## MINI PROJECT REPORT

**on**

**Twitter Data Analysis**

**Submitted in partial fulfilment for the completion of**

**BE-VI Semester**

**In**

**INFORMATION TECHNOLOGY**

**By**

**Mohammed Abdul Mateen (160118737023)**

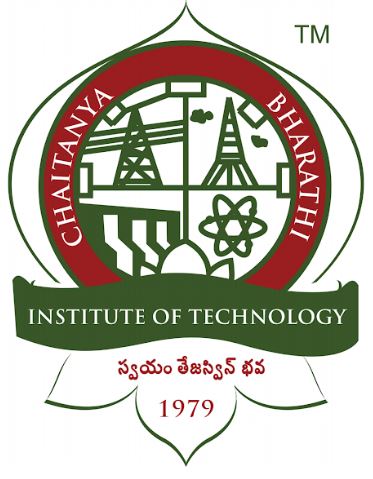
**Shaik Abdullah Adnan (160118737024)**

**Under the guidance of**

**Dr. K. Radhika**

**Professor, HoD**

**Dept. of IT, CBIT.**

****

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY (A)**

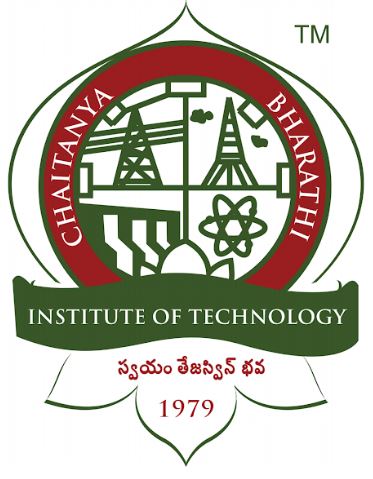
**(Affiliated to Osmania University; Accredited by NBA(AICTE) and NAAC(UGC), ISO Certified 9001:2015)**

**GANDIPET, HYDERABAD – 500 075**

**Website:** [**www.cbit.ac.in**](http://www.cbit.ac.in)

**2020-2021**

****

****

**CERTIFICATE**

This isto certify that the project work entitled “**Twitter Data Analysis**” submitted to **CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY,** in partial fulfilment of the requirements for the award of the completion of VI semester of B.E in Information Technology, during the academic year 2020-2021, is a record of original work done by **Mohammed Abdul Mateen (160118737023), Shaik Abdullah Adnan (160118737024)** during the period of study in Department of IT, CBIT, HYDERABAD, under our supervision and guidance.

**Project Guide**  **Head of the Department**

**Dr.K. Radhika**

Professor, Dept. of IT, Professor, Dept. of IT,

CBIT, Hyderabad. CBIT, Hyderabad.

**DECLARATION**

It is our privilege to acknowledge with deep sense of gratitude and devotion for keen personal interest and invaluable guidance rendered by our Project Guide, Dr. K. Radhika, Head of Department of Information Technology, Chaitanya Bharathi Institute of Technology.

We are grateful to our Principal Dr. G. P. S. Verma, Chaitanya Bharathi Institute of Technology, for his cooperation and encouragement. Finally, we also thank all the staff members, faculty of Dept. of IT, CBIT, and our parents who with their valuable suggestions and support, directly or indirectly helped us in completing this project work.

Our thanks to all members of the staff and our lab assistants for helping us to carry out the groundwork of this project.

We also take this opportunity to thank our parents for their support to complete the project.

Mohammed Abdul Mateen (160118737023)

Shaik Abdullah Adnan (160118737024)

**ABSTRACT**

**Data analytics** enables both prediction and knowledge discovery capabilities. It helps in understanding the current state of the business or any process and provides a solid foundation to predict future outcomes. It has made a huge impact in the industry in terms of growth and providing optimal solutions.

A very well-known social media platform, **Twitter** has become a platform for many people to express their views for a particular topic.

Now, imagine analysing data from twitter platform. If used in the right way, **KNOWLEDGE** can be extracted from this pile of data and hence one can arrive for better optimal solutions, be it in any field.

This project mainly focuses on analysing the data in different ways using different methods and create some functions which give us knowledge about the data.

When **python** programming language with its libraries is integrated with **machine learning algorithms**, this work becomes even more easier.

Hence, using python as the language to code, machine learning algorithms and the **Twitter API**, this mini project provides a platform which gives an insight of the data analysed. Some of the functionalities of the project include **emotion analysis, analysing the data, visualising the data, etc**.

**TABLE OF CONTENTS**

**Page No.**

**DECLARATION iii**

**ABSTRACT iv**

**LIST OF FIGURES vi**

**1. INTRODUCTION 7**

**1.1 INTRODUCTION 7**

**1.2 PROBLEM STATEMENT 8**

**2. LITERATURE SURVEY 9**

**3. SYSTEM DESIGN/METHODOLOGY 11**

**3.1 ARCHITECHTURE OF SYSTEM 11**

**3.2 REQUIREMENTS 13**

**3.3 METHODOLOGY 13**

**3.4 SOFTWARE DESCRIPTION 14**

**3.5 DATA ANALYSIS AND VISUALIZATION 14**

**4. IMPLEMENTATION 15**

**5. TESTING & RESULTS 20**

**6. CONCLUSION & FUTURE SCOPE 24**

**BIBLIOGRAPHY 25**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Figure | Description | Page No. |
| Figure 3.1 | Flow chart of architecture | 11 |
| Figure 3.2 | Flow chart | 12 |
| Figure 3.3 | Flow chart 2 | 13 |
| Figure 4.1 | Packages importing | 15 |
| Figure 4.2 | Authenticating the Twitter API | 15 |
| Figure 4.3 | Contents retrieved from tweets | 16 |
| Figure 4.4 | Duplicates removed & data saved | 16 |
| Figure 4.5 | Tweets Vs Device used graph | 17 |
| Figure 4.6 | Doing sentiment analysis | 17 |
| Figure 4.7 | Visualization of tweets | 18 |
| Figure 4.8 | Tokenization of tweet data | 18 |
| Figure 4.9 | Cloud format of data | 19 |
| Figure 5.1 | Data of 10 tweets | 20 |
| Figure 5.2 | Score calculation for each tweet | 21 |
| Figure 5.3 | Plot for better visualization | 22 |
| Figure 5.4 | Pie-chart visualization of score | 23 |

1. **INTRODUCTION**

**1.1 INTRODUCTION**

In the past few years, there has been a huge growth in the use of microblogging platforms such as Twitter. Spurred by that growth, companies and media organizations are increasingly seeking ways to mine Twitter for information about what people think and feel about their products and services. Companies such as Twitratr (twitrratr.com), tweetfeel (www.tweetfeel.com), and Social Mention (www.socialmention.com) are just a few who advertise Twitter emotion analysis as one of their services.

While there has been a fair amount of research on how emotions are expressed in genres such as online reviews and news articles, how emotions are expressed given the informal language and message-length constraints of microblogging has been much less studied. Features such as automatic part-of-speech tags and resources such as emotion lexicons have proved useful for emotion analysis in other domains, but will they also prove useful for emotion analysis in Twitter? This document answers to this question.

Another challenge of microblogging is the incredible breadth of topic that is covered. It is not an exaggeration to say that people tweet about anything and everything. Therefore, to be able to build systems to mine Twitter emotion about any given topic, we need a method for quickly identifying data that can be used for training. In this paper, we explore one method for building such data: using Twitter hashtags (e.g., #bestfeeling, #epicfail, #news), targeting specific twitter account (e.g. @Vice President Kamala Harris, @Joe Biden) to identify positive, negative, and neutral tweets to use for training three-way emotion classifiers.

The online medium has become a significant way for people to express their opinions and with social media, there is an abundance of opinion information available. Using twitter data analysis (emotion analysis), the polarity of opinions can be found, such as positive, negative, or neutral by analysing the text of the opinion. Sentiment analysis has been useful for companies to get their customer's opinions on their products predicting outcomes of elections and getting opinions from movie reviews. The information gained from sentiment analysis is useful for companies making future decisions.

Many traditional approaches in sentiment analysis uses the bag of words method. The bag of words technique does not consider language morphology, and it could incorrectly classify two phrases of having the same meaning because it could have the same bag of words. The relationship between the collection of words is considered instead of the relationship between individual words. When determining the overall sentiment, the sentiment of each word is determined and combined using a function. Bag of words also ignores word order, which leads to phrases with negation in them to be incorrectly classified.

**1.2 PROBLEM STATEMENT**

The main concept around which this mini project revolves is gaining knowledge from the data we have. A huge amount of data is available on the internet. With an exponential growth in technology and gadgets around us, the data which these technologies undergo have also increased. Now, the question is what can be done with this load of data generated. And furthermore, the problem arises in answering the question how this data is to be utilised. This mini project is an answer to these questions. The motive is simple to understand.

We use the twitter social media platform to extract the data available there, process the data and then come up with useful results after working on it. Data Analysis has made a huge impact in the market in terms of growth of an organisation or any company. Proper decisions can be made which is helpful for the company in many of its operations and hence it is an added advantage.

The data is processed, cleaned, various algorithms and tools are applied on the data and an analysis is made taking into consideration different things.

Few of the modules include:

* Analysing a particular users account.
* Analysing a particular topic which is trending at that specific time.
* Visualising this data in different ways to gain more knowledge, etc.

1. **LITERATURE SURVEY**

Sentiment analysis is the process of analysis of the text from many levels. First level is document level, the classification task determine the class of an object based on its attributes (Turney, 2002; Pang and Lee, 2004), and after that it can analysed at the sentence level[5] for classifying the sentence based on the negative, positive and neutral sentiments (Hu and Liu, 2004; Kim and Hovy 2004) and next level is the phrase level[4] for defining if an expression is unbiassed or polar and then remove uncertainty of Twitter Sentimental Analysis 344 Published By: Blue Eyes Intelligence Engineering Retrieval Number: E1989017519 & Sciences Publication meaning from the polarity of the polar expressions (Wilson et al., 2005; Agarwal et al., 2009. Berming ham and Smeaton (2010) and Pak and Paroubek (2010). Go et al. (2009) they used distant learning algorithm to obtain the sentiment data [8]. In this techniques, positive emoticons symbols in tweets such as “:)” “:-)”and negative emoticons symbols in tweets such as like “:(” “:-(”. They proposed the models using Naive Bayes algorithm for analysis the text and the report are generated and visualized.

They used two methods such as unigrams for identifying single word repeating over the context and bigrams for identifying double word repeating over the context along with Parts-of-Speech (POS) for analysing the tweets. But the unigram method had reached a better way of analysis, but the bigrams and POS had failed to attempt his purpose.

Pak and Paroubek (2010) collect the following tweets considered as data which really helped them in similar distant learning paradigm for setting a model for analysis. They perform classification of task such as subjective, objective. For subjective the information is taken from the user tweets by means of text or image or symbols as Go et al. (2009). For objective information the information is obtained from verification of the data such as famous newspapers like “Times of India”, “Washington Posts” etc. Information which is taken for analysis is casual sample of flowing tweets collected by using queries. In the past year there have been numerous documents observing the Twitter sentiment and buzz (Jansen et al. 2009; Pak and Paroubek 2010; O’Connor et al. 2010; Tumasjan et al. 2010; Bifet and Frank 2010; Barbosa and Feng 2010; Davidov, Tsur, and Rappoport 2010). Further scientists have started to discover the usage of part-of-speech structures, but results remain mixed. It has enormous, interesting chances to develop the innovative applications, because success of many business depends on accessible information on online sources such as blogs, twitter and other social networks.

Barbosa and Feng (2010) has analysed the sentimental classification on Twitter data. The test data of tweets are collected, they have taken some of the syntax features for analysis of tweets which contains symbols, retweet, emoticons, tags, link, punctuation and exclamation marks, semicolon are in the combination with structures for identifying the polarity of words.

Kamps et al. (2002) has analysed the data by using the lexical database. Lexical database is description of lexemes. Lexical database such as WordNet are used. This contains the emotional content of a word. The distance metric of words are used to determined semantic polarity of adjectives.

Researchers are also trying to find different ways of analysing tweets based on the ideas they had while understanding the concept. Researchers tried this analysis using some of the specified fields such as Machine learning which uses Naive Bayes, Maximum entropy and SVM alongside the Semantic Orientation based Word Net which extracts equivalent words and similitude for the content feature, then Lexicon based analysis based on the created dataset which consists of pre-processed tweets and lastly, Hybrid approach where some researchers combined the supervised machine learning and lexicon based approaches together to improve sentiment classification performance.

Gamon (2004) has done sentiment analysis on feedback data from the Global Support Services survey. They are used query to identify the role of features like Part of Speech tags. The accuracy of classifier can be obtained by some of factors such as feature selection, from the testing data and demonstrate the abstract linguistic analysis feature for accuracy of data.

Devakip, et al (2017) has done analysis on twitter data for election. It indicates the popularity of parties in the election based on positive tweets. This system uses Naïve Bayes classifier algorithm are used to classify the positive and negative tweets.

A comparative study of existing techniques for mining the data which includes machine learning, Interdependent Latent Dirichlet Allocation, lexicon-based approaches, together with cross domain, cross-lingual methods and some evaluation metrics. The concept level sentences analysis uses the Combining Lexicon and Learning based Approach. As the result of study, machine learning methods such as Support Vector Machine and Naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are amazingly effective in some cases.

More research is needed to determine whether the POS features are just of poor quality due to the results of the tagger or whether POS features are just less useful for sentiment analysis in this domain. Features from an existing sentiment lexicon were somewhat useful in conjunction with microblogging features, but the microblogging features (i.e., the presence of intensifiers and positive/negative/neutral emoticons and abbreviations) were clearly the most useful. In this paper, we perform extensive feature analysis of tweets using hashtags, ID’s and building model classifications.

1. **SYSTEM DESIGN/METHODOLOGY**

**3.1 ARCHITECHTURE OF THE SYSTEM:**

Chart, diagram

Description automatically generated

Fig.3.1 Flow chart of architecture

The system architecture consists of the components as shown in the figure 1.1 such as Tweets extraction from twitter, pre-processing of data, feature extraction, Training set are defined for the given analysis. The training set is obtained by predefined set of positive or negative tweets which can be done using naive Bayes or support vector machine (SVM) and output obtained is positive, negative tweets. The Classifier will classify the tweets according the training set and regulates the polarity of the tweet as the output.

Another image related to the system architecture would help in understanding it better. Hence a few more pictures different from the above one will give more clarity on it.

Diagram

Description automatically generated

Fig.3.2 Flow chart

**Feature Extraction**

The pre-processed dataset has many distinctive properties. In the feature extraction method, we extract the aspects from the processed dataset. Later this aspect are used to compute the positive and negative polarity in a sentence which is useful for determining the opinion of the individuals using models like unigram, bigram.

Machine learning techniques require representing the key features of text or documents for processing. These key features are considered as feature vectors which are used for the classification task. Some examples feature that have been reported in literature are:

1. **Words and Their Frequencies:** Unigrams, bigrams and n-gram models with their frequency counts are considered as features. There has been more research on using word presence rather than frequencies to better describe this feature. Panget al. showed better results by using presence instead of frequencies.

2. **Parts Of Speech Tags** Parts of speech like adjectives, adverbs and somegroups of verbs and nouns are good indicators of subjectivity and sentiment. We can generate syntactic dependency patterns by parsing or dependency trees.

3. **Opinion Words And Phrases** Apart from specific words, some phrases and idioms which convey sentiments can be used as features. e.g. cost someone an arm and leg.

4. **Position Of Terms** The position of a term with in a text can affect on how much the term makes difference in overall sentiment of the text.

5. **Negation** Negation is an important but difficult feature to interpret. The presence of a negation usually changes the polarity of the opinion.. International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016 8 e.g., I am not happy.

6. **Syntax** Syntactic patterns like collocations are used as features to learn subjectivity patterns by many of the researchers.

**3.2 REQUIREMENTS**

Few of the requirements which a person must have in hand in order to have a hands on experience of it are as follows. These include the programming languages used, the IDE platform where it is coded and other such related things:

* Python installed on the desktop on whics the programming is being done.
* Jupyter / Google Colab / Pycharm or other such related IDEs to code.
* Twitter developer account for accessing the twitter API and retrieving the tweets.
* Data visualisation libraries installed for better knowledge extraction.

**3.3 METHODOLOGY**

In this method we uses textblob as a method to find the polarity of the text (positive text, negative text or neutral text). The tweets are imported from the Twitter using the (API) provided by the Twitter Developer. From these API various fields like tweets, source, retweets, likes, language, user etc. can be scrapped. After collecting these data, we can analyse the various famous person thoughts on an event or occasion

Diagram

Description automatically generated

Fig.3.3 Flow chart 2

The figure explains the extraction of tweets id from twitter through API, then pre-process the data that are extracted. Pre-processing includes exclusion of unwanted fields, segregating the fields important for analysis. Once the fields are extracted and segregated CSV is created. Using this CSV, the length of the message, Likes, retweets for the id is extracted and various results are derived. With the scraped tweets, classify the tweets whether positive or negative or neutral.

**3.4 SOFTWARE DESCRIPTION**

In the system the graphs such as Table, Bar graph, Line graph are generated with the help of Jupyter notebook. The predefined functions are pandas, numpy, matplotlib, pyplot, list, Dictionary. Pandas is used for converting from csv file to dataset. Numpy is one of the essential library for scientific calculating in Python. It delivers a high-performance multidimensional array object, and apparatuses for experementing with these arrays. Python comprises of numerous built-in container categories: lists, dictionaries, sets, and tuples. A list is the Python equal of an array, but is resizable and can contain elements of different types. A dictionary stores (key, value) pairs, like a Map in Java or an object in JavaScript. Python library such as Text Blobare used for processing the textual data. It provides API for processing natural language processing (NLP) such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. Tweepy is used for accessing Twitter API and it is open sourced.

**3.5 DATA ANALYSIS AND VISUALIZATION**

In Twitter the various famous personalities tweet their thoughts on their opinion on an occasion. From their thoughts, importance of that occasion and the polarity of their tweet are analysed. Some of the analysis with the dataset as follows.

• Calculate the average length of the tweet and visualize the average length for a period.

• Visualize the favourites and retweets for each personality.

• Visualize the various source of the tweet.

• Calculate the polarity of the tweets given by the person

• Visualize the Polarity of tweet (positive, negative, neutral)

• Compare the tweets polarity of various famous personalities

1. **IMPLEMENTATION**

To understand the implementation in the best possible way, some of the screenshots of the code snippets are attached and the description/function related to the code snippet is explained. Note that the whole of the code is not explained here. Only an important part of which will give a good overview of the implementation.

**Graphical user interface, text, application

Description automatically generated**

Fig.4.1 Importing all the required libraries and downloading tweepy to work with twitter.

**Graphical user interface, text, application, email

Description automatically generated**

Fig.4.2 Authenticating the twitter API

**Text

Description automatically generated**

Fig.4.3 Contents retrieved from tweets

The data which is retrieved is not just the tweet content. Apart from the tweet description a lot of other data is also extracted from the twitter API. The figure above classifies shows and classifies the data into different categories.

**Graphical user interface, text, application

Description automatically generated**

Fig.4.4 Duplicates removed & data saved

Saving this data into a csv file and then working on that data.

Shape

Description automatically generated

Fig.4.5 tweets vs device graph

A graph showing from which device the user has tweeted. Ex: Android phone, iPhone, iPad, etc.

Graphical user interface, text

Description automatically generated

Fig.4.6 Doing Sentiment analysis

Chart, bar chart

Description automatically generated

Fig.4.7 Visualization of tweets

The above figure shows the visualization of tweets whose sentiment was analysed before. This helps us in better understanding.

Chart, histogram

Description automatically generated

Fig.4.8 Tokenization of tweet data

Tokenising the tweet data, i.e. extracting each and every word from the tweet and assigning a token to it. Later all such tokens are taken into consideration and an analysis is done. The graph above visualises the tokens which have been studied. It shows the count of words which have been used the most in the tweets which have been retrieved.

A picture containing text, newspaper

Description automatically generated

Fig.4.9 Cloud format of data

Visualising the data of the tweets in a word cloud format. The words that are seen in this word cloud are the most used words by users in their tweets.

**Note:** all the graphs shown above show the results when **“#covid19”** is taken into consideration

1. **TESTING & RESULTS**

Here we will see a set of few more images which shows us the results when some of the tweets are analysed.

Table

Description automatically generated

Fig.5.1 Data of 10 tweets

This image shows the output when we retrieve the latest 10 tweets from the official twitter account of Kamala Harris.

Now these tweets are cleaned and the noise is removed so that a good study can be done. Here, the polarity and subjectivity of the tweets are taken into consideration for determining the sentiment of these tweets.

**Polarity** is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement.

**Subjective** sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1].

Graphical user interface, text, application

Description automatically generated

Fig.5.2 Score is calculated for each Tweet

Based on that a **score** is calculated which is either positive, negative or neutral as shown in the picture.

Chart, scatter chart

Description automatically generated

Fig.5.3 Plot for better visualization

Now, in the above figure we can generate a plot using the matplot library which is available in python. Using this plot we have better view or better visualization of the data (tweets) and every classifier has different colour.

A similar analysis was done on few of the tweets from Donald Trump’s twitter handle and the results are shown below:

Chart, pie chart

Description automatically generated

Fig.5.4 Pie-chart visualization of the score

It should be clear that the examples shown above are not the only topics on which analysis can be done. Apart from this, any random person’s account or any random topic or hashtag can be taken into consideration and analysis can be made for the same. A few were shown above just as an example.

1. **CONCLUSION & FUTURE SCOPE**

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on analysing the sentiments of the tweets and feeding the data to a machine learning model to train it and then check its accuracy, so that we can use this model for future use according to the results. It comprises of steps like data collection, text pre-processing, sentiment detection, sentiment classification, training and testing the model. This research topic has evolved during the last decade with models reaching the efficiency of almost 85%-90%. But it still lacks the dimension of diversity in the data. Along with this it has a lot of application issues with the slang used and the short forms of words. Many analysers don’t perform well when the number of classes are increased. Also, it’s still not tested that how accurate the model will be for topics other than the one in consideration. Hence sentiment analysis has a very bright scope of development in future.

**BIBLIOGRAPHY**

This project mainly revolves around the data which has been retrieved from Twitter. However to mention a few other resources which came into use during the course of the project are listed below:

* Data: **Official Twitter API**
* Literature survey: articles published on the internet.
* Debugging and solving the errors in code: Websites like Stack Overflow, w3 Schools, Tutorials Point gave a lot of information about different libraries that could be used for analysing and visualising the data.
* Few datasets for testing various libraries for analysis and visualisation of data before applying them on actual data were taken from websites like: **Kaggle, Google Dataset Search, Github**. (Note: The actual data on which the analysis has been done has been exclusively retrieved from twitter using the official Twitter API.)