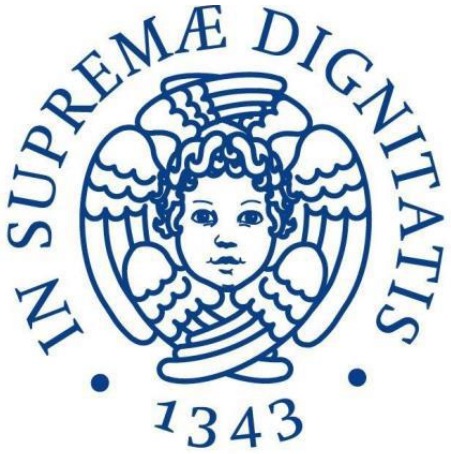


# UNIVERSITÀ DI PISA

**MSc Data Science and Business Informatics**
**Text Analytics Final Report**
**2022/2023**

**Team Members**

**Muzammil Raza Soomro - 665575**

**Thaharim Khan - 648207**

**Muhammad Adnan- 646867**

**Muhammad Shayan- 665344**

**Professor:**
**Lucia C. Passaro**

**Introduction:**

Twitter has grown to become one of the most prominent social media platforms, and its popularity is growing exponentially each day as the number of tweets reaches millions. A common Web information mining research topic is sentiment analysis of online reviews. Emotion dictionaries or machine learning are used in the classic method of text sentiment analysis.

People all over the world have become increasingly passionate about expressing their emotions and thoughts on the internet in recent years, thanks to the rapid expansion of the Internet. As a result, sentiment analysis has emerged as a critical research area in Natural Language Processing. Traditional database solutions are incapable of processing millions of information; hence big data is the newest buzzword. Capturing data, analyzing data, searching, storing, sharing visualization, and maintaining data privacy are all challenges addressed by big data

Twitter is used in this project to do predictive analysis and extract values from data. Using a standard relational database to process huge data will take more time, money, and resources. To improve the big data architecture, big data can be combined with data lakes, machine learning, and artificial intelligence to construct analytical algorithms. The data received from tweets can be used to categories tweets into good and negative categories. The most prevalent strategies for extracting feelings from textual input are used lexicons and machine learning approaches. Regular texts as well as texts with a lot of noise in the data can benefit from context-based sentimental analysis.

**Challenges in Sentiment Analysis**:

It's a difficult undertaking to conduct a sentiment analysis. Some of the difficulties encountered in Twitter Sentiment Analysis while doing the project are listed below.

i. Sarcasm Detection: - Sarcastic sentences use positive words to express a negative view about a target in an unusual way. For example, “When I die, I want Sarfaraz to lower me into my grave so he can let me down one last time.” Although the sentence contains solely positive words, it conveys a negative message.

ii. Domain dependence: - In different realms, the same language or phrase can have distinct meanings. For example, the phrase "unpredictable Match" has a favorable connotation in the realm of movies, plays, and the like, but it has a bad connotation when applied to a vehicle's steering.

iii. Identifying text's subjective elements: - Sentiment-bearing content is represented by subjective portions. In some cases, the same term can be viewed as subjective, while in others, it can be treated as objective. This makes it difficult to distinguish between objective and subjective text.

iv. Entity Recognition: - a. Separating material concerning a certain entity and then analysing sentiment towards it is necessary. b. “I like the way how Iftikhar Ahmed batted, but I enjoy the whole innings of Babar Azam,” for example. A simple bag-of-words method labels it as neutral, but it has a distinct attitude toward both entities in the statement.

v. Internationalization: - The current research focuses primarily on English content; however, Twitter has a diverse user base from all over the world.

**Workflow of the projects:**

Detailing workflow of our project is given below. where we can get a clear overview of the overall procedures of the project.

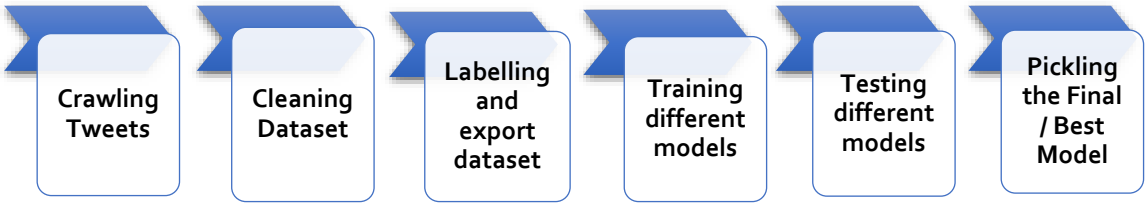


Figure 1: Working procedure of the project.

Figure 1 gives us the detailing overview of the project. Firstly, we need to collect the dataset from the twitter with using twitter API. After collecting the data, we have cleaned it and labelled it. we have also exported the dataset. we tried to applied various Machine learning model and also applying the NLTK for analyzing the sentiment again tested it finally choose the best model among all the tested model. we have also tried the manual testing for checking the validity of our model

**Dataset Collection:**

Data crawling is a method which involves data mining from different web sources. Data crawling is very similar to what the major search engines do. In simple terms, data crawling is a method for finding web links and obtaining information from them. For this project we have collected data from tweeter through API. Below figure 2 we give a screenshot for a piece of code which we have used to collect data from tweeter.

```
'''
You will be required to insert your own codes to complete this function.
Walk through this function and enter your own codes where instructed.
'''
def retrieve_tweets(api, keyword, batch_count, total_count, latitude, longitude, radius):
    """
    collects tweets using the Twitter search API

    api:         Twitter API instance
    keyword:     search keyword
    batch_count: maximum number of tweets to collect per each request
    total_count: maximum number of tweets in total
    """


    # the collection of tweets to be returned
    tweets_unfiltered = []
    tweets = []

    # the number of tweets within a single query
    batch_count = str(batch_count)

    '''
    You are required to insert your own code where instructed to perform the first query to Twitter API.
    Hint: revise the practical session on Twitter API on how to perform query to Twitter API.
    '''
    # per the first query, to obtain max_id_str which will be used later to query sub
    resp = api.request('search/tweets', {'q': keyword,
                                         'count': batch_count,
```

Figure 2 : API Code using for Data Collection

After using this piece of code, we are able to collect a bunch of datasets below figure 3 show in the dataset for our project.

| | id | text | hashtags | created_at | user followers count | replycount |
|---|---|---|---|---|---|---|
| 0 | 1584689793698791424 | Hangover of #INDvsPAK2022 is still around. #T2... | ['INDvsPAK2022', 'T20WorldCup'] | 2022-10-24 23:34:23+00:00 | 538 | 0 |
| 1 | 1584684371512291330 | @TheBarmyArmy luv it \n\n ❤️❤️❤️ \n\n#INDvPAK \n... | ['INDvPAK', 'indvspakmatch', 'INDvsPAK2022'] | 2022-10-24 23:12:50+00:00 | 24 | 0 |
| 2 | 1584684147339296768 | #INDvsPAK2022 | ['INDvsPAK2022'] | 2022-10-24 23:11:57+00:00 | 56 | 0 |
| 3 | 1584684139839905792 | No ball thi #INDvsPAK2022 | ['INDvsPAK2022'] | 2022-10-24 23:11:55+00:00 | 56 | 0 |
| 4 | 1584681884210319360 | @mufaddal_vohra @ImRo45 \n@BCCI \n\nWhen will\... | ['INDvPAK', 'INDvsPAK2022', 'T20Cricket'] | 2022-10-24 23:02:57+00:00 | 24 | 1 |

Figure 3: Dataset collected from tweets

We can get a clear understanding of the amount of our data with a quick glimpse at the below table

| **API** | **Twitter API** |
|---|---|
| **Total Tweets Collected** | **272270** |
| **Unique Tweets** | **260884** |
| **Duplicate Tweets** | **11386** |
| **Users** | **143879** |
| **Hashtags used** | **#PAKvsIndia, INDvPAK, #INDvsPAK2022** |
| **Date Range** | **2022-10-18 to 2022-10-24** |
| **Language classified by tweeter** | **54** |

Table 1: Classification of the data from the dataset

**Cleaning Dataset**

At the first part of cleaning dataset, we used to choose word cloud for visualizing our data. The main reason for visualizing data is to get a clear view about the data so that we can able to manage the cleaning process in a very good manner. We used word cloud and Word Cloud is **a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance**. Significant textual data points can be highlighted using a word cloud. After **visualizing we got some important textual data point like "Virat Kohli", "King Kohli", "Ind VS Pak", "t 20 world cup". Below figure 4 showing the word cloud from our dataset.**

Figure 4: Word cloud for the dataset

From the word cloud we get that we need to focus more on this set of word from which we could get a better understanding for our sentiment analysis.

The list of our cleaning work cloud be,

(i)remove the duplicate tweets from our dataset the amount of duplicate tweet on our dataset is around 11386 so that we can work on remaining 260884 data to avoid redundancy.

(ii) we have removed stop words.

(iii) we have removed html tags from the text and

(iv) We have removed urls as well.

(v) remove character references from the text

(vi) remove repeated characters in elongated words

Below figure 5 shows the cleaning text after proceedings the data cleaning procedure.

| | id | text | text_cleaned |
|---|---|---|---|
| 1 | 1584684371512291330 | @TheBarmyArmy luv it \n\n❤️❤️❤️\n\n#INDvPAK \n... | luv it indvpak indvspakmatch |
| 2 | 1584684147339296768 | #INDvsPAK2022 | |
| 3 | 1584684139839905792 | No ball thi #INDvsPAK2022 | no ball thi |
| 4 | 1584681884210319360 | @mufaddal_vohra @ImRo45 \n@BCCI \n\nWhen will\... | when will ever learn indvpak |
| 5 | 1584680546005037056 | What's your zodiac sign ?And what's your Fav S... | what s your zodiac sign and what s your fav si... |

Figure 5: Cleaned dataset

In the middle we have done some auto labelling without cleaning for getting an idea about the amount classified and not classified data. Below figure 6 showing the amount of data with their percentile amount.

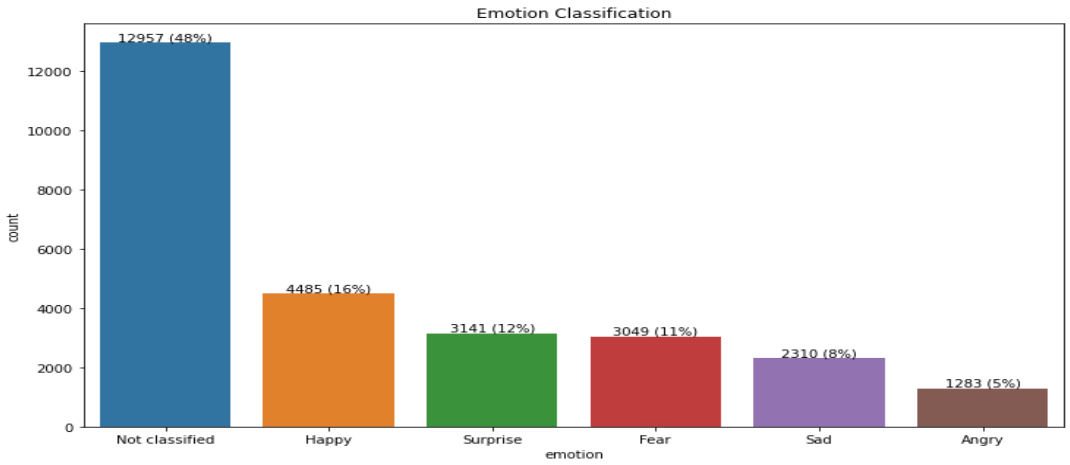

Figure 6: Auto labelling without cleaning

The figure stated above give us an idea that among all the dataset we have almost 49% of data which is not classified and rest of them are classified within the category of happy, fear, surprised, sad and angry. With the big amount of not classified data we cannot even analyze the sentiment so we have decided to label it so that we can analyze the sentiment

**Labeling:**

After getting the clean dataset we are trying to label our data so that we can get a list of emotions. Data labeling is the process of adding labels to raw data to give context or meaning to the data. The labeled data is then used to train the NLP models to make predictions or understand or generate speech. We used in our project *Text2emotion* python package for labeling our dataset. get_emotion () function is used here for doing this process after that we export it for getting all labeled data in an Excel file an. Below figure 7 showing the data which, we got after labelling our dataset

| text | sentiment | emotion |
|---|---|---|
| @mufaddal_vohra @ImRo45 \n@BCCI \n\nWhen will\... | 0 | Fear |
| What's your zodiac sign ?And what's your Fav S... | 1 | Surprise |
| @adidoescricket @TheRealPCB \n@ICC\n\nHe will ... | 1 | Happy |
| @RJ_Balaji kirukku koodhian.. Spoiled the imp... | 1 | Surprise |
| Bhai log hum hindu hein isiliye toh Pakistan a... | 0 | Surprise |

Figure 7: Labelled dataset.

**Sentiment Intensity Analyzer using NLTK:**

NLTK consists of the most common algorithms such as tokenizing, part-of-speech tagging, stemming, sentiment analysis, topic segmentation, and named entity recognition. NLTK helps the computer to analysis, preprocess, and understand the written text. NLTK contains useful tools for text preprocessing and corpora analysis we did not need to create our own stop words list or frequency function for this project.

After applying NLTK we define a function using NTLK that returns the number of proper nouns in a string below figure showing the noun count. then create a new feature for the number of proper nouns in each tweet after that we have checked the result.

| | id | text | text_cleaned | propn_count |
|---|---|---|---|---|
| 1 | 1584684371512291330 | @TheBarmyArmy luv it \n\n❤️❤️❤️\n\n#INDvPAK \n... | @TheBarmyArmy luv it \n\n\n\n#INDvPAK \n#indvs... | 2 |
| 2 | 1584684147339296768 | #INDvsPAK2022 | # | 0 |
| 3 | 1584684139839905792 | No ball thi #INDvsPAK2022 | No ball thi # | 0 |
| 4 | 1584681884210319360 | @mufaddal_vohra @ImRo45 \n@BCCI \n\nWhen will\... | @mufaddal_vohra @ \n@BCCI \n\nWhen will\n\n@kl... | 4 |
| 5 | 1584680546005037056 | What's your zodiac sign ?And what's your Fav S... | Whats your zodiac sign ?And whats your Fav Sig... | 8 |

Figure 8: proper nouns Noun count.

For analyzing the report, we have tested how NLTK worked with the first text after we have tested how NLTK works with the first text after lowercasing it.

**(i)Polarity**

Analyzing sentiment without context gets difficult as machines cannot learn about contexts if it is not trained explicitly. The most crucial disadvantage that arises from context is changing in polarity. So, we have checked the SentimentIntensityAnalyzer with polarity score which is most important while using the SentimentIntensityAnalyzer. Below figure 9 showing the amount of sentiment on the basis of positive and negative

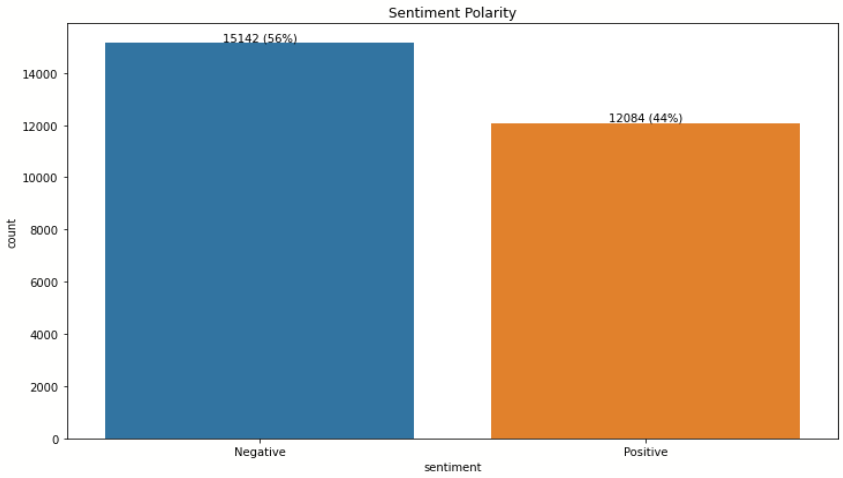


Figure 9: Sentiment analysis based on polarity score

From the figure above we could get the clear view about the quantity of positive and negative tweets based on sentiment score.

**Machine Learning for Evaluation:**

For this project we are using real data set which we have collected from tweeter after implementing all the process and for better understanding of our data we are trying to implement some machine learning model like MultinomialNB, SVC, LogisticRegression, KNeighborsClassifier, RandomForestClassifier, ExtraTreesClassifier, BaggingClassifier, XGBClassifier below figure showing us the accuracy report of all the classification model.

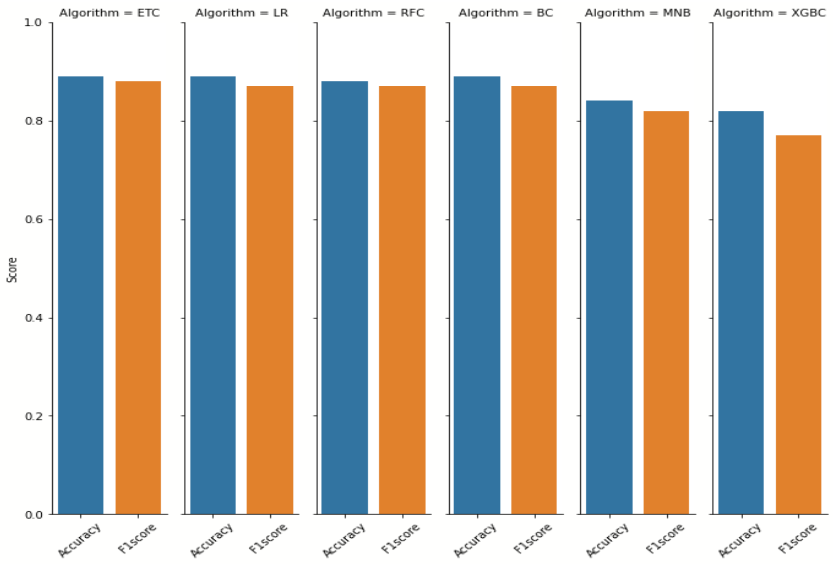


Figure 10: Accuracy chart of the classification model.

Below figure showing the accuracy score for each of the model.

| | Algorithm | Accuracy | F1score |
|---|---|---|---|
| 3 | ETC | 0.89 | 0.88 |
| 1 | LR | 0.89 | 0.87 |
| 2 | RFC | 0.88 | 0.87 |
| 4 | BC | 0.89 | 0.87 |
| 0 | MNB | 0.84 | 0.82 |
| 5 | XGBC | 0.82 | 0.77 |

Figure 11: Accuracy score

Among all the classification model we got top three model with MultinomialNB (), SVC, and logistic regression with a good accuracy so we proceed further with this top three model. Again, we have applied voting classifier for getting the best one among those three and stacking classifier as well. Below table showing the accuracy score after implementing the voting classifier and stacking classifier.

Voting Classifier

| Accuracy | F1score |
|---|---|
| 0.8868894601542416 | 0.8752025931928686 |

Stacking Classifier

| | |
|---|---|
| Accuracy | 0.8920308483290489 |
| F1score | 0.8822587104525432 |

**Vectorization:**

After going for further procedure, we have used the cleaned text for vectorization. While managing the vectorization we have used tf-idf vectorization the main task of this vectorization is converting text data to numerical vectors. Later we have used those vectors to fit into the logistic regression model. we fitted the model with it and check the processing time and again make predictions and check the processing time again finally we have checked the accuracy. Below figure 12 showing the vectorized text of our data.

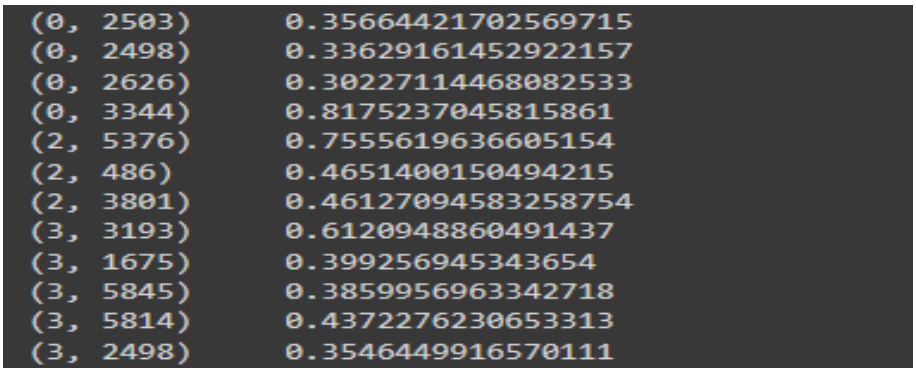


Figure 12 : vectorized text of our data

**Validation of the model:**

As we have worked on pretrained model with our dataset. So that for validating it we have chosen the way of validating it with some labelling data. we have taken the labelled data by our own for checking the accuracy of our model.

| text | sentiment | emotion |
|---|---|---|
| Two Good News On This Diwali, One Is The Victo... | 1 | Happy |
| ...احمد بھائی مجھے لگ تھا ہے کہ فخر زمان کو 1 ڈ | 0 | Fear |
| Nawaz is the champion.\nPakistan fought with t... | 1 | Happy |
| Inbox now for all account recovery services,Lo... | 0 | Surprise |
| By the end of this world cup, we will be aware... | 0 | Happy |

Figure 13: Manual labeling for model Validation

Manual Labelled Data validated on final model

| Accuracy | F1score |
|---|---|
| 0.8948374760994264 | 0.8801742919389979 |

**Conclusion:**

The age of getting meaningful insights from social media data has now arrived with the advance in technology. It's time for organizations to move beyond overall sentiment and count based metrics. Companies have been leveraging the power of data lately, but to get the deepest of the information, one has to leverage the power of AI, Deep learning and intelligent classifiers like Contextual Semantic Search and sentiment analysis.