2017 2nd International Conference on Telecommunication and Networks (TEL-NET 2017)

Clustering and Sentiment Analysis on Twitter Data

Shreya Ahuja[1]
Department of CSE
Amity University, Noida
India
shreyaahuja96@gmail.com

Gaurav Dubey[2]
Department of CSE
Amity University, Noida
India
gdubey@amity.edu

***Abstract*- Twitter is a social media platform is a great place where people from all parts of the world can make their opinions heard. Twitter produces around 500 million of tweets daily which amounts to about 8TB of data. The data generated in twitter can be very useful if analyzed as we can extract important information via opinion mining. Opinions about any news or launch of a product or a certain kind of trend can be observed well in twitter data. The main aim of sentiment analysis (or opinion mining) is to discover emotion, opinion, subjectivity and attitude from a natural text. In twitter sentiment analysis, we categorize tweets into positive and negative sentiment.**

Clustering is a protean procedure in which identically resembled objects are grouped together and form a pack or cluster. We conducted a study and found out that the use of clustering can quickly and efficiently distinguish tweets on the basis of their sentiment scores and can find weekly and strongly positive or negative tweets when clustered with results of different dictionaries. This paper surveys different approaches of clustering with respect to sentiment analysis and presents a way to find relationships between the tweets on the basis of polarity and subjectivity.

Keywords- Cluster, Opinions, Sentiments, Twitter.

I.INTRODUCTION

The way people think and express themselves have changed in last few years. Internet has changed their view points, expressions and platform at which they convey these expressions [9]. The platform such as twitter has now become a hub where the people make their opinions heard. Twitter is a great platform for even many businesses to connect with clients and an ideal place for celebrities or any entity to share their thoughts. Millions of users share their thoughts everyday by posting tweets, these tweets can be of no more than 140 letters that makes the users to write short and "to the point" [10]. This "to the point" approach makes it easy to identify sentiments of a given tweet.

The sentiments found in the tweets can be categorized as positive, negative or neutral. Positive sentiments are those, which contains good words or appraisals for a certain statement, news, expression, event, movie or product reviews. Similarly, the negative sentiments are those, which contains bad words or criticize any event, product, movie etc. Neutral sentiments are those which are neither positive or negative [9]. These sentiments along with tweet attributes are usually categorized into classes, but in this paper, I would draw the focus on clustering with sentiment analysis.

Clustering is a job in which we assign certain groups or classes to certain objects such that the objects within the same group or class are more similar than those in the other distinguished group or class [11].

In sentiment analysis, various things are considered as a group of things, example, sentiment scores, polarity, subjectivity, objectivity etc. I use unsupervised learning such as clustering to group such things with one another.

A. Sentiment Analysis

Identifying the mood or opinion of a person's view written in natural language is known as sentiment analysis. The positive or negative polarity is assigned after identification of the opinion [12]. There are many techniques which are applied to a natural text to determine the sentiment such as feature extraction, emoticon study, tokenization etc.

During sentiment analysis, usually positive and negative words are extracted from the text and are assigned a score from the dictionary of words.

The sentiment scores in this research are calculated with the help of two dictionaries, which are used after the preprocessing and tokenization of the tweets. The dictionaries are- AFINN and TextBlob. Sentiment analysis with these two dictionaries/tools are discussed in this paper later.

B. Clustering

Clustering is an unsupervised process which is based on similarity between data to form some pattern which can be useful to fetch appropriate results by training data. In this process, objects that are similar to each other are assigned a specific group or class from those that are dissimilar to them [16].

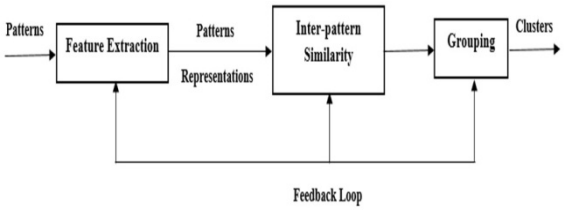

Figure 1: Stages of clustering

In this research, sentiment scores are calculated for each tweet and then unsupervised techniques are applied to them so that groups are formed naturally based on similarity, i.e. clustering. Cluster analysis is an essential unique proof of groups within the same set of objects or patterns [2]. For

instance, if we denote a data within set 'm', which is exactly like other set of classes, we may be distinctly bend towards group 'm' and form those sets of classes into an actual cluster. By increasing the group within similar and other intra groups, authors can intercept and assign the clusters that form an observation within certain space without any much effort. The stages of clustering are shown in fig. 1.

Some commonly used clustering algorithms are:

Partitional Clustering Algorithm-

A partitional clustering algorithm acquires a solitary partition of the data rather than a clustering structure, for example, the dendrogram delivered by a various leveled method. Partitional strategies have points of interest in applications including extensive data sets for which the development of a dendrogram is computationally restrictive [15].

Fuzzy C Means Clustering Algorithm-

In this algorithm, the distance between the cluster center and data point is used to designate membership function to each data point with respect to each cluster center. The higher the proximity of the data set to the center of the cluster, more the membership function would be progressing towards that cluster center [16]. Within each data point the total sum of membership values should be unity. The membership and cluster centers are updated after each iteration.

K Mean Clustering-

It is one of the simplest unsupervised techniques for clustering. In this algorithm, the sample observations are divided into K clusters which are far from each other. Then each data point gets assigned to the cluster that is nearest to it, i.e. with the minimum Euclidean distance. In this way, each data point is assigned one of the chosen clusters [17]. After that the cluster centers are calculated again and the same process goes on until the centers stop changing positions.

II.LITERATURE SURVEY

Bergsma et al. on the basis of twitter data, were able to forecast some invisible features such as ethnicity and gender by performing clustering on visible parameters such as name, location and friend-list[1]. Romero, et al. [2] emphasize the need to be classify data and study of prediction parameters for any data analysis project. Vanessa Friedmann et al. [3] with the help of PCA, pruned the data and altered it into a lower dimensional feature space. This feature can be used to be passed into the k-means algorithm to segment the samples into clusters. Go et al. [4] in 2009 used the emoticons as part of distant learning to obtain emotion or sentiment. The tweets which ended with "☺" were termed as positive tweets and the tweets which ended with "☹" were termed as negative tweets.

In another noteworthy attempt for the classification of twitter data sentiments, Barbosa et al. [5] proposed a polarity prediction, which used the data from 3 websites as unwanted labels so as to train a model and further use about thousand manually labelled tweets for adapting and another thousand manually labeled tweets for checking its reliability. Luiz F. S. Coletta et al. [6] considered the classifiers combination and clusters in the classification of tweets sentiment by using C3E-SL. They considered the auxiliary data provided by data divisions made from "bag-of-words" description with lexicon scores can improve the classification correctness generated by classifiers. In this way, the clusters can form a type of "topic structure" which is there in the encoding of data, which is in the form of meta-information.

A framework for grouping tasks which involved short and sparse text segments was presented byPhan et al. [7]. The main component of their framework is the collection of the appropriate section "universal dataset" from which hidden topics can be determinedand applied to the grouping or classification task, Gimpel et al. [8] introduced a POS tagger for twitter data. They developed features for Twitter POS tagging and conducted experiments to evaluate them, andprovided their annotated corpus and trained POStagger to the research community.

III.METHODOLOGY

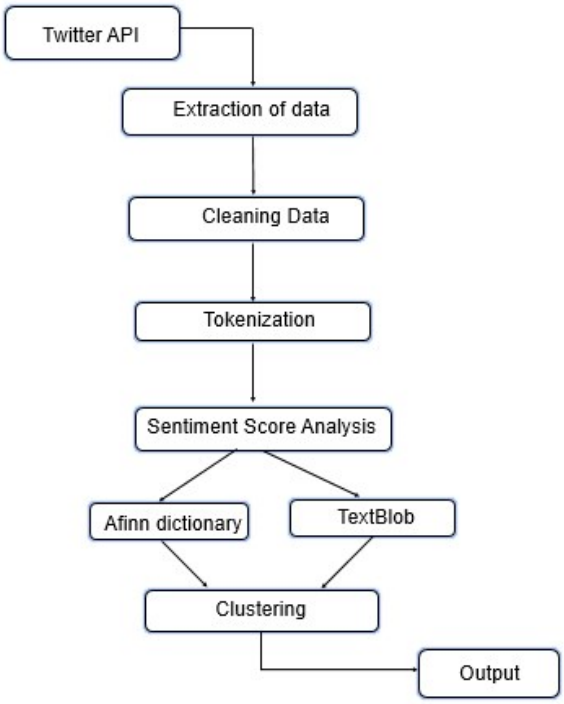

Figure 2: Methodology adopted in this research

A.Data Extraction

Twitter data can be extracted by copying and pasting it to your account but that will be a very tedious task. Hence to solve this issue there is a tool in Python known as Tweepy. To extract data, tweepy is imported and the code is run to get the data from twitter, based on the topic as required, as well as the time and language with the help of OAuth of Twitter API. In this research 1.5 lakh tweets have been extracted of IPL 2017.

B.Cleaning Data

The tweets which are streamed contain a huge amount of garbage background information which is not needed for analysis, or it contains attributes which are not required for the analysis, sometimes the tweets are not in the format which is specified, hence creating errors while analysis [18]. So, we observe the dataset to find such irregularities or inconsistencies and clean the data. The process of data cleaning is straight forward and simple, i.e., if we find inconsistent attributes or data, it is removed. Data which is not in specified format, i.e. JSON format in this research, is removed. Tweets with no text and only attributes are also removed.

C.Tokenization

For any natural language processing problem, tokenization is a must. It is a process which divides a simple natural language data from sentences into individual words or symbols known as tokens [18]. Tokenization is also known as lexical analysis; the phrases or sentences are broken down into lexicons so that the computer is able to understand our language. For example, a sentence "This is a test that is not so simple: 1.23." after tokenization will look like this, "This" "is" "a" "test" "that" "is" "not" "so" "simple" ":" "1.23" ".".

In this research, every tweet is tokenized in python using libraries and functions such as tokenizer and a natural language processing tool-NLTK which makes tokenization possible.

D.Sentiment Score Analysis

Sentiment refers to a personal point of view, feeling, emotion or expression towards something. A person's sentiments can be broadly classified into 3 categories, namely, positive, negative and neutral. Each sentiment is given a polarity, -ve polarity for negative sentiment and +ve polarity for positive sentiments (neutral sentiments being 0). These sentiment polarities are accompanied with individual token's score [12].

The analytical work starts by processing each token of each tweet. For analysis of natural language, unnecessary words like prepositions and articles that are called stop words are removed as they are not useful for sentiment analysis. Another important technique while doing sentiment analysis is to consider consecutive words occurring together, called bigrams. For example- I do not like this car, here "like" tells us it's a positive sentiment but as "not" comes before "like" it is a negative sentiment, hence if consecutive words had not been taken into account, the analysis would have given wrong results.

In this research, tweets of IPL 2017 are taken and are the sentiment scores are calculated using two tools/ dictionaries- Afinn and TextBlob.

AFINN Dictionary- Afinn is a lexicon which was introduced by Finn Årup Nielsen. It works on python and contains a dictionary with 2477 words and their respective sentiment scores [19]. Afinn is used to assign polarity scores to the data, where a –ve score indicates negative sentiment and positive score indicates positive sentiment. Higher the magnitude of the score, higher is the degree of positivity or negativity. There is no specified range within which Afinn scores are limited.

TextBlob- It is a library for natural language processing which work solely on python. TextBlob is used to assign both polarity and subjectivity scores to the data. "The sentiment property returns a named tuple of the form Sentiment (polarity, 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] where 0.0 is very objective and 1.0 is very subjective."

E.Cluster Analysis

K-means, Fuzzy C-means, Hierarchical clustering and Mixture of Gaussians are commonly used clustering techniques. For this research, as the data is very large and Euclidean distance needs to be computed in less time to make the analysis more efficient, K-means algorithm technique is used [17]. The chief objective is to assign k number of centroids which is for each cluster.

1. Choose K number of clusters randomly, far from each other.
2. Calculate the distance between each data point and the centroid, i.e. cluster center.
3. The data point is assigned to the cluster with whose centroid it has the minimum distance.
4. Once all data points have been assigned to the clusters, the centroids are recalculated using data from results obtained in previous step with the help of the formula:

$$C_i = (1/n) * \sum_{j=1}^{n} x_i^{(j)}$$

5. Repeat steps 2-4 until no new k centroids are formed.

To choose the appropriate value of k for k means clustering, elbow method and silhouette methods are used which are graphical methods [20]. Silhouette method requires a lot of computation and it saves the data after each step hence using a lot of memory of the system on the other hand elbow method uses less memory space. On the data obtained from manipulation elbow method and silhouette methods are applied. An example of elbow method is shown in fig.3, the point where the elbow bent is present in the graph, that is the appropriate value of K that should be chosen.

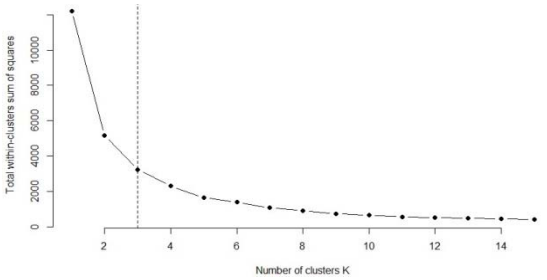

Figure 3: Graph showing elbow method to find 'K'.

Based on the type of dataset, k means uses 3 algorithms, namely Hartigan - Wong algorithm, Lloyd algorithm and MacQueen algorithm.

In this paper clustering has been applied to 2 problems, Afinn VS TextBlob and subjectivity VS polarity. For the former, Lloyd algorithm is used as the data is very large for computation and distinctly scattered on the plot. Lloyd algorithm optimizes the total sum of squares. For the latter, Hartigan - Wong algorithm is used as within sum of squares approach was appropriate to get fast initial convergence.

IV. Analysis and Results

On the collected data, sentiment scores are calculated using Afinn and TextBlob. The following table 4.1 shows the tweet id of the tweet (which is a unique id for every tweet) with their respective TextBlob polarity score, TextBlob subjectivity, Afinn polarity and the language of the tweet and time at which it was created.

On using this data obtained from sentiment score values of subjectivity from TextBlob and polarity from TextBlob and Afinn, from table 1, two graphs are constructed showing the plot between Afinn polarity VS TextBlob polarity and subjectivity vs polarity (Textblob) (ref. fig 4 and fig 5)

TABLE 1. RESULTS AFTER THE SENTIMENT ANALYSIS

Tweet ID	TextBlob Polarity	TextBlob Subjectivity	AFINN Score
"851890416630734850"	0.65	0.84375	14
"851890416567721984"	0	0	0
"851890417972924416"	1	1	-1
"851890442782232576"	-0.4	0.6	-5
"851890446338899968"	0	0	0
"851890453557288960"	0	1	0
"851890461497315331"	0	0	0
"851890479407013888"	0.4	0.25	5
"851890498641985536"	0.25	0.45	0

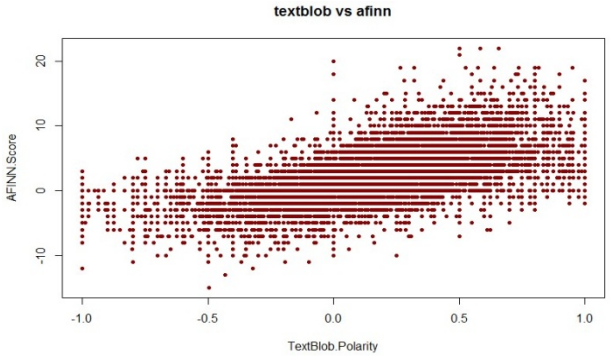

Figure 4: Plot between the sentiment scores of Afinn and Textblob applied on about 1.5 Lakh IPL 2017 tweets.

In fig.6 we can clearly see 4 centroids and 4 clusters. The cluster in cyan represents the tweets that have a very high degree of positivity in both the tools and hence are certainly positive. The second cluster in blue is most compact and shows tweets that are slightly positive and nearing to being neutral, but we cannot be certain about the polarity of this cluster as a whole. The cluster in green is again not certainly negative or neutral, it has tweets that are mainly negative and nearing neutrality. The cluster in red is most widespread and has high negative polarity.

We can also see that cyan (cluster 1) covers the maximum area and hence we can say that maximum tweets were positive.

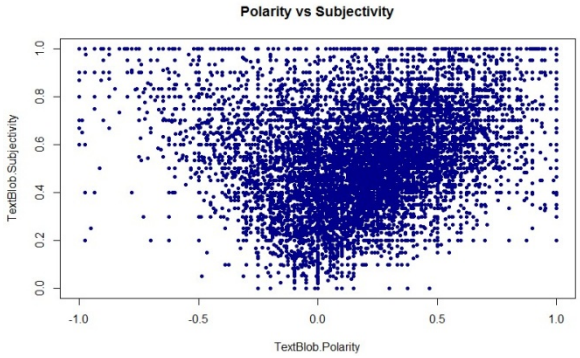

Figure 5: Plot between the subjectivity and polarity scores obtained from Textblob applied on about 1.5 Lakh IPL 2017 tweets.

Elbow method is applied on both polarity vs subjectivity and Afinn vs Textblob data to determine the appropriate value of K respectively.

The value of K came out to be 3 for subjectivity vs polarity and 4 for Afinn VS TextBlob. Using this value of 'k' we applied Lloyd Algorithm for the graph plot of Afinn versus TextBlob in figure 6 and HartinganWong algorithm for the graph plot of subjectivity vs polarity in fig.7. Both these algorithms were run in R to obtain the resulting visual graphs below.

Subjectivity is based on opinions, i.e. subjectivity scores in texttblob are based on keywords like 'I', 'my', 'our', 'mine' etc. whereas polarity is based on sentiments whether they are positive or negative. Hence by performing clustering on polarity and subjectivity, it can be inferred as to how many opinions are positive or negative and the degree of the same.

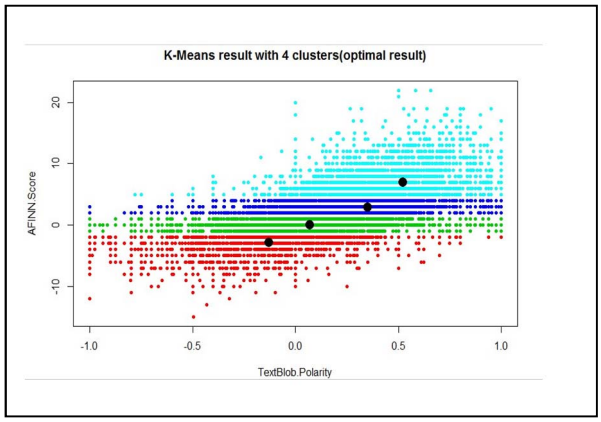

Figure 6: Resulting clusters for the plot b/w Afinn and Textblob scores using K-means clustering.

Cluster ‘blue’- It includes the tweets with subjectivity greater than 0.4 and going up to 1.0 and sentiment scores from -1.0 to +0.3, i.e. from extremely negative to very slightly positive. Hence this cluster covers highly opinioned, negative tweets and is the most widespread.

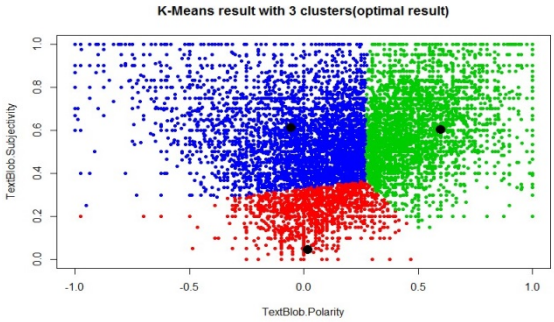

Figure 7: Resulting clusters for the plot b/w Polarity and Subjectivity scores using K-means clustering.

Cluster ‘green’- It includes the tweets with subjectivity greater than 0.2 and going up to 1.0 and sentiment scores from +0.4 to +1.0, i.e. this cluster as a whole contains only positive tweets with varying subjectivity.

V. Conclusion and Future Work

. From this analysis, we came to know about different clusters that our sentiment scores belong to both polarity wise and subjectivity wise. Pre-defined dictionaries or sentiment tools can’t contain the proper score of every word in the context to a sentence, hence forming a cluster of the results from both the tools’ score, we are able to group ‘definitely’ positive and ‘definitely’ negative tweets. Some tweets just contain news or some other person’s view on a certain expression but after finding the subjectivity (personal opinion, feeling or emotion of a particular person expressing it) of a tweet we can cluster it with sentiment scores and find the ratio of the real positive and negative ‘opinions’ and not just sentiments of a sentence.

As the data, I chose was specific to IPL 2017, I would like to make labels specific to the topic and then find sentiment scores related to this topic itself by
using a classifier

REFERENCES

[1] S. Bergsma, M. Dredze, “Broadly improving user classification vi communication-based name and location clustering on twitter,” 2013.

[2] C. Romero, S. Ventura, P. de Bra, and C. Castro, “Discovering prediction rules in aha! courses,” Proceedings of the International Conference User Modelling, 25-34.

[3] Vanessa Friedmann, “Clustering a customer base using twitter data,” CS-229, 2015.

[4] Alec Go, Richa Bhayani, and Lei Huang, “Twitter sentiment classification using distant supervision,” Technical report, Stanford, 2009.

[5] Luciano Barbosa and Junlan Feng, “Robust sentiment detection on twitter from biased and noisy data,” Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36-44, 2010.

[6] Luiz F. S. Coletta, N´adia F. F. da Silva, Eduardo R. Hruschka, and Estevam Rand Hruschka Jr, “Combining classification and clustering for tweet sentiment analysis,” Brazilian Conference on Intelligent Systems. 2014.

[7] X. Phan, L. Nguyen, and S. Horiguchi, “Learning to classify short and sparse text & web with hidden topics from largescale data collections,” In Proceeding of the 17th International conference on World Wide Web (WWW '08, ACM, New York, NY, USA, 91-100, 2008.

[8] Kevin Gimpel, Nathan Schneider, Brendan O’Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith, “Part-of-Speech tagging for twitter: annotation, features and experiments,” CMU, Pittsburg, USA.

[9] Manisha Rani and Jyoti Arora, “A review of data analysis of twitter,” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 5, May 2016 ISSN: 2277 128X.

[10] Alexander Pak, Patrick Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining”.

[11] P.-N.Tan, “Introduction to data mining,” Pearson, 2006.

[12] Svetlana Kiritchenko, Xiaodan Zhu and Saif M. Mohammad “Sentiment analysis of short informal texts,” Journal of Artificial Intelligence Research 50 (2014) 723–762.

[13] DoaaMohey, El-Din, and Mohamed Hussein, “A survey on sentiment analysis challenges,” Journal of King Saud University - Engineering Sciences, 2016.

[14] Mythili S andMadhiya E, “An Analysis on clustering algorithms in data mining,” International Journal of Computer Science and Mobile Computing, Vol.3 Issue.1, January- 2014, pg. 334-340.

[15] Guha, Meyerson, A. Mishra, N. Motwani, and O. C., “Clustering data streams: theory and practice,” IEEE Transactions on Knowledge and Data Engineering, vol. 15 pp. 515-528, 2003.

[16] J. C. Dunn, “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters,” Journal of Cybernetics 1973, 3: 32-57.

[17] J. B. MacQueen, “Some methods for classification and analysis of multivariate Observations,” Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281297.

[18] RishabhSoni and K. James Mathai, “Effective sentiment analysis of a launched product using clustering and decision trees,” International Journal of Innovative Research in Computer and Communication Engineering, 2016, ISSN: 2320-9798.

[19] Lars Kai Hansen, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, and Michael Etter, “Good friends, bad news - affect and virality in twitter,” The 2011 International Workshop on Social Computing, Network, and Services, 2011.

[20] Nicolas Bertagnolli, “Elbow method and finding the right number of clusters,” http://www.nbertagnolli.com /jekyll/update/Elbow.html, Dec 10, 2015.