

Sentiment Analysis on Twitter data using Machine Learning

Madikonda Jagadish¹, Cholleti Shiva Kumar², Dobbala Sandeep³, 4G Bhargavi

^{1,2,3}B. Tech Scholars, Department of Computer Science and Engineering, SNIST, Hyderabad-501301, India

⁴Assistant Professor, Department of Computer Science and Engineering, SNIST, Hyderabad-501301, India

Abstract - Twitter is a popular platform for people to express their thoughts and emotions on various occasions. Sentiment analysis is a method of analyzing data in order to extract the sentiment that it contains. Twitter sentiment analysis is the application of sentiment analysis to data from Twitter (tweets) in order to extract user sentiments. Over the last few decades, research in this field has steadily increased. The reason for this is the difficult format of the tweets, which makes processing difficult. Because the tweet format is so small, it creates a whole new set of issues, such as the use of slang and abbreviations. In this paper, we demonstrate the use of sentimental analysis as well as how to connect to Twitter and execute queries using sentiment analysis. We conduct tests on several issues, ranging from politics to humanity, and provide the intriguing findings. We discovered that the level of neutral sentiment for tweets is very high, which amply demonstrates the shortcomings of the existing works.

1. INTRODUCTION

Twitter has emerged as a major microblogging website, with over 100 million users daily sending out over 500 million tweets. Twitter's large audience has consistently drawn users to express their opinions and perspectives on any issue, brand, company, or another topic of interest. As a result, many organizations, institutions, and businesses use Twitter as a source of information.

Twitter users can express themselves in the form of tweets, which are limited to 140 characters. As a result, people condense their statements by using slang, abbreviations, emoticons, short forms, and so on. Along with this, people express themselves through sarcasm and polysemy. As a result, the term "unstructured" is appropriate for the Twitter language. To elicit emotion from Sentiment analysis involves determining the sentiment of a specific remark or sentence. It's a categorization technique that extracts opinion from tweets and creates a sentiment, which is individualized depending on the topic of interest. It's our responsibility to determine what characteristics will determine the feeling it conveys.

The class of entities that the person conducting sentiment analysis intends to find in the tweets is referred to as sentiment in the programming model. The sentiment class's dimension is a key aspect in determining the model's effectiveness. For instance, we may classify tweet sentiment

into two categories—positive and negative—or three categories (positive, negative and neutral). The class of entities that the person conducting sentiment analysis intends to find in the tweets is referred to as sentiment in the programming model. The sentiment class's dimension is a key aspect in determining the model's effectiveness.

Many businesses and organizations now utilize sentiment analysis to evaluate customer feedback on a product or their response to an event without the need for surveys or other pricey and time-consuming methods. One such social networking site, Twitter, one of the biggest networking sites, is considered in this thesis. According to the data, there are around 316 million active users monthly, and on average, 500 million tweets are sent each day.

II. LITERATURE SURVEY

Sentiment analysis involves determining the sentiment of a specific remark or sentence. It's a categorization technique that extracts opinion from tweets and creates a sentiment, which is individualized depending on the topic of interest. It's our responsibility to determine what characteristics will determine the feeling it conveys. The class of entities that the person conducting sentiment analysis intends to find in the tweets is referred to as sentiment in the programming model. The sentiment class's dimension is a key aspect in determining the model's effectiveness.

For instance, we may classify tweet sentiment into two categories—positive and negative—or three categories (positive, negative and neutral). The class of entities that the person conducting sentiment analysis intends to find in the tweets is referred to as sentiment in the programming model. The sentiment class's dimension is a key aspect in determining the model's effectiveness. learning approach uses feature extraction while training the model with a feature set and dataset.

III. DESIGN AND IMPLEMENTATION

Through the use of Twitter's own APIs, this technical paper documents the implementation of Twitter sentiment analysis. For text mining on social networks, there are excellent resources and tools. The full range of the libraries utilized in this project has been available.

We use following approaches to extract sentiment from the tweets.

1. Download and cache the sentiment dictionary first.
2. Download the testing data sets for Twitter and enter them into the software.
3. Remove the stop words from the tweets to clean them up.
4. Tokenize all of the dataset's words before feeding them to the program.
5. For each term, contrast it with the dictionary's definitions of words having positive and negative connotations. Then raise either the positive or negative count.
6. In order to determine the polarity, we can calculate the outcome percentage based on the positive and negative counts.

3.1 IMPLIMENTATION

Python was used in this study to implement sentimental analysis. Several packages have used it, notably textblob and tweepy. The commands listed below can be used to install the necessary libraries:

Install tweepy via pip.

Install textblob via pip.

Textblob:

A Python (2 and 3) package called TextBlob is used to process textual data. It offers a straightforward API for getting started with typical natural language processing (NLP) activities like part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and others.

Tweepy:

You may access the Twitter API with Python quite conveniently using the open source Tweepy library. Tweepy contains a collection of classes and methods that represent the models and API endpoints of Twitter. User's need to go to the apps.twitter.com/app/new and generate the API keys.

With following steps, we can connect twitter API with python:

Create a free Rapid API user account (or log in).

Open the Twitter API page on Rapid API.

As soon as you click "Connect to API," you can start entering the parameters and fields for your API Key.

Start the Twitter API Endpoints testing.

The below figure says about connecting of python with twitter API

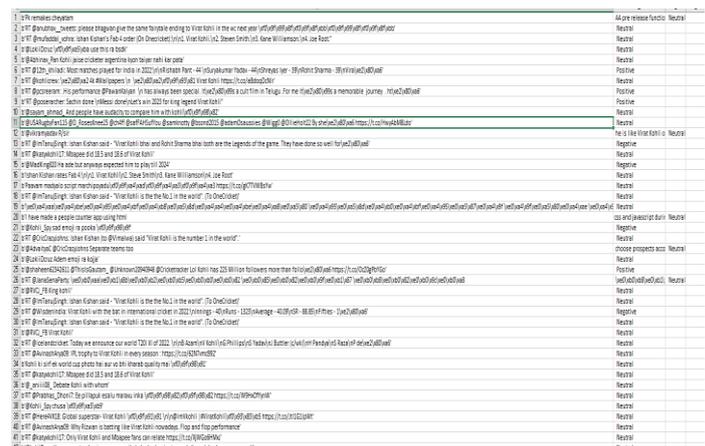
```
# complete authorization and initialize API endpoint
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)

# initialize stream
streamListener = StreamListener()
stream = tweepy.Stream(auth=api.auth, listener=streamListener, tweet_mode='extended')
```

Fig 1. User API Keys

3.1 Dataset:

The data is retrieved form the twitter using API and that is stored in the csv file for the data visualization. This stores the tweets that are retrieved form the twitter API and the user can see the data for clarification.



Tweet	Sentiment
1. The weekend streamer...	Neutral
2. I'm @Ashley... who's...	Neutral
3. I'm @Ashley... who's...	Neutral
4. I'm @Ashley... who's...	Neutral
5. I'm @Ashley... who's...	Positive
6. I'm @Ashley... who's...	Positive
7. I'm @Ashley... who's...	Positive
8. I'm @Ashley... who's...	Positive
9. I'm @Ashley... who's...	Positive
10. I'm @Ashley... who's...	Positive
11. I'm @Ashley... who's...	Positive
12. I'm @Ashley... who's...	Positive
13. I'm @Ashley... who's...	Positive
14. I'm @Ashley... who's...	Positive
15. I'm @Ashley... who's...	Positive
16. I'm @Ashley... who's...	Positive
17. I'm @Ashley... who's...	Positive
18. I'm @Ashley... who's...	Positive
19. I'm @Ashley... who's...	Positive
20. I'm @Ashley... who's...	Positive
21. I'm @Ashley... who's...	Positive
22. I'm @Ashley... who's...	Positive
23. I'm @Ashley... who's...	Positive
24. I'm @Ashley... who's...	Positive
25. I'm @Ashley... who's...	Positive
26. I'm @Ashley... who's...	Positive
27. I'm @Ashley... who's...	Positive
28. I'm @Ashley... who's...	Positive
29. I'm @Ashley... who's...	Positive
30. I'm @Ashley... who's...	Positive
31. I'm @Ashley... who's...	Positive
32. I'm @Ashley... who's...	Positive
33. I'm @Ashley... who's...	Positive
34. I'm @Ashley... who's...	Positive
35. I'm @Ashley... who's...	Positive
36. I'm @Ashley... who's...	Positive
37. I'm @Ashley... who's...	Positive
38. I'm @Ashley... who's...	Positive
39. I'm @Ashley... who's...	Positive
40. I'm @Ashley... who's...	Positive
41. I'm @Ashley... who's...	Positive
42. I'm @Ashley... who's...	Positive
43. I'm @Ashley... who's...	Positive
44. I'm @Ashley... who's...	Positive
45. I'm @Ashley... who's...	Positive
46. I'm @Ashley... who's...	Positive
47. I'm @Ashley... who's...	Positive
48. I'm @Ashley... who's...	Positive
49. I'm @Ashley... who's...	Positive
50. I'm @Ashley... who's...	Positive

Fig 2. Dataset with tweets

IV. TWITTER SENTIMENT ANALYSIS WITH PYTHON

4.1 Python:

Python is a preferred programming language because of its extensive capabilities, applicability, and simplicity. Due to its independent platform and widespread use in the programming community, the Python programming language is the most suitable for machine learning. The requirement for intelligent answers to practical issues needs the further development of AI in order to automate laborious processes that would be difficult to program without AI. The Python programming language is thought to be the ideal technique for automating these processes since it is more straightforward and consistent than other programming languages. Additionally, the vibrant Python community makes it simple for developers to discuss projects and offer suggestions for improving their code.

4.2 Tweepy:

You may access the Twitter API with Python quite conveniently using the open-source Tweepy library. Tweepy contains a collection of classes and methods that represent the models and API endpoints of Twitter.

It also handles following things like data encoding and decoding

The following figure say about setup of tweepy API.

```
from textblob import TextBlob
import tweepy
import csv
import matplotlib.pyplot as plt

consumer_key = '#####'
consumer_secret = '#####'

access_token = '#####'
access_token_secret = '#####'

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

pos = 0
neu = 0
neg = 0

positive = []
neutral = []
negative = []

tweet_s = input("Search pattern: ")
public_tweets = api.search_tweets(tweet_s, count=100)
with open('tweets_sentiment.csv', 'w') as file:

    for tweet in public_tweets:
        text_tweet = TextBlob(tweet.text)
        print(text_tweet)
        analysis = TextBlob(tweet.text)
        print(analysis.sentiment)
        if analysis.sentiment.polarity > 0:
            sent = "Positive"
            pos += 1
        elif analysis.sentiment.polarity == 0:
            sent = "Neutral"
            neu += 1
```

Fig 3. Tweepy API connection

4.3 Textblob:

A Python library for Natural Language Processing is called TextBlob (NLP). Natural Language Toolkit (NLTK) was a tool that TextBlob actively employed to complete its tasks. The NLTK library enables users to do categorization, classification, and a variety of other tasks while providing quick access to a large number of lexical resources. TextBlob is a straightforward library that provides intricate textual analysis and processing.

A sentiment is identified by its semantic orientation and the force of each word in the sentence for lexicon-based techniques. This calls for a pre-defined dictionary that divides words into negative and positive categories. A text message will typically be represented by a bag of words. Following the individual scoring of each word, the ultimate

sentiment is determined by performing a pooling procedure, such as averaging all the sentiments.

TextBlob returns a sentence's polarity and subjectivity. The polarity scale is [-1,1], where -1 represents a negative emotion and 1 represents a good emotion. Negative words turn the polarity around. Semantic labels in TextBlob facilitate detailed analysis. Emoticons, exclamation points, emoticons, etc. are a few examples. The range of subjectivity is [0, 1]. Subjectivity measures how much factual information and subjective opinion are present in the text. The content contains personal opinion rather than factual information due to the text's heightened subjectivity. One other setting for TextBlob is intensity. TextBlob uses the "intensity" to determine subjectivity. Whether a word modifies the next word depends on its intensity. Adverbs are used as modifiers in English, such as "extremely good."

4.3 NLTK:

The Natural Language Toolkit (NLTK) is a Python programming environment for creating applications for statistical natural language processing (NLP).

Steven Bird, Edward Loper, and Ewan Klein created the Natural Language Toolkit as an open-source library for the Python programming language with the intention of using it for development and education.

It is appropriate for linguists without extensive programming experience, engineers and researchers who need to delve into computational linguistics, students, and educators because it includes a hands-on guide that introduces topics in computational linguistics as well as Python programming fundamentals.

To gain insights from linguistic data, you can use these methods with NLTK using robust built-in machine learning procedures. Tasks like tokenization, stemming, lemmatization, punctuation, character count, word count, etc. can be accomplished with this library. This does an analysis of the data and produces the necessary results.

4.4 Matplotlib:

For Python and its numerical extension NumPy, Matplotlib is a cross-platform package for graphical data visualization and charting. This makes it a strong open-source substitute for MATLAB. The APIs (Application Programming Interfaces) of matplotlib can also be used by developers to integrate plots into GUI programmers.

The way a Python matplotlib script is written makes it possible to create a visual data plot in the majority of cases with just a few lines of code. Overlaying two APIs is the Matplotlib scripting layer.

The matplotlib object is the top-level Python code object in the pyplot API hierarchy.

An Object-Oriented API collection of objects that is more flexible than pyplot in how it may be put together. The backend layers of Matplotlib are directly accessible through this API.

IV. RESULT

The tweets are received from Twitter using the API, analyzed, and the results are shown in the below pie chart. The below figure shows the tweets that are retrieved from the twitter.

```

RT @kalydkohli17: Test cricket will never get a better captain than Virat Kohli https://t.co/dw0hPQ50M
Sentiment(polarity=0.5, subjectivity=0.5)

RT @kohlicrew: • dk Mailpapers
Virat Kohli https://t.co/andop0air
Sentiment(polarity=0, subjectivity=0.0)

RT @cric_beat: Most runs in Test matches
20140 - Ricky Ponting
17111 - Sachin Tendulkar
15977 - Virat Kohli
14827 - Jacques Kallis
Happy...
Sentiment(polarity=0.5, subjectivity=0.5)

RT @m147: Kriket experts are saying @m005 (India's best batsman over the years) should play Ranji before coming in the test cricke.
Sentiment(polarity=1.0, subjectivity=0.3)

@kohli_spy https://t.co/cH9y4xvml
Sentiment(polarity=0.0, subjectivity=0.0)

@kohli_p0_drf https://t.co/060kxw703
Sentiment(polarity=0.0, subjectivity=0.0)

@kohli_spy they deserve
Sentiment(polarity=0.0, subjectivity=0.0)

RT @m1au5j5ngb: Virat Kohli is the the most popular hashtags used on reels in 2022 in India. (According to Meta reports)
Sentiment(polarity=0.55, subjectivity=0.7)

RT @m1au5j5ngb: Virat Kohli is the the most popular hashtags used on reels in 2022 in India. (According to Meta reports)
Sentiment(polarity=0.55, subjectivity=0.7)

RT @sidsenIndia: virat kohli with the bat in international cricket in 2022:
Innings - 40
Runs - 1923
Average - 48.09
SR - 88.85
Fifties - 1
Sentiment(polarity=-0.075, subjectivity=0.19999999999999998)
    
```

Fig 4. Tweets retrieved form twitter

The analysis's findings, which show different people's opinions on numerous issues, are summarized in the pie chart below. In order to analyze their product or business, these tweets are analyzed and results are recorded. the analysis' findings are based on a variety of queries, including those related to movies, politics, fashion, and more. The data based on the tweets retrieved are illustrated by the pie chart in figure 5. Based on the tweets we retrieve, if we run the program at other times, we can receive slightly different results.

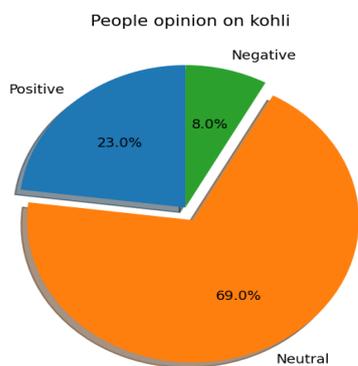


Fig 5. Output of the tweets analysis

This shows the analysis of the tweets about Virat Kohli based on the latest 100 tweets and more tweets can also be retrieved based on the user's needs.

Three different categories are defined as positive, negative and neutral tweets.

In the above pie chart, the results are as follows:

- Positive tweets percentage: 23.0 %
- Negative tweets percentage: 69.0 %
- Neutral tweets percentage: 8.0 %

The fraction of neutral tweets is notably high, as shown in the pie chart. It's also crucial to note that, depending on the experiment's data, we can obtain various conclusions because people's opinions can alter in response to external factors.

V. CONCLUSION

Twitter sentiment analysis comes under the category of text and opinion mining. It focuses on examining the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then test its precision, so that we may use this model going forward based on the results. It entails actions including gathering data, text pre-processing, sentiment categorization, sentiment detection, model training, and testing. The models used in this research have improved over the past ten years, attaining efficiencies of roughly 85%–90%. However, the dimension of data variety is still missing. In addition, it has numerous application problems due to the slang and abbreviations employed. The performance of many analyzers suffers as the number of classes rises. Therefore, there is a very promising future for the advancement of sentiment analysis.

REFERENCES

- [1] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREC (Vol. 10, No. 2010).
- [2] TextBlob, 2017, <https://textblob.readthedocs.io/en/dev/>
- [3] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
- [4] Neethu MS and Rajashree R, " Sentiment Analysis in Twitter using Machine Learning Techniques" 4th ICCNT 2013, at Tiruchengode, India. IEEE – 31661
- [5] Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research, 50, 723-762.

- [6] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38). Association for Computational Linguistics.
- [7] Pang, B. and Lee, L. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts". 42nd Meeting of the Association for Computational Linguistics[C] (ACL-04). 2004, 271-278.
- [8] Rosenthal, S., Farra, N., & Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017) (pp. 502-518).
- [9] Nehal Mangain, Ekta Mehta, Ankush Mittal and Gaurav Bhatt, "Sentiment Analysis of Top Colleges in India Using Twitter Data", (IEEE) ISBN -978-1-5090-0082-1, 2016.