imshow(neutral\_wc)

plt.figure(figsize = (12, 12))

plt.imshow(negative\_wc)

tweets.isnull().sum()

**##Removing StopWords and Punctuations**

Import the english stop words list from NLTK

stopwords\_english = stopwords.words('english')

print('Stop words\n')

print(stopwords\_english)

print('\nPunctuation\n')

print(string.punctuation)

tweets=tweets.dropna(axis=0)

tweets.isnull().sum()

**## Removing Stopwords**

tweets['clean\_text'] = tweets['clean\_text'].apply(lambda x: ' '.join([word for word in x.split() if word not in (stopwords\_english)]))

tweets['clean\_text'][0]

**## removing punctuations**

tweets['clean\_text'] = tweets['clean\_text'].apply(lambda x: re.sub(r'[^\w\s]', '', x))

tweets['clean\_text'][0]

**##Stemming The Words**

stemmer = PorterStemmer()

def stemming(word):

list1=[]

for i in word.split():

list1.append(stemmer.stem(i))

return ' '.join(list1)

tweets['clean\_text'] = tweets['clean\_text'].apply(lambda x:stemming(x))

tweets['category'] = [2 if x == -1 else x for x in tweets['category']]

tweets\_2=tweets.copy()

**##Tokenizing the words**

tokenizer = Tokenizer()

tokenizer.fit\_on\_texts(tweets.clean\_text)

word\_index = tokenizer.word\_index

vocab\_size = len(word\_index)+1

**##Padding the tweets**

**##padding the tokenized sequences to same length**

max\_length = 200

lines = pad\_sequences(tokenizer.texts\_to\_sequences(tweets.clean\_text),

maxlen = max\_length)

tweets.clean\_text = lines.tolist()

tweets.category.value\_counts()

tweets.head()

tweets.isnull().sum()

**##Splitting the data**

from sklearn.model\_selection import train\_test\_split

X\_train,X\_test,y\_train,y\_test = train\_test\_split(tweets['clean\_text'],tweets['category'],test\_size=0.2,random\_state=101)

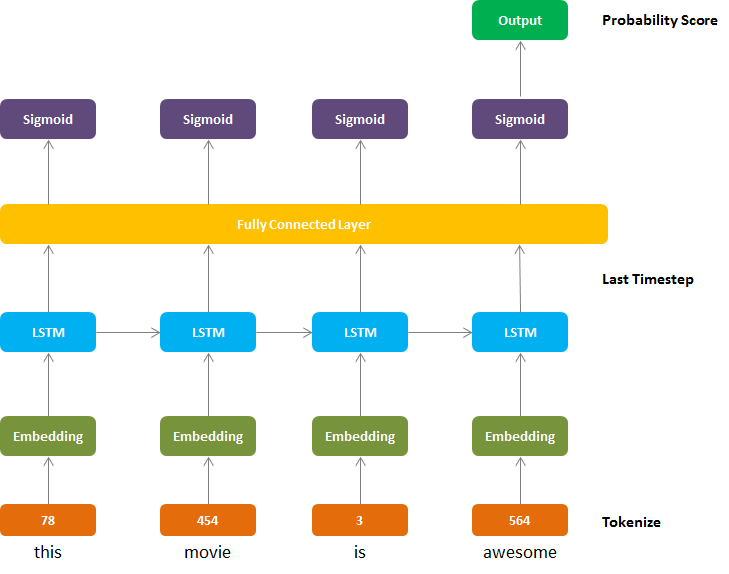
X\_train = np.vstack(X\_train.values)

y\_train = np.vstack(y\_train.values)

X\_val = np.vstack(X\_test.values)

y\_val = np.vstack(y\_test.values)

**##Data Modelling**



**##Sequential Modelling**

model = tf.keras.Sequential()

**#Input layer**

model.add(Input(shape=(None,)))

**#Embedding layer**

model.add(Embedding(input\_dim=vocab\_size,output\_dim=200,trainable=True))

**#LSTM layer**

model.add(LSTM(64, activation='relu'))

**#Fully connected layer**

model.add(Dense(64, activation='relu'))

model.add(Dropout(0.2))

**#Output layer**

model.add(Dense(3, activation='sigmoid'))

model.summary()

**##Adding Optimizer,loss function and Training the Model**

model.compile(optimizer='adam', loss=keras.losses.SparseCategoricalCrossentropy(from\_logits=True), metrics=['accuracy'])

history = model.fit(X\_train, y\_train,

epochs=2, batch\_size=512,

verbose=1,shuffle=True,validation\_data=(X\_val,y\_val))

**##Accuracy and Loss of Test data**

predictions = model.predict(X\_val)

print(history.history)

**##Model Evaluvation**

print("Evaluate on test data")

results = model.evaluate(X\_val, y\_val, batch\_size=128)

print("test loss, test acc:", results)

**CHALLENGES IN SENTIMENT ANALYSIS**

Sentiment Analysis is a very challenging task. Following are some of the challenges faced in Sentiment Analysis of Twitter.

1. **Identifying subjective parts of text:**
   * 1. Subjective parts represent sentiment-bearing content. The same word can be treated as subjective in one case, or an objective in some other. This makes it difficult to identify the subjective portions of text.
     2. For example: 1.The language of the Mr. Dennis was very crude. 2. Crude oil is obtained by extraction from the sea beds. The word “crude” is used as an opinion in first example, while it is completely objective in the second example.
2. **Domain dependence:**
   * 1. The same sentence or phrase can have different meanings in different domains. For Example, the word “unpredictable” is positive in the domain of movies, dramas etc, but if the same word is used in the context of a vehicle's steering, then it has a negative opinion.
3. **Sarcasm Detection:**
   * 1. Sarcastic sentences express negative opinion about a target using positive words in unique way.
     2. Example: “Nice perfume. You must shower in it.” The sentence contains only positive words but actually it expresses a negative sentiment.
4. **Thwarted expressions:** 
   * 1. There are some sentences in which only some part of text determines the overall polarity of the document.
     2. Example: “This Movie should be amazing. It sounds like a great plot, the popular actors, and the supporting cast is talented as well.”
     3. In this case, a simple bag of words approaches will term it as positive sentiment, but the ultimate sentiment is negative.
5. **Explicit Negation of sentiment:**
   * 1. Sentiment can be negated in many ways as opposed to using simple no, not, never, etc. It is difficult to identify such negations
     2. Example: “It avoids all suspense and predictability found in Hollywood movies.”
     3. Here the words suspense and predictable bear a negative sentiment, the usage of “avoids” negates their respective sentiments.
6. **Order dependence:** 
   * 1. Discourse Structure analysis is essential for Sentiment Analysis/Opinion Mining.
     2. Example: A is better than B, conveys the exact opposite opinion from, B is better than A.
7. **Entity Recognition:**
   * 1. There is a need to separate out the text about a specific entity and then analyze sentiment towards it.
     2. Example: “I hate Microsoft, but I like Linux”.
     3. A simple bag-of-words approach will label it as neutral however, it carries a specific sentiment for both the entities present in the statement.
8. **Building a classifier for subjective vs. objective tweets:**
   1. Current research work focuses mostly on classifying positive vs. negative correctly. There is a need to look at classifying tweets with sentiment vs. no sentiment closely.
9. **Handling comparisons:**
   1. Bag of words model doesn't handle comparisons very well.
   2. Example: “IIT’s are better than most of the private colleges”, the tweet would be considered positive for both IIT’s and private colleges using bag of words model because it doesn’t take into account the relation towards "better".
10. **Applying sentiment analysis to Facebook messages:**
    1. There has been less work on sentiment analysis on Facebook data mainly due to various restrictions by Facebook graph API and security policies in accessing data.
11. **Internationalization [16,17]:**
    1. Current Research work focus mainly on English content, but Twitter has many varied users from across.

**Chapter 5**

**CONCLUSION**

In this paper, we provide a survey and comparative study of existing techniques for opinion mining including machine learning and lexicon-based approaches, together with cross domain and cross-lingual methods and some evaluation metrics. Research results show that machine learning methods, such as Support Vector Machine (SVM) and naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require few efforts in human-labeled document .We also studied the effects of various features on classifier. We can conclude that more the cleaner data, more accurate results can be obtained. Use of bigram model provides better sentiment accuracy as compared to other models. We can focus on the study of combining machine learning method into opinion lexicon method in order to improve the accuracy of sentiment classification and adaptive capacity to variety of domains and different languages.

The task of sentiment analysis, especially in the domain of micro-blogging, is still in the developing stage and far from complete. Therefore we would like to propose a couple of ideas which we feel are worth exploring in the future and may result in further improved performance. For now, we have worked with only the very simplest unigram models we could improve those models by adding additional information like closeness of the word with a negation word. We could specify a window prior to the word (a window could for example be of 2 or 3 words) under consideration and the effect of negation may be incorporated into the model if it lies within that window.

The closer the negation word is to the unigram word whose prior polarity is to be calculated, the more it should affect the polarity. For example, if the negation is right next to the word, it may simply reverse the polarity of that word, and farther the negation is from the word the more minimized its effect should be. Apart from this, we are currently only focusing on unigrams and the effect of bigrams and trigrams may be explored. As reported in the literature review section when bigrams are used along with unigrams this usually enhances performance. However, for bigrams and trigrams to be an effective feature we need a much more labeled data set than our meager 9,000 tweets. So say instead of calculating a single probability for each word like P (word | obj) we could instead have multiple probabilities for each according to the Part of Speech the word belongs to used a somewhat similar approach and claims that appending POS information for every unigram results in no significant change in performance (with Naive Bayes performing slightly better and SVM having a slight decrease in performance), while there is a significant decrease in accuracy if only adjective unigrams are used as features. However, these results are for the classification of reviews and maybe verified for sentiment analysis on micro blogging websites like Twitter. One more feature we that is worth exploring is whether the information about the relative position of the word in a tweet has any effect on the performance of the classifier explored a similar feature and reported negative results, their results were based on reviews which are very different from tweets and they worked on an extremely simple model. In this project, we are focusing on general sentiment analysis.

For example, we noticed that users generally use specific types of keywords which can be divided into a couple of distinct classes, namely: media/movies/music, celebrities, products/brands, sports/sportsmen, politics/politicians. So we can attempt to perform separate sentiment analysis on tweets that only belong to one of these classes (i.e. the training data would not be general but specific to one of these categories) and compare the results we get if we apply general sentiment analysis on it instead.

**FUTURE WORKS**

Currently, this project is done using the Naive Bayes Algorithm which is one of the Machine Learning Algorithm which only got us an accuracy of around 85%. In the future, we will be exploring and implementing the Deep Learning Algorithms to our NLP model in order to increase the accuracy of our model and to get better predictions from our model.

**Weekly Task**

The report of project work allocated by the supervisor is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week No.** | **Date:**  **From-To** | **Work Allocated** | **Work Completed**  **(Yes/No)** | **Remarks** | **Guide Signature** |
| 1 | 24-04-23 to 29-04-23 | Information Gathering, Make the Dataset | Yes |  |  |
| 2 | 01-05-23 to 06-05-23 | Methods used for Pre-processing | Yes |  |  |
| 3 | 08-05-23 to 13-05-23 | Pre-processing of Dataset | Yes |  |  |
| 4 | 15-05-23 to 20-05-23 | Make report | Yes |  |  |

**Refrences**

[1] Efthymios Kouloumpis and Johanna Moore,IJCSI International Journal of Computer Science Issues.

[2] S. Batra and D. Rao, “Entity Based Sentiment Analysis on Twitter”, Stanford University.

[3] Saif M.Mohammad and Xiaodan zhu , Sentiment Analysis on of social media texts.

[4] Ekaterina kochmar, University of Cambridge, at the Cambridge coding Academy Data Science.

[5] Manju Venugopalan and Deepa Gupta, Exploring Sentiment Analysis on Twitter Data.

[6] Brett Duncan and Yanqing Zhang, Neural Networks for Sentiment Analysis on Twitter.

[7] Afroze Ibrahim Baqapuri, Twitter Sentiment Analysis: The Good the Bad and the OMG! Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media.

[8] Kishori K. Pawar, Pukhraj P Shrishrimal, R. R. Deshmukh,” Twitter Sentiment Analysis: A Review” International Journal of Scientific & Engineering Research.

[9] [www.kaggle.com](http://www.kaggle.com) --Kaggle