A Project Report on

Twitter Sentiments Analysis using Machine Learning

Submitted in partial fulfillment of the requirements for the award
of the degree of

Bachelor of Engineering

in

Computer

by

Parth B. Shah.(17102035)
Khush M. Shah.(17102033)
Vishnu Ezhuthassan(17102071)

Under the Guidance of

Prof. Jaya D. Gupta.

Department of Computer
NBA Accredited
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI

Academic Year 2021-2022

Approval Sheet

This Project Report entitled ***"Twitter Sentiments Analysis using Machine Learning"*** Submitted by ***"Parth B. Shah."(17102035),"Khush M. Shah."(17102033),"Vishnu Ezhuthassan"(17102071)*** is approved for the partial fulfillment of the requirenment for the award of the degree of ***Bachelor of Engineering*** in ***Computer*** from ***University of Mumbai***.

(Prof. Jaya D. Gupta.)
Guide

Prof. Sachin H. Malave.
Head of Department, Computer.

Place : A.P.Shah Institute of Technology, Thane
Date : 08/11/2021

CERTIFICATE

This is to certify that the project entitled ***"Twitter Sentiments Analysis using Machine Lerning"*** submitted by ***"Parth B. Shah." (17102035),"Khush M. Shah." (17102033),"Vishnu Ezhuthassan" (17102071)*** for the partial fulfillment of the requirement for award of a degree ***Bachelor of Engineering*** in ***Computer***, to the University of Mumbai, is a bonafide work carried out during academic year 2021-2022.

(Prof. Jaya D. Gupta.)
Guide

Prof. Sachin H. Malave.
Head of Department, Computer.

Dr. Uttam D. Kolekar.
Principal

External Examiner(s)

1.

2.

Place : A.P.Shah Institute of Technology, Thane
Date : 08/11/2021

Acknowledgement

We have great pleasure in presenting the report on **Twitter Sentiments Analysis using Machine Learning** We take this opportunity to express our sincere thanks towards our guide **Prof. Jaya D. Gupta.** Department of Computer, APSIT thane for providing the technical guidelines and suggestions regarding line of work. We would like to express our gratitude towards his constant encouragement, support and guidance through the development of project.

We thank **Prof. Sachin H. Malave.** Head of Department, Compute, APSIT for his encouragement during progress meeting and providing guidelines to write this report.

We thank **Prof. Amol Kalugade** BE project co-ordinator, Department of Computer, APSIT for being encouraging throughout the course and for guidance.

We also thank the entire staff of APSIT for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Student Name1 : Parth B. Shah.
Student ID1 : 17102035

Student Name2 : Khush M. Shah.
Student ID2 : 17102033

Student Name3 : Vishnu Exhuthassan
Student ID3 : 17102071

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

_____

(Signature)

_____

(Parth B. Shah. (17102035))
(Khush M. Shah. (17102033))
(Vishnu Exhuthassan. (17102071))

Date : 08/11/2021

Abstract

Analysis of sentiments is the method of deciding whether the sentiments in the text are positive, negative or neutral. It is also known as material polarity or mining of opinions. The growth and advancement in social media platforms engaged a huge number of users. Social media platforms like twitter where users can post their tweets in 280 characters. Because of the limited number of characters in tweets, it becomes easy for sentiment analysis. On Twitter 550 millions of tweets are posted daily. Twitter also represents all age group people and also a fair representation of gender. Therefore, the sentiment analysis of twitter data becomes somewhat general sentiments of society.

Contents

List of Tables

List of Abbreviations

ML : Machine Learning
TM : Text Mining
SA : Sentiment Analysis
OM : Opinion Mining
GUI : Graphical User Interface
NLP : Natural Language Processing

Chapter 1

Introduction

Twitter sentiments analysis makes use of Natural Language Processing (NLP) to evaluate a speaker's, writer's, or other person's mood and emotions through the piece of text. Sentiment analysis refers to identifying as well as classifying the sentiments that are expressed in the text source. Tweets are often useful in generating a vast amount of sentiment data upon analysis. These data are useful in understanding the opinion of the people about a variety of topics.

Through Sentiments Analysis, we can determine if a tweet of a user is positive, negative or neutral. Social networking platforms such as Twitter, Facebook, Instagram, YouTube, etc. have been so popular for days now. They allow people to communicate, create networks, and share thoughts easily and promptly. Twitter has become an excellent medium for opinion creation and presentation.

Chapter 2

Literature Review

Sentiment analysis in the domain of micro-blogging is a relatively new research topic so there is still a lot of room for further research in this area. Decent amount of related prior work has been done on sentiment analysis of user reviews,documents, web blogs/articles and general phrase level sentiment analysis. These differ from twitter mainly because of the limit of 280 characters per tweet which forces the user to express opinion compressed in very short text. The best results reached in sentiment classification use supervised learning techniques such as Naive Bayes and Support Vector Machines, but the manual labelling required for the supervised approach is very expensive. Some work has been done on unsupervised and semi-supervised approaches, and there is a lot of room of improvement. Various researchers testing new features and classification techniques often just compare their results to baseline performance.

2.1 Objectives

To utilize Twitter data due to its widespread global acceptance. The rapid expansion and acceptance of social media has opened doors into opinions and perceptions that were never as accessible as they are with today's prevalence of mobile technology. Harvested Twitter data, analyzed for opinions and sentiment can provide powerful insight into a population. This insight can assist companies by letting them better understand their target population. The knowledge gained can also enable governments to better understand a population so they can make more informed decisions for that population.

2.2 Problem definition

A target defines the chosen person as the source of the emotion. A target determines the individual chosen as the origin of the sentiment. Polarity can be described in both positive and negative aspects, or can be represented in three positive, negative and neutral aspects. Feature defines the particular aspect or attribute of the objective to which the emotion is expressed. Take the following example1: Samsung Phone cost lower than ten thousand, does not gives a good performance. Aspect is also an essential part of sentiment analysis. There are multiple aspects of assessments, since a tweet's sentiment can be conveyed in entire text (Document-level analysis), semantically and syntactically related texts (Sentence-level analysis) or aspects are extracted from the text (Aspect-level analysis).

2.3 Scope

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 sentiment analysis, the polarity of opinions can be found, such as positive, negative, or neutral by analyzing 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. Other techniques discussed in sentiment analysis include Naive Bayes.

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 sentiment 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) to identify positive, negative, and neutral tweets to use for training threeway sentiment classifiers. While there has been a fair amount of research on how sentiments are expressed in genres such as online reviews and news articles, how sentiments are expressed given the informal language and message-length constraints of microblogging has been much less studied.

2.4 Technology stack

1. Python.
2. Web scrape using Twitter's API.
3. Natural Language Processing.
4. Naïve Bayes Classifier.

2.5 Benefits for environment and society

Twitter Sentiments Analysis can be used for real-time applications which can be a very helpful business. It can be used for people's sentiments on current political topics or trends. It can also be used for the review of movies using the trends on twitter.As the audience on the media platforms grows continuously data from these sites can be used to analyses the sentiments of the people.

- Manufacturers or developers of the products of the can review their product by analyzing the sentiments of the people. That is how people reacting to their products.

- Marketing personal can see how people are reacting to their advertising campaign. They can analyze the sentiments related to this.
- Political parties can see how their political campaign is running and how people reacting to it. They can analysis which issue to be raised to not.
- Filmmakers can see how people are reacting to their newly released movie, by analyzing the sentiments of the people.

Chapter 3

Project Design

- Retrieval of tweets : As twitter is the most enlarged part of social networking site, it consists of various blogs which are related to various topics in the world. Instead of taking whole blogs, a search for a particular topic will be done and then download all tweets then extract them in the form of csv files.
- Pre-processing of removed data: After retrieval of tweets Sentiment analysis tool is applied on untested tweets but in most of cases results to very poor performance. Therefore, preprocessing techniques are necessary for obtaining better results as given. We extract tweets i.e. short messages from twitter which are used as untested data. This data needs to be preprocessed. So, preprocessing involves following steps
- Filtering: Filtering is nothing but extraction of raw data. In this step, URL links (E.g. http://twitter.com), special words in twitter, user names in twitter e.g. @Ron - @ symbol indicating a user name, emoticons are extracted.
- Tokenization: Tokenization is nothing but partitioning of sentences. In this step, we will tokenize or segment text with the help of partitioning text by spaces and punctuation marks to form container of words.
- Removal of Stopwords: Articles like “a”, “an”, “the” and other stopwords such as “to”, “of”, “is”, “are”, “this”, “for” removed in this step.
- Construction of n-grams: n-grams can make out of consecutive words. Negation words such as “no”, “not” is attached to a word which follows and precedes it. For Instance: “I do not like remix music” has two bigrams: I do+not, do+not like, not+like remix sentence, So the correctness of the classification improves by such procedure, because negation plays an important role in sentiment analysis.
- Output sentiment: Based on the dictionary assignment of score, the proposed system interprets whether the tweet is positive, negative or neutral.

3.1 Design(Flow Of Modules)

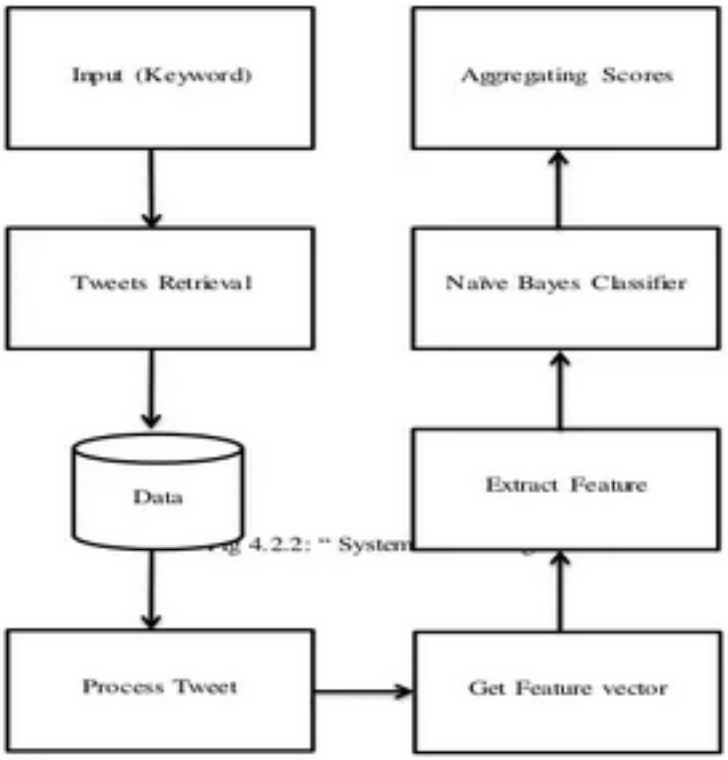
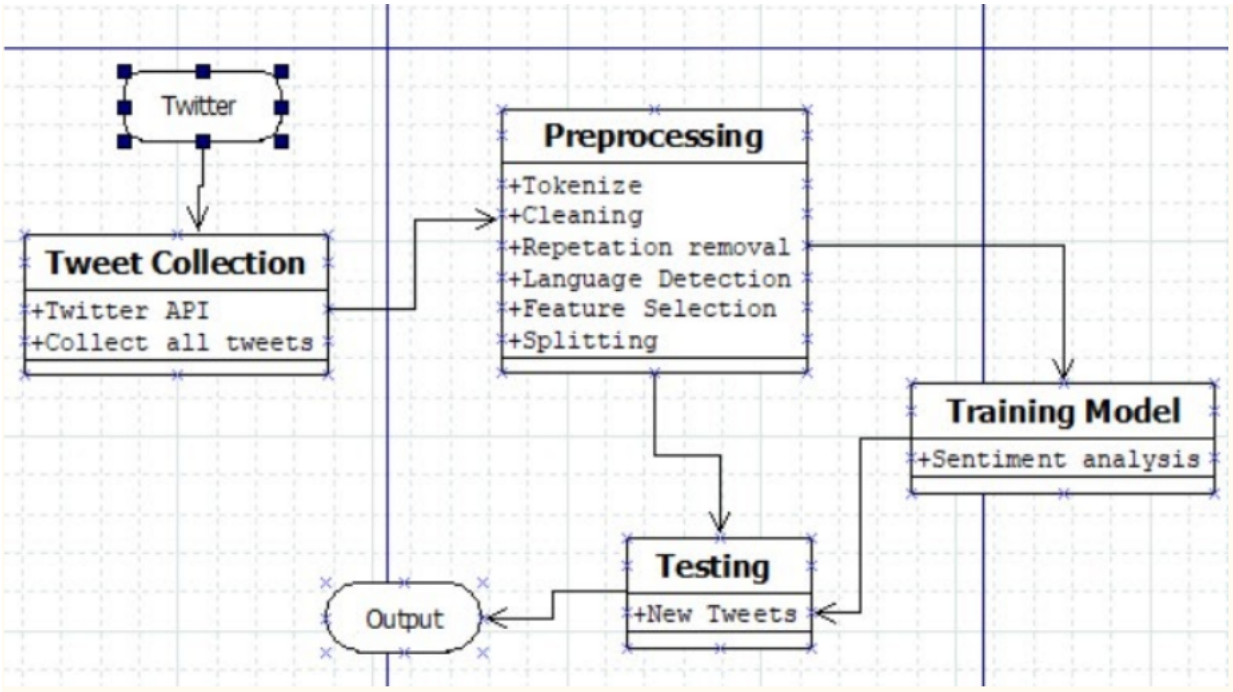


Fig 4.2.2: " System

3.2 Class Diagram



Chapter 4

Conclusions and Future Scope

A machine learning technique of data analysis of twitter are discussed i.e. Naïve Bayes. The analysis of twitter data is being done in various aspects to mine the sentiments. This study defines the concept of opinion in sentiment analysis of Twitter. Sentiment analysis deals with opinion classified into positive, negative and neutral. The study shows that the machine learning method such as Naïve Bayes has the highest accuracy and can be consider as the baseline learning methods. More work in future is needed to improve the performance measures.

1. As of now we are trying to execute on the dataset which we have, in that we are trying to remove all the non - english words, converting uppercase characters to lowercase, removing all the hash and hashtags, etc. And trying to increase the accuracy of the model.
2. After that, we will try to web scrape the data using Twitter API and try to implement our application.
3. We'll try to make our application work, in niche field of different domains.
4. We will try to create GUI for our application, which will be user friendly and has smooth User Interface.

Bibliography

[1] Evolutionary Machine Learning Techniques: Algorithms and Applications (Algorithms for Intelligent Systems) by Seyedali Mirjalili, Hossam Faris, Ibrahim Alijarah.

[2] Sentiment Analysis Using Support Vector Machine 1Aamera Z. H. Khan, 2Dr. Mohammad Atique, 3Dr. V. M. Thakare in International Journal of Advanced Research in Computer Science and Software Engineering(http://ijarcsse.com/BeforeAugust2017/docs/papers/SpecialIssue/ITSD2015/25).

[3] Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers International Journal of Scientific and Research Publications(http://www.ijsrp.org/research-paper1018.php?rp=P827862).

[4] SENTIMENT ANALYSIS OF TWEETS USING SUPPORT VECTOR MACHINE Suman Rani1 , Jaswinder Singh Suman Rani et al, International Journal of Computer Science and Mobile Applications, Vol.5 Issue. 10, October- 2017, pg. 83-91.

[5] Sentiment Analysis using Maximum Entropy Algorithm in Big Data Durgesh Patel , Sakshi Saxena , Toran Verma, International Journal of Innovative Research in Science, Engineering and Technology(http://www.ijirset.com/upload/2016/may/24649Sentiment.pdf).

[6] International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 08 — Aug-2016 www.irjet.net, Sentiment Analysis Using SVM and Maximum Entropy Snehal L. Rathod, Sachin N.Deshmukh.

[7] International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8 June, 2019, Sentiment Analysis Using Naïve Bayes Classifier Sentiment Analysis Using Naïve Bayes Classifier.

Appendices

Naïve Bayes Classifier :

Naïve Bayes is the supervised machine Learning algorithm which uses Bayes theorem for classification problems. It is mostly used in text classification which includes a dataset for high-dimensional training. It is one of the simplest and most powerful classification algorithms that helps to create fast machine learning models that can make predictions quickly. It is mostly used for text analysis, sentiments analysis and classifying articles. It make use of following Bayes theorem :-

Prob(A/B) = (Prob(B/A)*Prob(A))/Prob(B)

Where:
x Prob(A—B) is know as (posterior probability) of hypothesis A which occurs when some condition is already provided.
x Prob(B—A) it is also called likelihood probability it is the probability of evidence E when we presumes that given hypothesis is true.
x Prob(A) it is the prior known probability of A and does not include any condition.
x Prob(B) it is the prior known probability of A and does not include any condition.

Formula for Calculation of Accuracy and Precision :
So, based on our study we have divided the tweets from the users on the basis of the sentiments that is positive, negative or neutral. Here we have created a contingency matrix for reference to our below tabular data.

		Actual Value		
Predicted Value		Positive Tweets	Negative Tweets	Neutral Tweets
Predicted Value	Positive Tweets	True Negative	False Positive	False Positive
Predicted Value	Negative Tweets	False Negative	True Negative	False Negative
Predicted Value	Neutral Tweets	False Neutral	False Neutral	True Neutral

Table 4.1: CONTINGENCY MATRIX

Here, True Positive are the positive tweets which are actually classified positive. Where else, False Positive are the tweets which are positive but, classified as negative or neutral. Similarly True Negative are the negative tweets which are classified as negative where else False Negative are the Negative Tweets which are classified as positive or neutral tweets. True Negative are the tweets which are actually found neutral and classified as neutral where else False Negative are the tweets which are neutral tweets but, classified as positive or negative.

1. Precision Calculation :

Precision = True Positive/(True Positive + False Positive)

2. Accuracy Calculation :

Accuracy = Number of Current Predicted Data/Total Number of Data

Accuracy = [(True Positive + True Negative + True Neutral)/Total]*100