A PROJECT REPORT

on

Sentiment Analysis on Twitter Data using Machine Learning

Submitted by

NEELAMPALLE NIKHIL KUMAR (18BEC031)

YALAPALLI SANTHI SWARUP (18BEC051)

MEGAVATH VINOD (18BEC027)

in partial fulfillment for MINI PROJECT – II

of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS AND COMMUNICATION ENGINEERING

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY
DHARWAD

ITTIGATTI ROAD, NEAR SATTUR COLONY

DHARWAD – 580009, KARNATAKA

AUGUST 2021/DECEMBER 2021

Acknowledgement

We would like to express our sincere gratitude to Dr. Ramesh Athe, Asst.Professor, Department of Data Science and Intelligent Systems, IIIT Dharwad for his guidance and constant support throughout the course of this minor project. We would also like to thank all the faculty and administration of the institute who ensured the needs fulfilled for the completion of this project.

Date: May 2021
Place: Dharwad

NEELAMPALLE NIKHIL KUMAR (18BEC031)
YALAPALLI SANTHI SWARUP (18BEC051)
MEGAVATH VINOD (18BEC027)

TABLE OF CONTENTS

S.NO	CONTENTS	PAGE NO.
1	ABSTRACT	4
2	INTRODUCTION	4-5
3	LITERATURE SURVEY	6-7
4	PROPOSED METHODOLOGY	8-11
5	RESULTS and OBSERVATION	11-12
6	CONCLUSION	12
7	REFERENCES	13

1. ABSTRACT

Over few decades Information Technology has evolved a lot. The advent of social media and its progress over the years, has enabled internet users to express their opinions, share their views, thoughts, and reviews etc. The use of social media apps by the people has been increasing rapidly. Huge amounts of data are being generated in various social media platforms. Over the years impact of the social media platforms like Facebook, Twitter and Instagram has been increased a lot. Twitter is an American micro blogging and social networking service. Users interact with messages called 'tweets' in it. Twitter gives a platform on which discussions on various topics. Twitter is much better for sentiment analysis than Facebook, because of the fact that Twitter data i.e., tweets can be easily extracted by emoticons (emotion icons). Tweets can be extracted from a twitter API. This platform offers many organizations a fast and effective way to analyze the opinion of customers towards the products, information etc. via their tweets, retweets, comments and likes.

Sentiment analysis is an approach to be used to measure customer's perception which helps organizations to better themselves. For developing these type of programs, Python which is the most popular programming languages is used because of its rich libraries such as Tweepy, pandas, TextBlob, wordcloud, numpy and more. The sentiments of the customers via twitter tweets will be visualized. We try implement it with help of Natural Language Processing techniques and we use Machine Learning to train and test, how better the sentiments were classified.

2. INTRODUCTION

Twitter Platform is one of the most used and popular social media platforms now-a-days. It is gaining more popularity towards its tweets, information etc.. People express a lot of sentiments through this platform. Social media is receiving more attention. This platform offers many organizations an effective way to interact with the opinion of customers. We can analyze these tweets of users or customers with sentiment analysis process and provide useful insights which provides information for organizations about their performance, impact etc. Sentiment analysis is the process of categorizing opinions expressed in a piece of text or tweet, especially whether r the tweet is positive, negative or neutral. Sentiment Analysis can be effectively used to analyze the opinion or sentiment. We use Twitter API to fetch twitter data. Python provides a library called Tweepy, through which we can access twitter data. On that, fetched data we try to perform our sentiment analysis process and try to draw out useful insights from our analysis also we check how accurately the sentiments we

Classified using machine learning algorithms.

A. Problem Statement

Our Projects is about, calculating the sentiments of the tweets that were fetched into either positive, negative or neutral and to draw out insights about the sentiment towards the target subject and try to check how accurately the sentiments were calculated using machine learning – classification algorithms.

B. Proposed work

In our proposed method we use Jupyter notebook to work on. We will fetch twitter data using Twitter API keys and Tweepy library. Further we perform some pre-processing tasks to simplify the tweets to perform sentiment analysis and to draw out insights from it. And to we check how accurate were the sentiments calculated by training and testing our dataset to various machine learning models.

C. Scope of work

Twitter sentiment analysis allows us to check what all are being said about your product or service or any other topic on social media. As there will many reviews it is not is to read all of them. So, sentiment analysis takes over the manual task of reading stuff and analyzing them. Using Sentiment analysis we can make out opinions of many people quickly.

D. Organization

This paper flows by explaining the research that already happened on the present investigation. And tells the need of present research. Followed by architecture of the entire research and implementation. Then with a brief description about workflow carried out throughout this project. Then explains Sentiment analysis and finishes with observations and conclusions.

3. LITERATURE SURVEY

A. Review of literature

Sentiment analysis is an approach for Natural Language Processing which identifies the emotion or sentiment underlying in body of text. This is popular method for organizations to determine and classify the sentiment or opinion about a product, service or idea. Social media monitoring, Customer support, Customer feedback, Brand monitoring etc. are some applications of Sentiment analysis

3.1. The effect of preprocessing techniques on Twitter sentiment analysis

Akrivi Krouska at el. Proposed the necessary information to get preprocess the tweets related to a topic in order to find the sentiments. Discussed role of text preprocessing and extended sentiment polarity classification methods in depth. Finally, they concluded that feature selection and representation can affect the classification performance positively.

3.2. Twitter Sentiment Analysis System

Shaunak Joshi and Deepali Deshpande have tried to classify human sentiment into two categories - positive and negative. Which helps to better understand human thinking and gives us an insight which further can be used. They collected dataset with labelled Sentiment from kaggle.com via internet. Data provided included with emoticons, usernames and hashtags which are required to be processed. They also extracted features like unigrams and bigrams which is form of representation of the tweet. They finally reported experimental results after all the process

3.3. Machine Learning-Based Sentiment Analysis for Twitter Accounts

Ali Hasan at el. considered tweets that are posted by the users with hashtags to express their opinions on political trends at those times. They also considered Urdu tweets which were translated further and preprocessed the dataset. They calculated subjectivity and polarity using three different libraries SentiWordNet, W-WSD and TextBlob. They used Naïve Bayes and SVM classifier and built a classification model that model was tested on training dataset to obtain accuracy result of every classifier used

3.4. Twitter Sentiment Analysis : A case study of Apparel Brands

Rasool at el. used streaming API Tweepy to extract twitter data from twitter they retrieved tweets using apparel brand names as a keyword including ‘Nike’ and ‘Adidas’. They performed necessary preprocessing steps for further analysis. They used Naïve Bayes classifier for sentiment classification. They compared user’s opinions with respect to the two apparel brands to help out the marketing and decision strategy for the

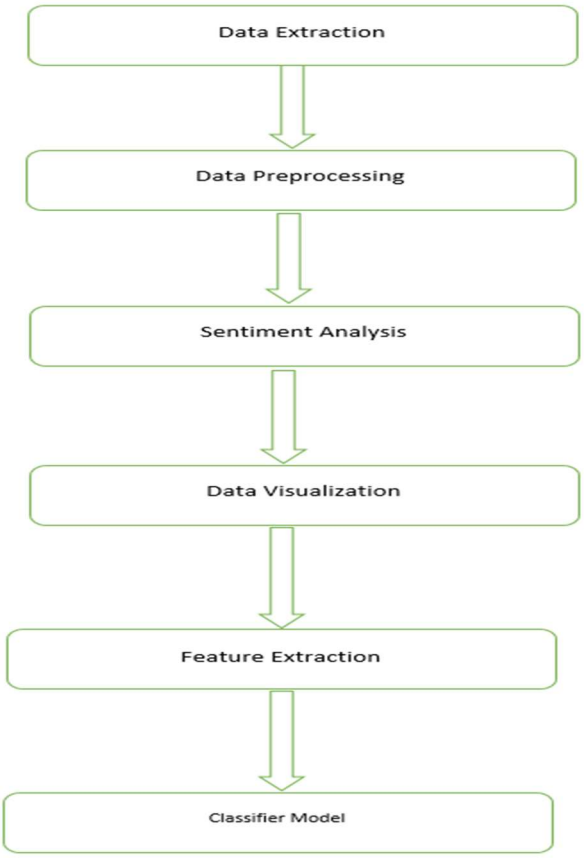
apparel industry

3.5. Twitter Sentiment Analysis During Covid-19 Outbreak in Nepal

B Prakash proposed to perform sentiment analysis over COVID-19 in Nepal country. He was interested in finding how citizens of Nepal were reacting towards the pandemic and they tried to perform sentiment analysis using related hashtags. He used TextBlob to calculate sentiments of the tweets. Data extracted from twitter was during the pandemic period.

4. PROPOSED METHODOLOGY

We are trying to implement sentiment analysis over twitter data related any product like 'Windows 11' or 'iphone13' i.e. tech related topics.



A. DATA EXTRACTION

To get access for twitter data we need create a twitter developer account through which we can get access to twitter API. Tweepy is an open source python package to access the twitter API with Python. We use API key, API secret, access token, access token secret through which we get

from API details of our twitter developer account.
These keys are very much confidential and it must not be shared with others.
By accessing twitter API with python we can extract twitter data using python programming. We can get information like tweet, username, location etc. we will import the collected twitter data into csv files and export it into pandas DataFrame, which is a very convenient way to perform further analysis.

date-time	tweetdata	username	location
2021-11-23 15:12:01	Windows 11 optional update released with new Fluent emoji https://t.co/sXo9bls7N5 #Windows11 https://t.co/9Nmdt8i3iY	WindowsLatest	USA
2021-11-23 15:06:20	When you realise that #Windows11 won't allow you to upgrade because your pc isn't modern enough for their standards… https://t.co/Z9Y2CKl5fV	FrequencyFormat	King's Lynn, England
2021-11-23 15:00:31	@Chrisket18 Are you running #Windows10 or #Windows11 there 👀	TheSnakeDocTV	United Kingdom
2021-11-23 15:00:08	#windows11 tip: Auto HDR and Dynamic Refresh Rate. During the Windows 11 unveiling, Microsoft announced support for… https://t.co/A0oUnVboUo	HowTimeNet	NaN
2021-11-23 14:59:43	Microsoft's new emoji are now available in Windows 11 https://t.co/3jLHQTMuDV via @Verge \r\n@DrJDrooghaag @BillMew… https://t.co/uRBVjGXfca	pettet50	EU

B. DATA PREPROCESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data Quality Assessment, Data Cleaning, Data Reduction are some prominent steps used. In tweets users can write anything and post anything. Sentences or messages tweeted contain many in statements like mentions, hashtags, punctuations, emoji's and some unwanted characters. By performing preprocessing steps over the tweets we make our data more accurate and understandable. We try to remove unwanted characters emoji's, URL's, Punctuations etc. So that we make our data clean for further analysis. We also perform some steps like tokenization, stop words, stemming to preprocess our data, we perform these steps after cleaning the data.

Tokenization – Tokenization is nothing but breaking raw text or statements into small chunks. Tokenization breaks out the raw text into words called tokens. These tokens helps to understand the context or developing model for NLP.

Stop Words – stop words are mostly used in text mining and NLP to eliminate which are mostly used that carry very little useful information. 'a', 'the', 'is' etc. are some of the examples of stop words. Removing stop words, data size decreases and time to train the model will also decrease as there will be only few meaningful tokens when compared to the tokenized data.

Stemming – is a method which is used make out base format of the words after removing suffixes/prefixes from them. Simply it is just like cutting down all branches of a stem. For example the stem of the words fast, faster, fastest is fast. This helps in reducing computation time and dimensionality of data.

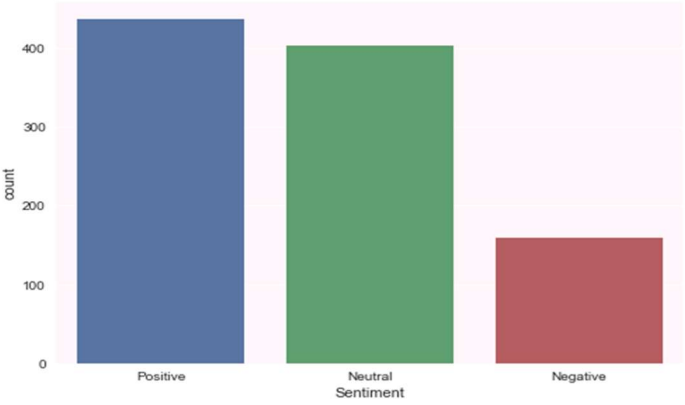
C. SENTIMENT ANALYSIS

After performing all the preprocessing steps we make our data readily available for NLP. We use an open source python package - TextBlob to perform sentiment analysis. It is a library for processing complex textual data and it supports NLP tasks like parts-of-speech tagging, noun phrase extraction, sentiment analysis, translation etc.
We use a polarity based approach to perform sentiment analysis. We use TextBlob library to calculate Polarity and Subjectivity of each and every tweet.
Polarity – it defines the orientation of how a text or statement is being expressed i.e. it determines whether the text is expressed in either positive negative or neutral way. This polarity ranges between -1 to +1. Where polarity values greater than 0 can be considered as positive, less than 0 can be considered as negative and equal to 0 is considered as neutral.

Subjectivity – it quantifies the text or statement's personal information and factual information. This subjectivity is a float value ranging between 0 and 1. It signifies that higher the subjectivity means that the text contains more personal opinion rather than factual information.
Based on polarity we classify all the tweets in the dataset either as positive, negative or neutral.

D. DATA VISUALIZATION

Data visualization is the representation of data or information in an easy understandable way. Graphs, charts, maps etc are some of ways of data visualization, which provides an easy way to understand and analyze the trends, patterns and outliers in data.
For our analysis we create visualizations like Pie Charts and Bar Graphs to view how many tweets are classified into each positive, negative and neutral category. We create word clouds to view most used words in the data, also we can check word clouds for separate positive, negative and neutral tweets and check what most common used words in the tweets in all three classifications. Scatter plots to analyze polarity vs subjectivity scores for all the tweets gathered.



E. FEATURE EXTRACTION

We now try to build a model from the data with sentiments. We need some features to be extracted from the data to build the model. We mainly extract two types of features from our processed data – Unigrams and Bigrams (N– grams).

N-grams are simple continuous sequence of words/tokens in a document. They can be defined as consecutive sequence of items in a document. It is very much useful while we deal with NLP tasks on text data. Unigrams – single token, bigrams – two token at a time, trigrams – 3 tokens etc. such features can be extracted. When consider n-grams, it makes up into a phrase rather than a word which becomes more meaningful in some cases. We use Scikit-learn library in further analysis. After extracting n-grams, we use CountVectorizer which a tool provided by scikit-learn module in python. It is used to convert given text or statement into a vector on the basis of frequency of each and every word that occurs in the entire text. We use these features in model building.

F. MODEL BUILDING

After feature extraction we will split our data set into test data and train data. We use these data to train and test with various machine learning classifier algorithms.

I. Naïve Bayes Algorithm

Naïve Bayes classifiers are a simple classifiers which are probabilistic and based on applying Bayes theorem with naive independence assumption among the features. This model can be used for text classification. In following model a class c^ is assigned to tweet or text t, where

$$\hat{c} = P(c|t) argmax_{\,c}$$

$$P(c|t) \propto P(c) \prod_{i=1}^{n} P(f_i|c)$$

Fi represents i-th feature among all n features. P(c/t) and P (fi/c) will be obtained through maximum likelihood estimates. We used MultinomialNB from sklearn.naivebayes

II. Logistic Regression

Logistic regression makes better prediction using maximum likelihood technique. Sigmoid is a mathematical function which can take any real value between – infinity to + infinity and map it to the real value between 0 to 1. If it is less than 0.5 it can classified as negative, greater than 0.5 as positive and at 0.5 as neutral class.

III. Decision Tree Classifier

Decision trees are one of the most used supervised machine learning algorithm. It is a prediction model made in the form of tree that is made using recursive splitting of internal nodes which represents testing on particular features where branches of tree denote outcome of each test and each node denotes final outputs. We will use this algorithm to perform classification.

IV. Support Vector Machine

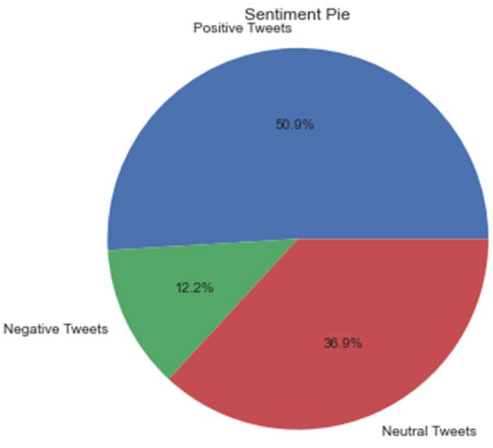
It is used for distinguishing or classification between several groups/classes. This algorithm uses a technique called kernel trick to transform data and based on that transformations it finds optimal bounds between possible outputs.

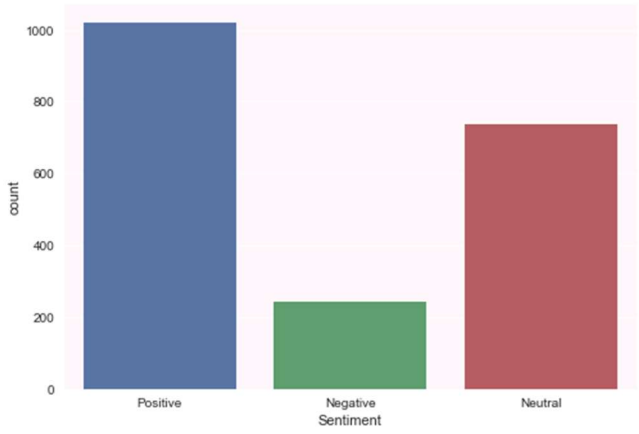
F1. Evaluation metrics

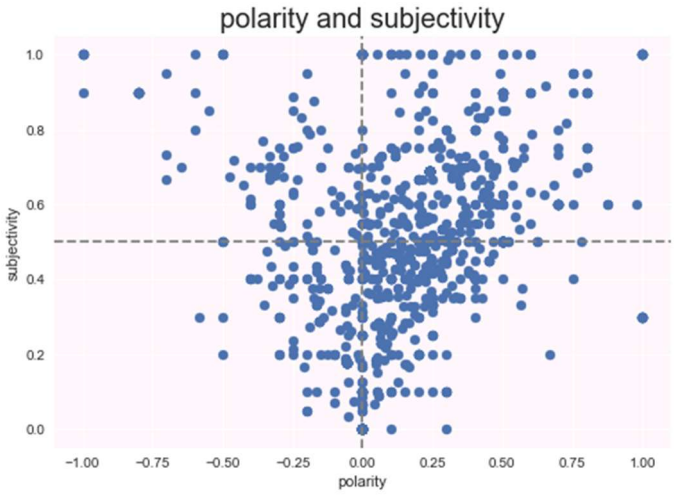
We try to check how accurate were classification was done over the testing dataset. We use sklearn.metrics to check how accuracy of classification done by each and every machine learning algorithm.

5. RESULTS AND OBSERVATIONS

In the whole process as we said earlier we try to analyze sentiment analysis of twitter data related any tech information. We performed our analysis over #windows11. We can observe how were sentiments of the users in fetched dataset, for that we created bar graphs, pie charts.







We can see the polarity vs subjectivity scatter plot, through which we can analyze both polarity and subjectivity at a time.

6. CONCLUSIONS

We implemented sentiment analysis over our target tweets. We have make out opinions of the users as either positive, negative or neutral. We have created few data visualizations to view user opinions.

From this we can conclude majority of users are positive and neutral towards our target subject and majority of the opinions are personal rather than factual. As we also tried to check how accurate the classifications were done we found logistic regression has better performance in our classification with accuracy scores.

```
Accuracy with Logistic Regression:  79.5
Accuracy with Naive-bayes:  71.0
Accuracy with DecisionTreeClassifier:  75.25
Accuracy with SVM:  75.25
```

7. REFERENCES

[1] Krouska, A., Troussas, C. and Virvou, M., 2016, July. The effect of preprocessing techniques on Twitter sentiment analysis. In 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA) (pp. 1-5). IEEE.

[2] Joshi, S. and Deshpande, D., 2018. Twitter sentiment analysis system.

[3] Hasan, A., Moin, S., Karim, A. and Shamshirband, S., 2018. Machine Learning- based sentiment analysis for twitter accounts. Mathematical and Computational Applications, 23(1), p.11.

[4] Rasool, A., Tao, R., Marjan, K. and Naveed, T., 2019, March. Twitter sentiment analysis: A case study for apparel brands. In Journal of Physics: Conference Series (Vol. 1176, No. 2, p. 022015). IOP Publishing.

[5] Verma, P., Khanday, A.M.U.D., Rabani, S.T., Mir, M.H. and Jamwal, S., 2019. Twitter sentiment analysis on Indian government project using R. Int J Recent Technol Eng, 8(3), pp.8338-41.

[6] Miranda, E., Aryuni, M., Hariyanto, R. and Surya, E.S., 2019, August. Sentiment Analysis using Sentiwordnet and Machine Learning Approach (Indonesia general election opinion from the twitter content). In 2019 International Conference on Information Management and Technology (ICIMTech) (Vol. 1, pp. 62-67). IEEE.

[7] Pokharel, B.P., 2020. Twitter sentiment analysis during covid-19 outbreak in nepal. Available at SSRN 3624719.

[8] Shamrat, F.J.M., Chakraborty, S., Imran, M.M., Muna, J.N., Billah, M.M., Das, P. and Rahman, M.O., 2021. Sentiment analysis on Twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indonesian Journal of Electrical Engineering and Computer Science, 23(1), pp.463-470.

[9] Bhumika Gupta, Monika Negi, Kanika Vishwakarma, Goldi Rawat, Priyanka Badhani , “Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python ”, *International Journal of Computer Applications (0975 – 8887) Volume 165 – No.9, May 2017* .

[10] Joylin Priya Pinto1, Vijaya Murari T.2 , ” Real Time Sentiment Analysis of Political Twitter Data Using Machine Learning Approach”, International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 04 | Apr 2019