Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 2458-9403
Vol. 7 Issue 3, March - 2020

Sentimental Analysis of Twitter Data to Rank Best ICC Player

Md. Toukir Ahmed*, Md. Niaz Imtiaz, Md. Masum Billah
Department of Computer Science and Engineering
Pabna University of Science and Technology, Pabna, Bangladesh
*Corresponding Author: toukirahmedreal@gmail.com

***Abstract*— Sentiment analysis is a type of natural language processing which categorizes a sentence into positive, negative or neutral. Nowadays, sentiment analysis is being used extensively in review analysis, predicting outcome of a particular event in social media, incentivized marketing detection and so on. In this work, a tool for sentimental analysis of twitter data to rank International Cricket Council (ICC) authorized best players is presented. The dataset contains 5, 000 tweets regarding ten most popular cricketers. The search Application Program Interface (API) allows developers to query specific twitter content, whereas the Streaming API is able to collect twitter content in real-time. Automated collection of ICC best player dataset for sentiment analysis and opinion mining purposes is performed and an experimental evaluation and classification is also assessed. An implementation of unsupervised method for sentimental analysis is carried out. Lastly the best players are predicted based on the tweets shared by their fans using Twitter.**

Keywords—: Classification, Machine Learning, Natural Language Processing, Opinion Mining, Sentiment Analysis, Twitter Data

I. INTRODUCTION

Nowadays, we live in an age which people are used to internet and others technologies for help. Majority percent people are dependent on technologies. Twitter is one of those microblogging sites and widely used platform for emotions manifestation & flooding the views to intended community. This assistance of twitter has turned as the habit of users. Cricket is like religion for subcontinent and widely appreciated in other regions too. So it cannot stay away untouched from tweeting trends. Throughout the past, it is observed that Indians are very emotionally attached with cricket. This gives us idea to capture these flowing emotions of Indian cricket lovers. Here, we use data collected from twitter which is in form of messages. The content of messages varies from personal to social views. Natural Language opinions are expressed in restrained and multifarious ways, which are difficult to solve by basic text processing methodologies. Recognizing the sentiment and sub-events correctly is more tedious due its unrestricted message format. Sentiment analysis is a fancy word that data scientists use instead of emotion detection. Sentiment or opinion mining refers to the type of natural language processing used to understand the moods, opinions and sentiments of the public regarding a particular product or a movie or an event [1]. In this work, we show how to calculate ICC Best Player Rating use sentimental analysis in tweeter data set. We intended to build a system which generates a rating of ICC best cricketer . This rating is solely dependent on public opinion on a player. Depending on public opinion, a products rating would be given.

II. LITERATURE SURVEY

Sentiment analysis is basically the automation of the analysis of a given text in order to determine the feelings conveyed in it. Sentiment analysis and opinion mining have become known as interchangeable terms. Sentiment analysis is defined by Subhadra Mukherjee [1] as a “Natural Language Processing and Information Extraction task that aims to obtain writer’s feelings expressed in positive or negative comments, questions and requests, by analyzing a large number of documents”. In other words, sentiment analysis intends to define the feelings of the writer regarding a particular topic based on the writer’s opinion. Sentiment analysis is important as it can help to provide insight into different fields. Even when sentiment analysis is not perfect, because the sentiment itself is really 3 subjective, there is no doubt that processing and analyzing existent opinionated data has only just begun. Even when social media monitoring tools such as Tracker and Mention claim that their Sentiment Analysis accuracy is over 70 percent most of the information found, such as [2] claim that anyone who says that they are getting more than 70 percent accuracy is lying. This is in agreement with [3] which states that human raters typically only agree 79 percent of the time, making this really difficult to automate systems to achieve high accuracy. Sentiment Analysis is a field that is growing fairly rapidly. 81 percent of Internet users (or 60 percent of Americans) have done online research on a product at least once meaning [4] every year there are more articles targeting different text domains over years, where the reviews represent around the 49.12% of the articles [5] One would not always want to apply sentiment to product reviews; there are too many other fields. One good example of this that has been experimented [6] is the comparison of Twitter sentiment versus Gallup polls of consumer confidence. The results yielded were positive and the correlation was 0.804, inferring that we can use Twitter to measure public opinion. This is precisely what we are going to use Twitter for during this

experiment: to extract opinions from it and determinate the tweets' polarity in real-time.

III. METHODOLOGY

This technical paper reports the implementation of the Twitter sentiment analysis, by utilizing the APIs provided by Twitter itself. There are great works and tools focusing on text mining on social networks. In this project the wealth of available libraries has been used.

A. The approach to extract sentiment

The approach used in this study to extract sentiment from tweets is as follows:

- Downloading and caching the sentiment dictionary
- Downloading twitter testing data sets, input it in to the program.
- Cleaning the tweets by removing the stop words.
- Tokenizing each word in the dataset and feeding it in to the program.
- For each word, comparison with positive sentiments and negative sentiments word in the dictionary. Then incrementing positive count or negative count.
- Finally, based on the positive count and negative count, generating result percentage about sentiment to decide the polarity.

B. Implementation Model

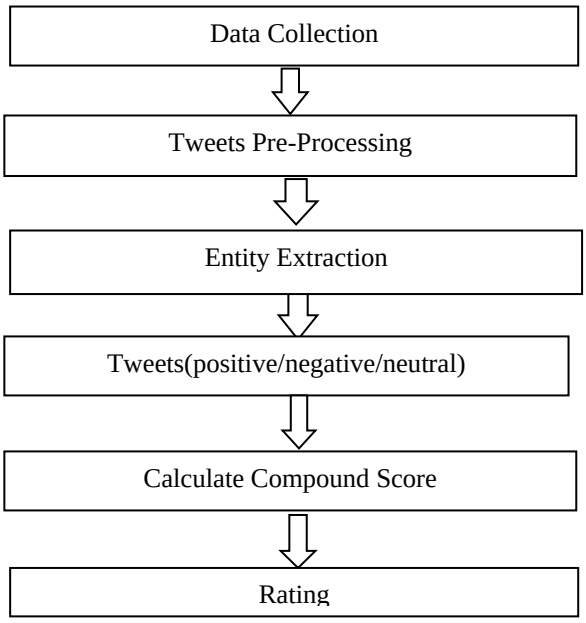

Fig. 1: Building a model for players ranking based on twitter data.

IV. DATA PRE-PROCESSING

As we are dealing with text data for sentiment analysis, data preprocessing plays a vital role on research to make the model understand the data. Text data contains a lot of noise. As a result, it's a challenge to clean the texts smartly. Data pre-processing reduces the size of the input text documents significantly. The following steps are performed for pre-processing.

A. Stop-word Elimination

A **stop word** is a commonly used **word** (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

Table 1. example of stop word elimination

Sample text with stop words	**Without Sop Words**
I like Reading, so I read	Like, Reading, Read
Can listening be exhausting	Listening, Exhausting
Living with hearing loss	Living, hearing, loss

B. Tokanization

In Python tokenization basically refers to splitting up a larger body of text into smaller lines, words or even creating words for a non-English language. The various tokenization functions in-built into the nltk module itself and can be used in programs as shown in Table 2 below.

Table 2. Example of Tokenization

Before Tokenization	After Tokenization
Shakib Al Hasan really star of Bangladesh	“Shakib”,” Al”,” Hasan”, “Really” “star”,” of”, “Bangladesh”
Babar Azam The Run Machine	“Babar”,” Azam”,” The”, “Run”, “Machine”
What is the role of Imad wasim in the team?	“What”, “is”,” The”,” role”, “of”, “Imad”,” wasim”,”in”,” the”, “team”

C. NER using Spacy

SpaCy [7] is an open-source library for advanced Natural Language Processing in Python. It is designed specifically for production use and helps build applications that process and “understand” large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning. Some of the features provided by SpaCy are- Tokenization, Parts-of-Speech (PoS) Tagging, Text Classification and Named Entity Recognition’s provides an exceptionally efficient statistical system for NER in python, which can assign labels to groups of tokens which are contiguous. It provides a default model which can recognize a wide range of named or numerical entities, which include person, organization, language, event etc. Apart from these default entities, SpaCy also gives us the liberty to add arbitrary classes to the NER model, by training the model to update it with newer trained examples.

D. Sentimental Classification using Text Blob

Sentiment analysis is natural language processing method to identify and quantify the subjective information. It can classify some information into three reaction, they are positive, negative and neutral. Creation of simple sentiment analysis is relatively easy with python. All dependency that is needed is textblob. The output is seen with the two value, they are polarity and subjectivity. The polarity score is a float within the range [-1.0, 1.0]. The subjectivity is a float within the range [0.0, 1.0]. Polarity indicates the sentiment, minus is for negative, 0 is for neutral and positive is for positive statement. Subjectivity indicates that if it’s close to 0 means objective statement, but if it’s close to 1 means the statement is very subjective.

Table 3. Extraction of tweets with NER

S.N	Entity Name	Entity Type
1	Bangladesh	GPE
2	Shahid Afridi	Person
3	Facebook	Organization
4	Samsung	Brand
5	Year	Measure
6	Nationalities	NORP
7	Companies	Organization

V. Results And Discussions

A. Sentiment Compound Scoring

Sentiment Analysis is the process of detecting the contextual polarity of text. In other words, it determines whether a piece of writing is positive, negative or neutral. After clustering the data, we did sentiment analysis on the datasets. For sentiment analysis I used the Naïve Bayes classifier algorithm in TextBlob [7]. This algorithm is used for predicting the probability of words being in any particular class. This is used due to its ease during both training and classifying steps. Preprocessed data is given as input to train the classifier and that model is applied on test to generate positive or negative or neutral sentiment. In Table 4, we can see the Spyder IDE console giving outputs for the tweets that

were used as input (filtered CSV file). The two values under each tweet represents the polarity and subjectivity of the sentence respectively.

Table 4. Calculation of Compound Score

Tweet	score	compound
Thank you Shahid Afridi lala love you	'neg': 0.0, 'neu': 0.511, 'pos': 0.489	0.7717
@Afridi_Shahid__Good Night bhai	'neg': 0.0, 'neu': 0.508, 'pos': 0.492	0.4404
Thank God someone said this….https://t.co/7oqczYSTDv	'neg': 0.0, 'neu': 0.787, 'pos': 0.213	0.5574
Virat won't be… https://t.co/E6qK6zSto3"	'neg': 0.0, 'neu': 0.933, 'pos': 0.067	0.168
@Vj_Cyborg U can't hate Hashim Amla	'neg': 0.0, 'neu': 0.572, 'pos': 0.428	0.4585
@JAfridi10 Hashim amla is coming to pakistan,	'neg': 0.0, 'neu': 1.0, 'pos': 0.0	0.0
Musfiqur Ra… https://t.co/QafyywFlf3",	'neg': 0.0, 'neu': 0.822, 'pos': 0.178	0.5994

The experiment of this algorithm in this thesis is performed in the environment using Spyder IDE. The codes were done in python v3.6.2. I have used TextBlob library for building the Naïve Bayes classifier. We used pythons NLTK (natural language toolkit) for natural language processing basics training dataset for naïve Bayes classifier were manual datasets. We have collected from twitter and later we have used a dataset of tweeter [8] In the classifier we have calculated the polarity. The polarity range is (-1.0 to 1.0) and if the polarity is less than 0 then the sentence is negative. If the polarity of the sentence is 0.0 then the sentence is neutral. Thus, if the polarity greater than 0.0 then the sentence is positive.

i) Positive (>0.0)
ii) Neutral (0.0)
iii) Negative (<0.0)

Fig 2 depicts a pie chart which describes the number of tweets have been found by recognizing the sentiment. The green, orange and blue color portion shows the most found tweets positives, negative and neutral.

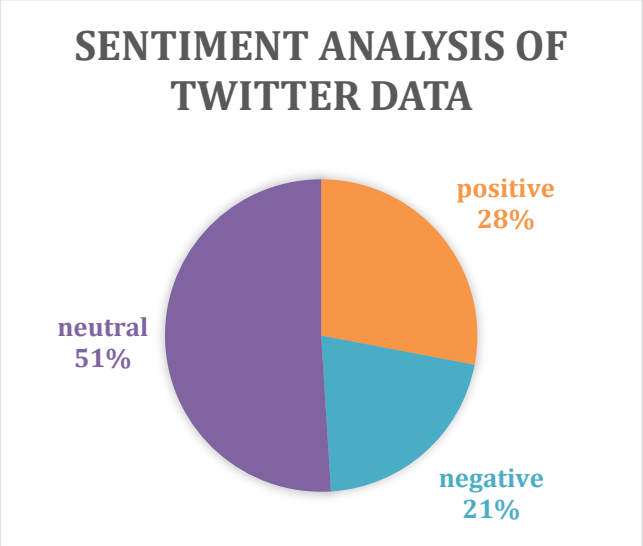

Fig 2. Pie chart for sentimental analysis

B. Rating generation

Rating is generated by averaging the polarity(positive, negative and neutral) off all attributes. Neutral polarities are considered to have a polarity of 0 and it does not affect the average polarity count these values are not taken in order to make the system more efficient. We have saved the number of positive ,negative tweets and averaged them to find out the rating in scale 1 to 10.

Table 5. Player Rating

layer Name	Mean	Rating
Shahid Afridi	0.23404	3
Amla	0.08123	2
Tamim Iqball	0.00824	1
Virat Kohli	0.25325	2.5
Morgan	0.25325	2.27
Shakib Al Hasan	0.37037	2.9

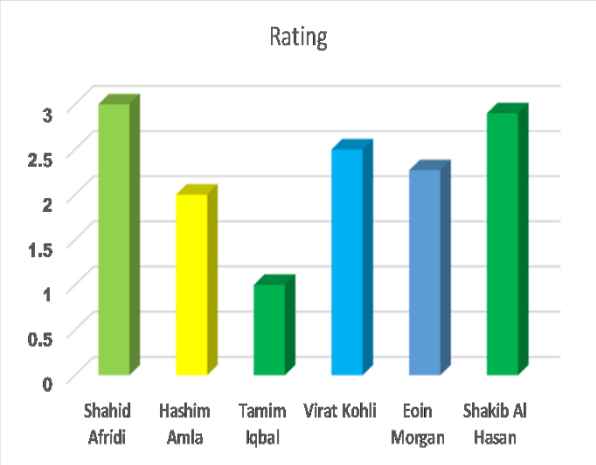

Fig 3. Player Rating

The above figure shows the Ranking of the player based on resent tweets. We have calculate mean of compound score it is a scale has been defined to find the ranking. The figure shows that Shahid Afridi's ranking point is 3 and the Hashim Amla's ranking point is 2. Tamim Iqbal's ranking point is 1, Virat kohli ranks with point 2.5, Eoin morgan's ranking point is 2.27 and Shakib's ranking point is 2.9.

VI. Conclusion And Future Work

In this work, we have proposed a general player rating system based on public opinion. This can be widely used in the future to get proper reviews of any cricketer to get the best review possible for a player. This system is reliable as the rating is basically generated based on public opinion and this rating will reflect the actual state of the player in the ICC. In future we intend to work to integrate the system so that it can generate rating of players of other sports too.

References

[1] R. Karim, "An Effective Machine learning Approach for," *Sentiment Analysis of Restaurant Reviews,* p. 9, 2016.

[2] Q. Li, Examining the accuracy of sentiment analysis by brand monitoring, Netherlands: Enschede, 2015.

[3] B. &. L. L. Pang, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval,* vol. 2, no. 1-2, pp. 1-135, 2008.

[4] Ogneva, How Companies Can Use Sentiment Analysis to Improve Their, Mashable: Mashable, 2012-12-13..

[5] B. B. O'Connor, Linking Text Sentiment to Public Opinion Time Series, mumbai: Center for Applied, 2010.

[6] P. Stone, Sentiment lexicon General Inquirer: A Competetive Approach to content Analysis,, The MIT, 1966.

[7] B. &. Z. L. Liu, " A survey of opinion mining and sentiment analysis. In Mining text data," in *Springer*, Boston, MA., 2012.

[8] mukharju, nataural language processing.

[9] B. Srinivasa-Desikan, "A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd.," *Natural Language Processing and Computational Linguistics,* 2018.