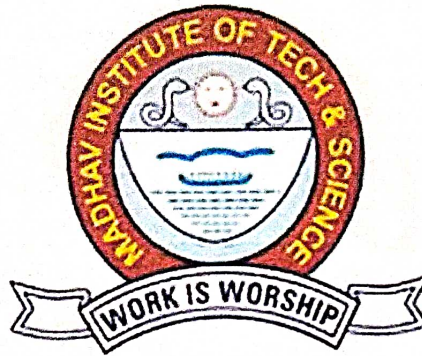


MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)



Project Report

on

Model of Twitter Data analysis using Machine learning

A project report submitted in partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by:

Geeteshwari Pal

0901cs191038

Faculty Mentor:

Mr. Mir Shahnawaz Ahmad

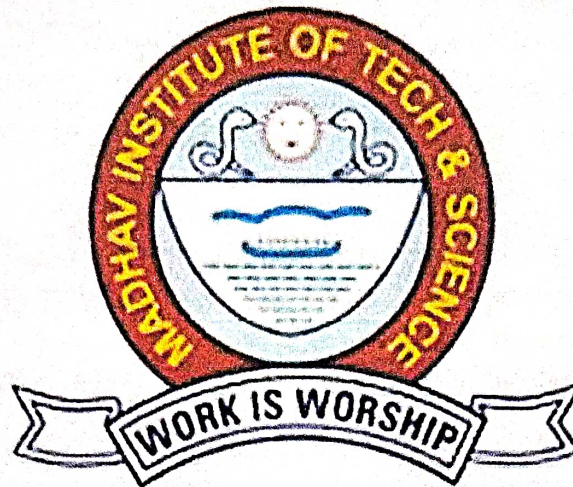
Assistant Professor, Computer Science and Engineering

Submitted to:

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

MAY-JUNE 2022

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)



Project Report

on

Model of Twitter Data analysis using Machine learning

Submitted By:

Geeteshwari Pal

0901cs191038

Faculty Mentor:

Mr. Mir Shahnawaz Ahmad

Assistant Professor, Computer Science and Engineering

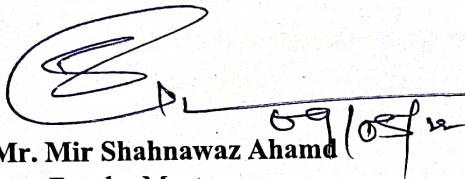
DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

MAY-JUNE 2022

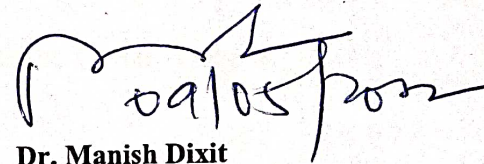
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Geeteshwari Pal** (0901cs191038) has submitted the project report titled Model of Twitter Data analysis using Machine learning under the mentorship of **Mr. Mir Shahnawaz Ahmad**, in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior.



Mr. Mir Shahnawaz Ahmad
Faculty Mentor
Assistant professor
Computer Science and Engineering



Dr. Manish Dixit
Professor and Head,
Computer Science and Engineering
Dr. Manish Dixit
Professor & HOD
Department of CSE
M.I.T.S. Gwalior

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Mr. Mir Shahnawaz Ahmad, Assistant Professor, Computer Science and Engineering.**

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.



Geeteshwari Pal

0901cs191038

3rd Year,

Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Computer Science and Engineering**, for allowing me to explore this project. I humbly thank **Dr. Manish Dixit**, Professor and Head, Department of Computer Science and Engineering, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Mr. Mir Shahnawaz Ahmad**, Assistant professor, Computer science and engineering, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.



Geeteshwari Pal
0901CS191038

3rd Year,

Computer Science and Engineering

ABSTRACT

This project addresses the problem of sentiment analysis in twitter; that is classifying tweets according to the sentiment expressed in them: positive, negative or neutral. Twitter is an online micro-blogging and social-networking platform which allows users to write short status updates of maximum length 140 characters. It is a rapidly expanding service with over 200 million registered users - out of which 100 million are active users and half of them log on twitter on a daily basis - generating nearly 250 million tweets per day. Due to this large amount of usage we hope to achieve a reflection of public sentiment by analyzing the sentiments expressed in the tweets. Analyzing the public sentiment is important for many applications such as firms trying to find out the response of their products in the market, predicting political elections and predicting socioeconomic phenomena like stock exchange. The aim of this project is to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream. The widespread and different types of information on Twitter make it one of the most appropriate virtual environments for information monitoring and tracking. In this paper, the authors review different information analysis techniques; starting with the analysis of different hashtags, twitter's network-topology, event spread over the network, identification of influence, and finally analysis of sentiment. Future research and development work will be addressed.

Key words: Big data, data analysis, social media, Twitter.

सार :

यह परियोजना ट्विटर में भावना विश्लेषण की समस्या का समाधान करती है; जो उनमें व्यक्त भावना के अनुसार ट्विट्स को वर्गीकृत कर रहा है: सकारात्मक, नकारात्मक या तटस्थ। ट्विटर एक ऑनलाइन माइक्रो-ब्लॉगिंग और सोशल-नेटवर्किंग प्लेटफॉर्म है जो उपयोगकर्ताओं को अधिकतम 140 अक्षरों की शॉर्ट स्टेटस अपडेट लिखने की अनुमति देता है। यह 200 मिलियन से अधिक पंजीकृत उपयोगकर्ताओं के साथ तेजी से विस्तार करने वाली सेवा है - जिसमें से 100 मिलियन सक्रिय उपयोगकर्ता हैं और उनमें से आधे दैनिक आधार पर ट्विटर पर लॉग इन करते हैं - प्रति दिन लगभग 250 मिलियन ट्वीट उत्पन्न करते हैं। इस बड़ी मात्रा में उपयोग के कारण हम ट्विट्स में व्यक्त भावनाओं का विश्लेषण करके जनता की भावनाओं का प्रतिबिंब प्राप्त करने की उम्मीद करते हैं। कई अनुप्रयोगों के लिए सार्वजनिक भावना का विश्लेषण करना महत्वपूर्ण है जैसे कि फर्म बाजार में अपने उत्पादों की प्रतिक्रिया का पता लगाने की कोशिश कर रहे हैं, राजनीतिक चुनावों की भविष्यवाणी कर रहे हैं और स्टॉक एक्सचेंज जैसी सामाजिक आर्थिक घटनाओं की भविष्यवाणी कर रहे हैं। इस परियोजना का उद्देश्य एक अज्ञात ट्वीट स्ट्रीम के सटीक और स्वचालित भावना वर्गीकरण के लिए एक कार्यात्मक क्लासिफायर विकसित करना है। ट्विटर पर व्यापक और विभिन्न प्रकार की जानकारी इसे सूचना निगरानी और ट्रैकिंग के लिए सबसे उपयुक्त आभासी वातावरण में से एक बनाती है। इस पत्र में, लेखक विभिन्न सूचना विश्लेषण तकनीकों की समीक्षा करते हैं; विभिन्न हैशटैग, ट्विटर के नेटवर्क-टोपोलॉजी, नेटवर्क पर फैली घटना, प्रभाव की पहचान, और अंत में भावना के विश्लेषण के विश्लेषण के साथ शुरू। भविष्य के अनुसंधान और विकास कार्यों को संबोधित किया जाएगा।

मुख्य शब्द: बड़ा डेटा, डेटा विश्लेषण, सोशल मीडिया, ट्विटर।

Table of Content

Title	Page No
<i>Abstract</i>	<i>iv</i>
सार	<i>v</i>
Chapter 1: Project Overview	
1.1 Introduction	1
1.2 Requirements for Report Writing:	1
1.2.1 Hardware requirements	1
1.2.2 Software requirements	1
1.3 Objective	1
1.4 System scope	2
1.5 Report & Analysis	2
1.6 System Context	2
1.7 System Function	2
Chapter 2 : Literature review	
2.1 Datasets	3
2.2 Data Retrieval	3
2.3 Ranking and Classifying twitter users	3
2.4 Homophily	4
2.5 Reciprocity	4
Chapter 3 : Requirements	
3.1 Functional Requirements	5
3.2 Non-Functional Requirements	5
3.3 Product Requirements	5
3.4 Organizational Requirements	6
Chapter 4 : Sentiment Analysis	
4.1 Natural Language Processing Approach	7
4.2 Machine Learning Approach	7
4.2.1 Naïve Bayes	7
4.2.1.1 Application of Naïve Bayes	8
4.3 Challenges	8
4.4 Conclusion & Future Scope	8
Working Model's screenshots	9
Conclusion	12
References	12

Chapter 1: PROJECT OVERVIEW

1.1. Introduction

We have chosen to work with twitter since we feel it is a better approximation of public sentiment as opposed to conventional internet articles and web blogs. The reason is that the amount of relevant data is much larger for twitter, as compared to traditional blogging sites. Moreover, the response on twitter is prompter and also more general (since the number of users who tweet is substantially more than those who write web blogs on a daily basis). Sentiment analysis of public is highly critical in macro-scale socioeconomic phenomena like predicting the stock market rate of a particular firm. This could be done by analyzing overall public sentiment towards that firm with respect to time and using economics tools for finding the correlation between public sentiment and the firm's stock market value. Firms can also estimate how well their product is responding in the market, which areas of the market is it having a favorable response and in which a negative response (since twitter allows us to download stream of geo-tagged tweets for particular locations. If firms can get this information, they can analyze the reasons behind geographically differentiated response, and so they can market their product in a more optimized manner by looking for appropriate solutions like creating suitable market segments. Predicting the results of popular political elections and polls is also an emerging application to sentiment analysis. One such study was conducted by Timespan et al. in Germany for predicting the outcome of federal elections in which concluded that twitter is a good reflection of offline sentiment.

1.2. Requirements for Report Writing:

1.2.1 Hardware Requirements

CPU/GPU	Intel i3 dual core
RAM	512 Mb (Recommended)
Memory	(Depends on amount of data) 15gb (Recommended)

1.2.2 Software Requirements

OS	Windows 7/Linux 5.4/ Mac
Language	Python 3.10.1
Platform	Jupyter Notebook and pip for package installation

1.3. Objective:

The aim of this project is to detect hate speech in tweets i.e. hate speech contains racist or sexist sentiment associated with it. Whole project can be broken down in two major parts:

- Text Preprocessing
- Data Exploration
- Feature Extraction
- Model Building

1.4. System Scope

System shares a few applications in real world basically it is helpful to eliminate hate on twitter. This hate is basically spread by fake accounts so with the help of this project we can identify those tweets and take actions accordingly.

It can be used in applications like Domain dependence, Explicit negation of sentiment etc.

1.5. Report and analytics

Sentiment analysis is a growing area of Natural Language Processing with research ranging from document level classification (Pang and Lee 2008) to learning the polarity of words and phrases (e.g., (Hatzivassiloglou and McKeown 1997; Esuli and Sebastiani 2006)). Given the character limitations on tweets, classifying the sentiment of Twitter messages is most similar to sentencelevel sentiment analysis (e.g., (Yu and Hatzivassiloglou 2003; Kim and Hovy 2004)); however, the informal and specialized language used in tweets, as well as the very nature of the microblogging domain make Twitter sentiment analysis a very different task. It's an open question how well the features and techniques used on more well-formed data will transfer to the microblogging domain. Just in the past year there have been a number of papers looking at 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). Other researchers have begun to explore the use of part-of-speech features but results remain mixed. Features common to microblogging (e.g., emoticons) are also common, but there has been little investigation into the usefulness of existing sentiment resources developed on non-microblogging data.

1.6. System Context :

Sentiment analysis can be defined as a process that automates mining of attitudes, opinions, views and emotions from text, speech, tweets and database sources through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like "positive" or "negative" or "neutral". It's also referred as subjectivity analysis, opinion mining, and appraisal extraction.

1.7. System Functions :

Our system provides primarily two functionalities:

1. Analysis and visualization:
 - a. Fetching data that are tweets
 - b. Using Libraries like pandas, seaborn, Matplotlib and Plotly for visualization.
 - c. Performing different types of analysis for predictions.
2. Machine Learning Model training:
 - a. Preparing data for training and testing.
 - b. Predict the sentiment of tweets.
 - c. Visualize the prediction results.

Chapter : 2 LITERATURE REVIEW

To track and monitor different datasets, most studies began with collecting the desired datasets from twitter, and applied filtering techniques to remove redundant data or spam tweets. Then parsed the data into a structured form. Finally analyzed the data. Below we review several types of analyses that most researchers have used.

2.1. Datasets

Analyzing structured data have been widely used. In such case, the traditional Relational Database Management System (RDBMS) can deal with the data. With the increasing amounts of unstructured data on various sources (e.g. Web, Social media, and Blog data) that are considered as Big Data, a single computer processor cannot process such huge amount of data. Hence, the RDBMS cannot deal with the unstructured data; a nontraditional database is needed to process the data, which is called NoSQL database. Most studies focused on tools, such as R (the programming language and the software environment for data analysis). R has limitations when processing twitter data, and is not efficient in dealing with large volume of data. To solve this problem a hybrid big data framework is usually employed, such as Apache Hadoop (an open source Java framework for processing and querying vast amounts of data on large clusters of commodity hardware) . Hadoop also deals with structured and semi-structured data, XML/JSON files, for example. The strength of using Hadoop comes in storing and processing large volume of data, while the strength of using R comes in analyzing the already-processed data. There are different types of twitter data such as user profile data and tweet messages. The former is considered static, while the latter is dynamic. Tweets could be textual, images, videos, URL, or spam tweets. Most studies do not, usually, take spam tweets and automatic tweets engines into account as they can, often, affect the accuracy and add noise and bias to analysis results. The mechanism of FireFox add-on and Clean Tweet filter was employed to remove users that have been on twitter for less than a day and they removed tweets that contain more than three hashtags.

2.2. Data Retrieval

Before retrieving the data, some questions should be addressed: What are the characteristics of the data? Is the data static, such as the profile user information “name, user Id, and bio”; or dynamic such as user’s tweets, and user’s network? Why is the data important? How is the data will be used? And how big the data is? It is important to note that it is easier to track a certain keyword attached to a hashtag rather than a keyword not attached to it. Twitter-API is a widely used application to retrieve, read and write twitter data. Other studies, have used GNU/GPL application like YourTwrapperKeeper tool, which is a web-based application that stores social media data in MySQL tables. However, YourTwrapperKeeper in storing and handling large size of data exhibits some limitations in using, as MySQL and spreadsheets databases can only store a limited size of data. Using a hybrid big data technology might address such limitations as we suggested above.

2.3. Ranking and Classifying Twitter Users

There are different types of user’s networks; a network of users within a specific event (hashtag), a network of users in a specific user’s account, and a network of users within a group in the network, that is, Twitter Lists. Lists are used to group sets of users into topical or other categories to better organize and filter incoming tweets . To rank twitter users, it is important to study the characteristics of twitter by studying the network-topology (number of followers/ followed) for

each user in the dataset. Many techniques have been employed in ranking analysis. Twitter users are ranked by identifying the number of followers by studying the PageRank, and by the retweet rate. In that study, 41.7 million user profiles, 1.47 billion social relations, and 106 million tweets were used. A new methodology is introduced to rank twitter users by using the Twitter Lists to classify users into the Elite users (Celebrities, Media news, Politicians, Bloggers, and Organizations) and the Ordinary users.

2.4. Homophily

Homophily is defined as the tendency that contacts among similar users occur at a higher rate than among dissimilar users, that is, similar users tend to follow each other. It requires studying the static characteristics of twitter data, such as the profile name and the geographic feature of each user in twitter network. We studied the homophily in twitter; studied the geographical feature in twitter to investigate the similarity between users based on their location. Additional work had been investigated in, homophily was studied using Twitter Lists to identify the similarity between the elite and ordinary users.

2.5. Reciprocity

The characteristic nature of twitter as being both directed and undirected social network has made most studies analyze reciprocity. Reciprocity is the property of following a user and being followed back (mutual relationship). For instance, celebrities tend to follow each other, so are politicians, bloggers, and ordinary users. We can conclude that homophily and reciprocity have the same logical behavior. The reciprocal relationship is measured by analyzing the number of followers, PageRank, and retweet rate. Additional methodology is investigated in, where the users follower-graph is studied to infer users reciprocities.

Chapter 3: REQUIREMENTS

3.1. Functional Requirements

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, data manipulation and processing and other specific functionality. In this system following are the functional requirements: -

Following are the functional requirements on the system:

1. The entire control model set must be translated to C output Code.
2. Inputs must be models designed using CLAW design components along with standard design components,
3. Multiple design models must be processed, and the result must be combined to obtain a single output file.

3.2. Non - Functional Requirements

Nonfunctional requirements are the requirements which are not directly concerned with the specific function delivered by the system. They specify the criteria that can be used to judge the operation of a system rather than specific behaviors. They may relate to emergent system properties such as reliability, response time and store occupancy.

Nonfunctional requirements arise through the user needs, because of budget constraints, organizational policies, the need for interoperability with other software and hardware systems or because of external factors such as: -

- Product Requirements
- Organizational Requirements
- User Requirements
- Basic Operational Requirements

3.3. Product Requirements :

Platform Independency: Standalone executables for embedded systems can be created so the algorithm developed using available products could be downloaded on the actual hardware and executed without any dependency to the development and modeling platform.

Correctness: It followed a well-defined set of procedures and rules to compute and also rigorous testing is performed to confirm the correctness of the data.

Ease of Use: Model Coder provides an interface which allows the user to interact in an easy manner.

Modularity: The complete product is broken up into many modules and well- defined interfaces are developed to explore the benefit of flexibility of the product.

Robustness: This software is being developed in such a way that the overall performance is optimized and the user can expect the results within a limited time with utmost relevance and correctness Nonfunctional requirements are also called the qualities of a system. These qualities can be divided into execution quality & evolution quality. Execution qualities are security & usability of the system which are observed during run time, whereas evolution

quality involves testability, maintainability, extensibility or scalability.

3.4. Organizational Requirements

Process Standards: The standards defined by DRDO are used to develop the application which is the standard used by the developers inside the defense organization.

Design Methods: Design is one of the important stages in the software engineering process. This stage is the first step in moving from problem to the solution domain. In other words, starting with what is needed design takes us to work how to satisfy the needs.

User Requirements:

- The coder must request the name of the model file to be processed
- In case of multiple files, the coder must ask the names of the files sequentially.
- The output file must be a C code translated from the model.
- Only a single output file must be created even if multiple input files are provided.

Chapter 4 : SENTIMENT ANALYSIS

Sentiment analysis is the measure of people's opinions on the level of agreement on a specific topic, a product, or a service, or even elections. Two approaches had been employed to study the sentiment analysis: natural language processing, and machine learning algorithms. To assess the customers' opinions in the past some paper-based surveys had been used, but it is difficult to monitor and collect all customers' opinions. With the increasing phenomena of social media it has become easier and more accessible to crawl all customers' feedbacks and analyze their sentiments as positive or negative.

4.1. Natural Language Processing Approach

Natural language processing (NLP) is the interaction between computers and human (natural) languages. To evaluate sentiment of users online, particularly on twitter, effective sentiment annotation should be used. Most studies use the three common sentiment labels: positive, neutral, and negative. new feature had been used to effectively annotate sentiments of users; "Mixed Sentiment label", it exists in tweets that have two different meanings. For example "I love iPhone, but I hate iPad". "iPhone" entity is annotated with positive sentiment label, and "iPad" entity is annotated with negative sentiment label, that means the tweet has a mixed sentiments.

4.2. Machine Learning Approach

Machine learning (ML) is a scientific discipline that explores the construction and the study of algorithms that can learn from data. We used the machine learning approach in analyzing the sentiment of twitter users. Machine learning is the study of algorithms that can learn from and make predictions on data. It is also called as related to prediction-making on some data. There are many machine learning algorithms. But this paper explains about one of them., as it is used in my project And That is -

- Naïve baye's

4.2.1. Naïve baye's

Naïve bayes is one of the most improved classification (classifier) methods. First in order to perform classification, we must select the features from the data set. All the tweets in the data sets will be processed by the classifiers. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. A naïve bayes algorithm is very easy to build up and mainly used for a large set of data. It provides a way of calculating $p(c|x)$ from $p(c)$, $p(x)$ and $p(x|c)$. Here $p(c|x)$ is called the posterior probability and it is given by the formula,

$$p(c|x) = \frac{p(x|c) p(c)}{p(x)} \text{ where,}$$

$P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes).

$P(c)$ is the prior probability of class.

$P(x|c)$ is the likelihood which is the probability of predictor given class.

$P(x)$ is the prior probability of predictor.

4.2.1.1. Applications of Naïve Bayes's

The applications of naïve bayes-

Real time Prediction: Naive Bayes algorithm is also a fast learning algorithm. Thus, it is used for making predictions in real time.

Multi class Prediction: This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes also.

Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers mostly used in text classification as it has a better result in multi class problems and have higher success rate as compared to other algorithms. This is also used to identify spam e-mail. Main application is sentimental analysis it is used to predict whether a user would like a given resource or not.

4.3. Main challenges that are faced by and sentiment analysis

Detection of spam and fake reviews: The web contains both authentic and spam contents. For effective Sentiment classification, this spam content should be eliminated before processing. This can be done by identifying duplicates, by detecting outliers and by considering reputation of reviewer.

- **Limitation of classification filtering:** There is a limitation in classification filtering while determining most popular thought or concept. For better sentiment classification result this limitation should be reduced. The risk of filter bubble gives irrelevant opinion sets and it results false summarization of sentiment.

- **Asymmetry in availability of opinion mining software:** The opinion mining software is very expensive and currently affordable only to big organizations and government. It is beyond the common citizen's expectation. This should be available to all people, so that everyone gets benefit from it.

- **Incorporation of opinion with implicit and behavior data:** For successful analysis of sentiment, the opinion words should integrate with implicit data. The implicit data determine the actual behavior of sentiment words.

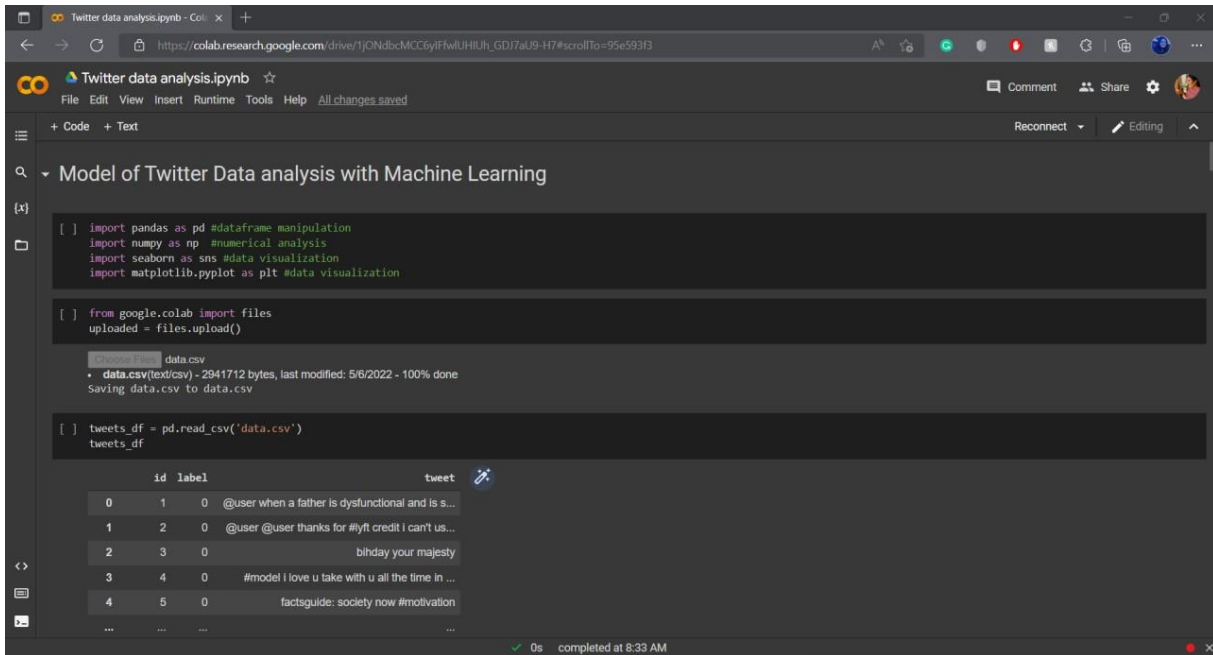
- **Domain-independence:** The biggest challenge faced by opinion mining and sentiment analysis is the domain dependent nature of sentiment words. One features set may give very good performance in one domain, at the same time it perform very poor in some other domain.

- **Natural language processing overheads:** The natural language overhead like ambiguity, co-reference, Implicitness, inference etc.

4.4. CONCLUSION AND FUTURE SCOPE

Applying sentimental analysis to extract the sentiment became an important work for many organizations and even individuals. Sentiment analysis is an emerging field in decision making process and is developing fast. Our project goal is to analyze the sentiments on a topic which are extracted from the Twitter and determine its nature (positive/negative/neutral) of the defined topics. The development of techniques for the document-level sentiment analysis is one of the significant components of this area. Recently, people have started expressing their opinions on the Web that increased the need of analyzing the opinionated online content for various real-world applications. A lot of research is present in literature for detecting sentiment from the text. Still, there is a huge scope of improvement of these existing sentiment analysis models. Existing sentiment analysis models can be improved further with more semantic and commonsense knowledge.

Working Models screenshots :



The screenshot shows a Google Colab notebook titled "Twitter data analysis.ipynb". The code cell contains the following Python code:

```
[ ] import pandas as pd #dataframe manipulation
import numpy as np #numerical analysis
import seaborn as sns #data visualization
import matplotlib.pyplot as plt #data visualization

[ ] from google.colab import files
uploaded = files.upload()

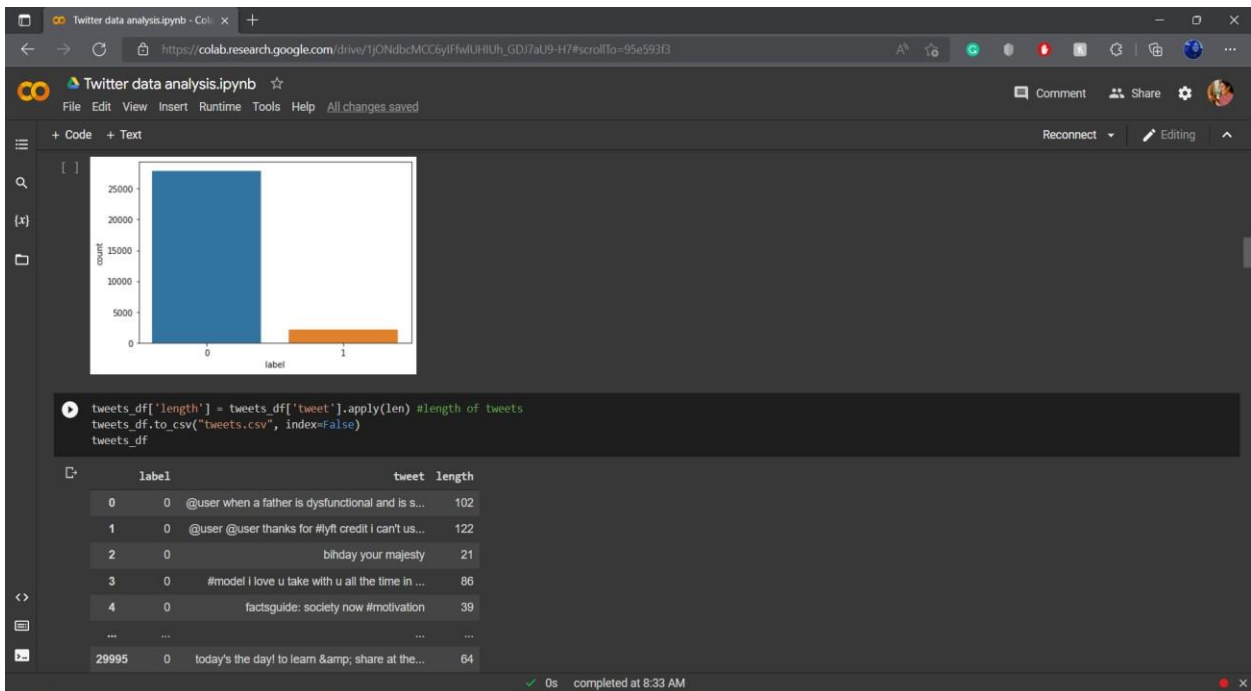
data.csv
• data.csv(text/csv) - 2941712 bytes, last modified: 5/6/2022 - 100% done
Saving data.csv to data.csv

[ ] tweets_df = pd.read_csv('data.csv')
tweets_df
```

The output shows a preview of the DataFrame:

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation
...

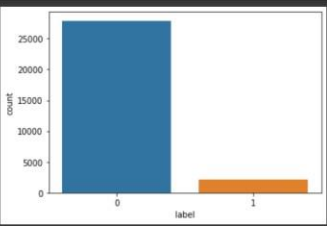
The status bar at the bottom indicates "0s completed at 8:33 AM".



The screenshot shows the same Google Colab notebook. The code cell contains the following Python code:

```
[ ] tweets_df['length'] = tweets_df['tweet'].apply(len) #length of tweets
tweets_df.to_csv("tweets.csv", index=False)
tweets_df
```

The output shows a bar chart of tweet lengths:



label	count
0	25000
1	2000

The output also shows a preview of the DataFrame with a 'length' column:

	label	tweet	length
0	0	@user when a father is dysfunctional and is s...	102
1	0	@user @user thanks for #lyft credit i can't us...	122
2	0	bihday your majesty	21
3	0	#model i love u take with u all the time in ...	86
4	0	factsguide: society now #motivation	39
...
29995	0	today's the day! to learn & share at the...	64

The status bar at the bottom indicates "0s completed at 8:33 AM".

Twitter data analysis.ipynb - Colab

https://colab.research.google.com/drive/1jONdbcmCC6yIFwLUHUh_GD77aU9-H7#scrollTo=95e593f3

Twitter data analysis.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

Reconnect Editing

Assess trained model performance:

```
[ ] from sklearn.metrics import classification_report, confusion_matrix
```

```
[ ] # predicting the test set results
y_predict_test = NB_classifier.predict(x_test)
cm = confusion_matrix(y_test, y_predict_test)
sns.heatmap(cm, annot = True)
```

```
[ ] print(classification_report(y_test, y_predict_test))
```

	precision	recall	f1-score	support
0	0.96	0.97	0.97	5584

0s completed at 8:33 AM

Twitter data analysis.ipynb - Colab

https://colab.research.google.com/drive/1jONdbcmCC6yIFwLUHUh_GD77aU9-H7#scrollTo=95e593f3

Twitter data analysis.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

Reconnect Editing

```
[ ] # predicting the test set results
y_predict_test = NB_classifier.predict(x_test)
cm = confusion_matrix(y_test, y_predict_test)
sns.heatmap(cm, annot = True)
```

```
[ ] print(classification_report(y_test, y_predict_test))
```

	precision	recall	f1-score	support
0	0.96	0.97	0.97	5584
1	0.58	0.50	0.54	416
accuracy			0.94	6000
macro avg	0.77	0.74	0.75	6000
weighted avg	0.94	0.94	0.94	6000

0s completed at 8:33 AM

CONCLUSION

Nowadays, sentiment analysis or opinion mining is a hot topic in machine learning. We are still far to detect the sentiments of a corpus of texts very accurately because of the complexity in the English language and even more if we consider other languages such as Chinese. In this project we tried to show the basic way of classifying tweets into positive or negative category using Naive Bayes as baseline and how language models are related to the Naive Bayes and can produce better results. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the naïve Bayes classifier, or trying another classifier all together.

Our heartfelt appreciation goes to Professor Mir Shahnawaz Ahmad with regards to his feedback across the course of project from the initial proposal up to the conclusion and for the valuable lessons learned along the way .

References

1. Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow Concepts, Tools, and Techniques to Build Intelligent Systems by Aurélien Géron,
2. Introduction to Machine Learning with Python by Andreas C. Müller & Sarah Guido
3. Toward Science/medium.com
4. Twitter official Documentation, Pandas Documentation
5. Other web resource: Github, Stackoverflow.