International Journal of Engineering &Technology, 7 (3.12) (2018) 1168-1174

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Sentiment Analysis using Machine Learning through Twitter Streaming API

P. Akilandeswari[1*], R. Harshita[2], Sumanth.KO.M[3]

[1,2,3]Department of Computer Science and Engineering, Srm Institute of Science and Technology, India
**Corresponding Author Email: [1]akilandeswari.p@ktr.srmuniv.ac.in , [2]harshitaravichandran@gmail.com ; [3]sumikom927@gmail.com*

Abstract

Social media allows to share the experiences with many best suggestions and provides opportunities to share the ideas about any topics at any time. In the current trending, twitter is used to gather different kinds of information as user need and it is a social network service which enables the user for better communication and gaining of knowledge. Accurate representation of the user interactions can be done based on the facts sematic content. The pre-processed tweets which are stored in database are been identified and classified whether it relates to the user keywords related posts. The best suggestion using polarity can be predicted using the user keywords. For the interactive automatic system which predicts the tweets posted by the user this system deals with the challenges that appears during the sentimental analysis. It deals with effective study prior to the subjective information. The basic task in this is to identify the polarity of a given tweet in the sentence whether it is positive, negative or neutral. However the polarity of the tweets has been identified, it was difficult for us to check with the meaningless data. To address this challenge the extracted tweets are been pre-processed by replacing the full form instead of short term words. The better performance can be achieved using more training data. However the analysis was frequently done using the previously stored data, it was a challenging task to do it using the streaming data. There are very few works related to the sentiment analysis using online streaming data. In this paper, we propose that the sentiment analysis can be improved using the online streaming data. For online streaming data all the data related to the given topic will be collected according to the current data in the twitter. For better up-to-date analysis, the streaming data is used and can achieve better results. In contrast by conducting the continuous learning from the streaming data, this approach provides better results than the traditional way of using the training data and it achieves the overall performance and computational efficiency.

Keywords: *Continuous learning, Opinion mining, Sentiment analysis, Social media analysis, Streaming data.*

1. Introduction

Data mining is the process of determining and learning the new patterns in a large dataset which involves the merging of the machine learning, statistics and database systems. The target of the data mining is to collect the information of the dataset and convert it to an understandable structure. It deals with information processing which involves the collection, extraction, warehousing, analysis and statistics including the artificial intelligence and in machine learning. The data is extracted in an interesting patterns such as the group of a data records like cluster analysis and sequential pattern mining. By the result of a decision support system, the data mining can identify the multiple group of data which obtains more accurate prediction results. As a result of the improvement in the dataset which has grown in size and complexity, the hands-on data analysis can be increasingly augmented with direct automated data processing such as the cluster analysis and support vector machines. The larger datasets are been used for the more efficiency in the learning. The dataset learning process consists of: Clustering- It is a process of grouping the data and also grouping its structures in the data which are similar in any way. This work subjects to the clustering using the hashtag provided to the related topics which are grouped together as a related cluster. Classification-It is a process of grouping the known structures which applies to an additional data.

This work classifies the subject related to the topic into positive, negative, mixed and neutral. Summarisation- This work summarises the related terms that are collected based on the given topic using a hashtag. Summarisation is done for collecting all the information by applying the input. The input is entered as a hashtag followed by the word. Then it summarises all the related tweets related with the input. Data cleaning is based on the removal of the unwanted data or any interpolating missing values. It also deals with finding the hidden correlation in the data and identifies the source of data which are more accurate and determines the most appropriate value for the analysis. Exploring data includes the calculation of the maximum and minimum value and to calculate the mean and standard deviations. SQL server Data Quality services are used to analyse the distribution of the data such as wrong or any missing data. The parameters are used in the algorithms and use the training data to use a subset of data which creates different results. For the verification of the model to check whether it specifies to your data, statistical techniques like cross-validation is used to create automatic subsets of the data and set the model for each subset. This work avoids the usage of the meaningless data such as the short term words, repeated usage of the letters and the words which are not related to the topic. Twitter data – In general twitter data gives the insights into a type of information stored in an account. It constitutes a rich sources which are used for capturing the information about any topic.

This data's can be used in many cases such as finding trends related to a particular keyword or by measuring the brand sentiment or by gathering the feedbacks. In this work it limits the usage of the extracted data for 50 users according to the current timing of the discussion about the topic by a user. Subjectivity classification is also used in this work as a single sentence may contain multiple opinion and subjective clauses. In this work few sentences does not contain a sentimental speech and few sentences may contain a sentimental speech. In such use cases the below statements gives an example based on the subjectivity using sentiment and subjectivity without using the sentiment. SUBJECTIVITY SENTENCE WITHOUT USING SENTIMENT- Example: "I believe that she went home yesterday". SUBJECTIVITY SENTENCE WITH USING SENTIMENT- Example: "I am so happy for you". Classifying a subjective sentence provides some conclusions, the pure subjectivity sentences have a clear idea towards the text whether it relates to either a positive or a negative sentiment.

Dataset Constructs And Sentiment Labelling: The training datasets are used as the application of the domain knowledge for training the lifetime learning model. When a new task comes, that first checks for the interrelation between the new and every previous tasks. If the two tasks are identical then we combine the knowledge of both the tasks into a single task and we acknowledge it as a new task. It stores the knowledge collected from all the previous tasks. Finally it compares the knowledge from the previous tasks and predict on the future tasks for a better learning technique. It recollects the knowledge form the previous learning and it helps for better future learning.

Challenges In Sentiment Analysis: Sentiment analysis is useful nowadays in many ways. The method in which delivering the information from person to person plays a important role in customer buying decisions. In commercial ways like sharing the opinions or attitude about any business or product or in any other social issues. Applications such as movie reviews, business, political leaders, government officials, stock market, consumer market and social issues. The challenges in the sentiment analysis are:

Indirect Sentiment- Certain sentence may have an indirect sentiment even without any presence of the sentiment words. Example: "How can you afford to buy this product?", "One should question the writer's point of view who wrote this novel". Here both the sentences do not express any negative words but both the sentences contains negativity and it denotes a negative sentence. Hence identifying the semantics is very essential in the sentiment analysis.

Impede Expectations - Certain sentence builds up the sentence at the first and refuse it at the end of the statements. Example: "The product is amazing, it has a very good specification and there are many colors available and has very good usages, However it can't hold up". Inspite of the word which has a positive orientation, the total sentiment is negative since the second statement is crucial and it plays an important role. The term frequency of the sentence is more important than the term presence.

Practical Sentiment - It is very important to understand the practicality of the sentence because, sometimes it changes the whole meaning of the sentence. It changes the sentiment of the sentence completely. Example: "I just finished watching the football match. The finals completely destroyed me". In this case of practical sentiment, the first statement describes about the watching of football match and the negative orientation is discussed in the second statement. There are many ways to represent the practical of the sentiment. As per the practical knowledge the overall orientation inferred here is negative.

Negation - In sentiment analysis, handling the negation task is challenging and is very difficult. The negation terms can be expressed even without the usage of the negative word. In the sentence "I do not like the product", there is a negation operator (like not). But in the sentence "I do not like the acting but I like the location of the movie", there is a combination of "not" and "only". The combination of "not" and "only" deter-mines the orientation of the sentiment. Such negated words should be handled carefully.

Role of Expressions: Role of expressions is a type in which the word is assigned using its symbolic functions. It plays a crucial role in assigning each word with its functions.

Adjectives- It is a very important parts of speech used and is used very frequently. There is a relation between the adjectives and the subjectivity. In general, people most commonly use the adjectives to express their sentiment towards the topic. Expressing the thoughts using the adjectives has high accurate results and the user can express their own point of view about the topic in an expressive manner. Example for wordlist of positive adjectives: "Brilliant, Excellent, Awesome, Fantastic, Cool, Exciting". Example for wordlist of negative adjectives: "Bad, Slow, Terrible, Stupid". These examples indicates the list of words containing the positive and the negative adjectives.

Adjective Adverb Mixtures- In sentiment analysis the adjective-adverb mixture performs the crucial role in analysing the sentiment in the sentence. To find the overall polarity of the tweets, it is necessary to evaluvate all the sentence of the tweets. Each sentence plays an important role in finding the sentiment (polarity). In general the users posting the tweets may use the adjective/adverb in a sentence to represent the thoughts about any topic. So it is necessary to calculate even the adjectives and adverbs for finding the sentiment. There are many types in adverbs representing the sentiment. There are declaration adverbs where the words used are affirmative and proclaimed words. In adverbs with doubts, there are certain words which represents as a doubt statement and does not represent a declarative sentence. There are also strong adverbs, weak adverbs and negation adverbs. For each type of adverbs there is each sentiment and the sentiment for each type is also calculated. Few types are: Adverbs with declaration: Certainly, Completely. Adverbs with doubts: Maybe, Probably. Strong adverbs: Extremely, Distinctly. Weak adverbs: Slightly, Hardly. Negation usage: Never, Certainly not. In the adjective-adverb mixture the sentiment is been calculated as, if a word in a sentence denotes the adverb and other word in the same sentence denotes an adjective, the sentiment score value of the sentence is altered by the adverb adjoining with it. And if the sentence containing more than one adverb and adjective, the score is calculated by altering the score of the adjective as each adverbs gets added to that. To calculate the score of the adjective-adverb mixture, a score value is allocated based on the adjective and adverb. There are certain cases like the sentence representing the weak adverb and adverb with doubts has the score less than or equal to the strong adverbs and adverbs with declaration.

Example for strong adverb and adverb with declaration: "Extremely good is more positive and has high score when compared with the adjective good". The word extremely good has more positive score when compared with good.

Example for weak adverb: "Hardly good has more negative score when compared with the adjective good which has a positive score". In this, hardly good indicates a negative orientation and good indicated a positive orientation.

Example for adverb with doubt: "Perhaps good has less score point when compared with the word extremely good". The first word "perhaps good" indicates a word containing a doubt and the second word "extremely good".

2. Related Works

Mart'inez-C'amara et al, [1] describes about the semantic orientation of opinion words in tweets, it presents a hybrid approach using both the corpus and dictionary based methods. These methods deals with the usage to find the semantic

orientation of objectives and the semantic orientation of verbs and adverbs. The major disadvantage of the paper is that there is a maximum length of 140 characters in a passage and in our project there is no such limitations on the characters. Alec Go et al, [2] explains the main contribution involved is to use the tweets with emoticons for distant supervised learning. It states the effective way of performing the distant supervised learning which can be done using emoticons as noisy labels for training data. The major disadvantage in this paper is that the neutral classification of words is not been considered for the efficient way of finding the polarity. In our project neutral class of words is not been ignored and it is been taken as a classification. Martinez-Camara et al, [3] proposed the constrained system which follows the supervised approach and the unconstrained system which follows the unsupervised approach. Due to the restriction of usage of the words, only 140 characters are been used and the users generally represents the concept in an irregular form. In our work, there is no limitation of characters and the polarity can be done easily without any data sparsity and longer texts are also taken for the checking of the polarity. Lei Zhang et al, [4] deals with the entity level sentiment analysis method in twitter. Here the new entity level sentiment analysis method is been proposed, to perform a entity level analysis this method adopts a lexicon based approach. For better accuracy results, the learning based method which used maximum entropy as the supervised learning algorithm, which performs in a poor manner. Jiang et al, [5] proposed a novel content representation for target dependent in sentiment analysis. It is independent of syntactic analyzers and it incorporates the sentiment lexicon information and distributed word representations. This paper generates two new context matrices LS and RS by keeping the embeddings of words in a sentiment lexicon but it does not checks the polarities of each word. In our work the polarity of each word in a sentence is been checked and calculated to achieve better results. P'erez-Rosas et al, [6] explains the opinion extraction tasks which is performed on subjectivity and analysis in languages other than English. Our work deals with the subjectivity classification label texts as to be either subjective or objective. The sentiment classification adds a granularity by further classifying the subjective texts into a positive, negative or neutral. Trivedi et al, [7] deals with the document level analysis can be benefited from fine grained subjectivity so that the sentiment polarity judgements are been taken based on the relevant parts of the document. It allows to learn the discourse connectors annotations without subjectivity annotations. This improves the level of sentiment analysis in English and in Spanish and it fails to represent in the richer representation of discourse structure and the sentence level valence. In our project we use the heuristic for automatic identification connectors when there is no list available and includes richer representation of sentence level valence and subjectivity. Polanyi et al, [8] explained mainly on the negative or positive attitude discussed by the individual terms which is incomplete and may give wrong results when implemented directly. It fails to explain that the context insensitivity evaluation. In our work we use this work for more sensitive evaluation of texts which gives an expected results. Wilson et al, [9] proposes the phrase level sentiment analysis which first determines whether an expression is neutral or polar and then removes the ambiguity of the polar expressions. The contextual polarity for a large subset of expressions is been automatically identified and thus achieves the results that are significantly better when compared to the baseline. In our work it automatically classifies the contextual polarity for a long sentence having large subset of expressions and achieves high quality and high precision. Tang et al, [10] explains the avoiding of the problem such as the expensive data labelling efforts, this paper deals with the categorization of reviewing the progress on transfer learning and related machine learning technique such as classification, regression and clustering problems. It also deals with the domain adaptation, multitask learning and sample selection bias and the covariate shift. The disadvantage is that it fails to solve how to avoid negative transfer which results on avoiding the transferring process itself. In our work we avoid those negative transfer by focusing on the distant measures and can classify the related cluster domains or a task. Severyn et al, [11] proposes a new model for initializing the parameter weights of the convolutional network which is difficult to train the accurate model while avoiding the need to add any additional features. It does not learn the rank for the microblogs retrieval and answer the re-ranking for question answering. In our paper, we initialize the process which includes the noisy labels which are inferred using emoticons found in the tweets. Cambria et al, [12] explains the off-topic passages which contains the unwanted affective information and it tends to mislead the results for the global sentiment polarity about the overall topic. The weakness of the knowledge based approach is that it poorly recognizes the affect when the linguistic rules are to be involved. Our work uses hybrid approach which aims better knowledge of the conceptual rules that mainly involves sentiments and the clues that can convey
the concepts from realization to verbalization from the human mind. Chen et al, [13] discussed about the current dominant machine learning paradigm works in an isolation. No attempt is been made to retain the learned knowledge and using that in future learning. The disadvantage is that it cannot backtrack and fix the errors which causes wrong inferences which was made based on the errors. In our paper we deal with the correctness and the applicability of the knowledge. It also deals with the learning with tasks of multiple types taken from different domains. Tang et al, [14] deals with the big data assisted customer analysis and advertising architecture. It aims to seek the potential users and it improves the efficiency of the advertisement delivery. The disadvantage of the paper is that it utilize the telecom data in a city and checks for the arrival rate, but the superior advertising exposure rate were low. In our paper, we deal with the precise advertising scheme such as data modularized application, analysis and mining and the collection and storage of the data's. Gimpel et al, [15] explains the previous tasks on twitter sentiment analysis using distant supervision. The existing requires the huge computation resource for analyzing the large number of tweets. Here it proposes a technique to speed the computation process for the sentiment analysis. The tweet subjectivity is been used to select the right training examples. The concept of effective word score is also been used which derives from a polarity score of frequently used words. It creates a prior probabilities using the datasets for average sentiment tweets in different spatial, temporal and authorial contents. The disadvantage is that it takes high computational time with high accuracies. In our work we develop the lower computational time with very high accuracies when compared to the baseline model.

3. Methodology

The methodology includes the Twitter Extraction, Classification of the tweets, Sentiment Analysis, Polarity Prediction.

Module Description:

Twitter Extraction - It facilitates interrelation among the system and the client. The client has privileges to create account and access his/her feeds from the system.

Classification of the tweets - Using Naïve baye's algorithm, the data can be analyzed and can recognize the patterns used for the classification and regression analysis. After pre-processing the extracted data, it is then classified into keyword related tweets. It classifies and predicts the group and clusters according to the user group. The same clusters are grouped under a classification. All the positive tweets are clustered into a positive classification, all the negative tweets are clustered into a negative classification, tweets containing a mixture of both the positive and the negative

are clustered into a mixed classification and all the neutral tweets are clustered into a neutral classification. The classification of the tweets makes easier to find the score point for each classification. And hence the polarity of each classification can be represented.

Sentiment Analysis - The sentiment analysis is used comparatively to categorise the positive, negative, mixed and neutral comments related with the text categorization. Sentiment analysis has the complexity like conveying the assumptions in different ways. In opinion texts, the lexical content might get misleaded. Intra-textual and sub-sentential reversals and negation topics can be commonly interpreted. The below are the possibilities that need to be classified as Users, Texts, Sentences (paragraphs, chunks of text), Predetermined descriptive phrases(<ADJN>,<N N>,<ADV ADJ>, etc), Words, Tweets/updates. Sentiment-oriented data sets are scope sensitive and it is challenging to create/collect the data from large domain. Representations of the sentiment needs more attention on elements to classify and scale the domain-appropriate annotated data is available or not. This work deals with the analysis of the tweets and it checks for the behaviour of the tweets posted by the user.

Polarity Prediction - The analysis of the tweets is done and then it is classified based on polarity of the words as positive, negative, mixed and neutral etc. The number of positive, negative, mixed and neutral tweets are identified based on the polarity. Naïve Bayes algorithm finds the polarity of the tweets by classifying it based on the positive, negative, mixed and neutral tweets. For correct predictive accuracy, set of training data is necessary. Bayes algorithm finds the polarity of the tweets by classifying it based on the positive, negative, mixed and neutral tweets For correct predictive accuracy, set of training data is necessary. Example: Analysing of the hotel reviews, government official feedback or any social reviews. The task is to predict the overall polarity based on the users comment about a topic. Example: If the specific topic is been represented using a hashtag, it extracts the tweets which are related to that topic and after the pre-processing technique the overall polarity is been determined. For finding the polarity, first the score point for each classification should be calculated. According to the threshold value assigned for each word, similarly all the words belonging to each classification are been calculated with its threshold value and the overall score point is represented as a percentage. And thus the score point are calculated. After finding the score point, the polarity should be calculated and the overall polarity is represented in the form of a graph. The score point for each classification is identified and the polarity for it is also expressed in the form of a bar graph. For representing the polarity in the bar graph, canvas js is used. Canvas javascript is used to visualise the data. It creates a rich dashboards that works on drives and does not compromise the functionality of the web application. It is a Hypertext Markup Language5 (HTML5) .

Architecture:

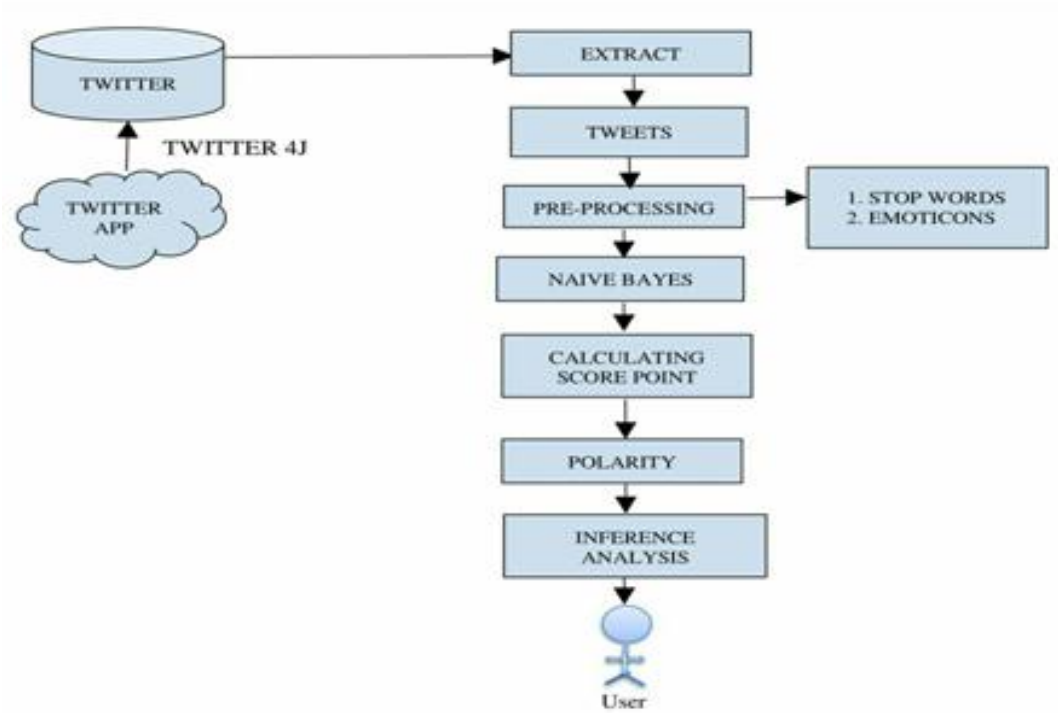

Fig. 1: Architecture

Pre-processing of the tweets- The pre-processing stages includes the removal of the stopwords and the emoticons.

Removal of stopwords: The pre-processing is been done to refine the extracted tweets, and refining process makes easier to do the sentiment analysis and to find the overall sentiment. Removal of stopwords is the first step of pre-processing stage. The removal of stopwords includes the removal of punctuations, usage of comma, usage of full stops, usage of the repeated letters, the double space is been replaced with the single space and it also involves the removal in re-occurrence of the full stops. It also involves in the removal of words like (a, about, after, across, and, is, for) etc, since it does not help with finding the polarity. These type of words does not deal with finding the polarity and hence those words are been removed in the pre-processing stage.

Full stops: In some tweets the user uses the casual languages which is often in the form of a repeated words or punctuations. For example: "This is so nice... wow", in such cases two or more occurrence of full stops can be removed.

Parse hashtag: The previous work deals with using a hashtag followed by the single word, it just removes the hashtag and adds the word to a feature vector. In some cases there is a usage of multiple words followed by a hashtag like "#BestEventEver". Hashtags like "happy", "depressed" are frequently used for the purpose of the sentiment classification.

Repeated letters: In general few texts may contain a casual usage of a language. For example- the word wow in an arbitrary numbers of "O" in the middle (“woooow”, “woow”) this results in

a non- empty result set. This process occurs if there is a reoccurrence of word more than two times. The negative words like cannot, wont, don't are replaced by not which effectively maintains the sentiment table.

Removal of Emoticons: The removal of emoticons involves the removal of special symbols like @ which will be replaced by AT. Special symbols like :) and :(is been replaced with HAPPY and SAD respectively. All the special symbols are been replaced in the form of words. In the training dataset around 250 such commonly used symbols is been trained. After the pre-processing stage, the tweets are refined and now it can be used for finding its polarity for each classification (positive, negative, mixed and neutral).

Working Process- Initially the data is been collected from the twitter service using twitter API. The twitter API is a database which collects the data from the twitter. The streaming data is been extracted from the twitter service. The tweets can be extracted by accessing the permission from the Twitter Application Management (TAM) twitter application management. For accessing the tweets, the permission should be sent to the twitter application management to generate the keys, where the keys are generated for our particular twitter account. After generating the keys, it can be used for extracting the streaming tweets from the twitter. The keys generated are consumer key, consumer secret key, access token key and access token secret key. The online streaming data is used to collect the up-to-date information from the twitter data. In this work 50 users streaming tweets can be collected and sentiment analysis can be done. The input is represented as a hashtag followed by the word, and the 50 users streaming tweets can be collected according to the topic. The user latest tweets are collected and it depends upon the time in which the users discuss. If the input which we enter does not have the latest related tweets, then there is an occurrence of exceeding the rate limit of the keys which are generated from the twitter service. Sometimes it tries to collect the maximum tweets if available. After the extraction of the tweets the pre-processing of the tweets is been done. The pre-processing is been done to refine the extracted tweets, and refining process makes easier to do the sentiment analysis and to find the overall sentiment. The pre-processing stages includes the removal of stopwords and emoticons. After the pre-processing of the tweets for identifying the clusters of words belonging for each classification, the splitting of words is been done. A single tweet is represented as a sentence and each and every word in a sentence gets splitted separately and it counts every word and evaluvates which words belongs to which classification. The training dataset contains a set of trained data and it also has a set of trained emotions which performs the crucial role in sentiment analysis. After splitting the words, each word gets compared with the training dataset and compares its knowledge with it and if both merges it gets clustered as a classification. After the pre-processing stage, the next step is to classify the tweets. In the sentiment analysis the classifications used are positive, negative, mixed and neutral. Using naive bayes algorithm the classification of the tweets is been done. All the positive related words are clustered as a positive classification. All the negative related words are clustered as a negative classification, the tweets having both the positive and the negative related words are clustered as a mixed classification and all the neutral related words are clustered as a neutral classification. The calculation of the score point is been analysed to find the overall reviews of which the user has been discussed related to the topic, given in the hashtag. The score point for each classification (positive, negative, mixed and neutral) is been calculated and identified. In finding the score point for each classification, it first calculates the points assigned for each word under each classification. Classification contains a clusters of similar data (words) and each word is assigned with a threshold value ranging from -4 to +4. The threshold value for all data in every classification is calculated and the score point for each classification (positive, negative, mixed and neutral) is been analysed. Finally the overall polarity of the tweets are classified as positive, negative, mixed and neutral, it expresses the overall review about the given topic in the form of a bar graph and the user who wishes to see the review of any topic can analysis the data using the resultant graph and it gives the user an efficient result about that topic.

4. Results

This section explains the experimental results of the sentiment analysis using online streaming data, and expresses the performance of the sentiment analysis in the form of a bar graph based on the various current trending topics. This section explains the experimental results of the sentiment analysis using online streaming data, and expresses the performance of the sentiment analysis in the form of a bar graph based on the various current trending topics.

Experimental results for the sentiment analysis discussed on the online streaming topic related to the current product review #iphone10.

Overall Polarity	Calculation
Mixed	5.436573 %
Polarity	**Calculation**
Posivite Polarity	5.107084 %
Nagative Polarity	2.306425 %
Mixed Polarity	5.436573 %
Neutral Polarity	1.8121911 %

Fig. 2: The overall score point discussed about the current online streaming product topic #iphone10

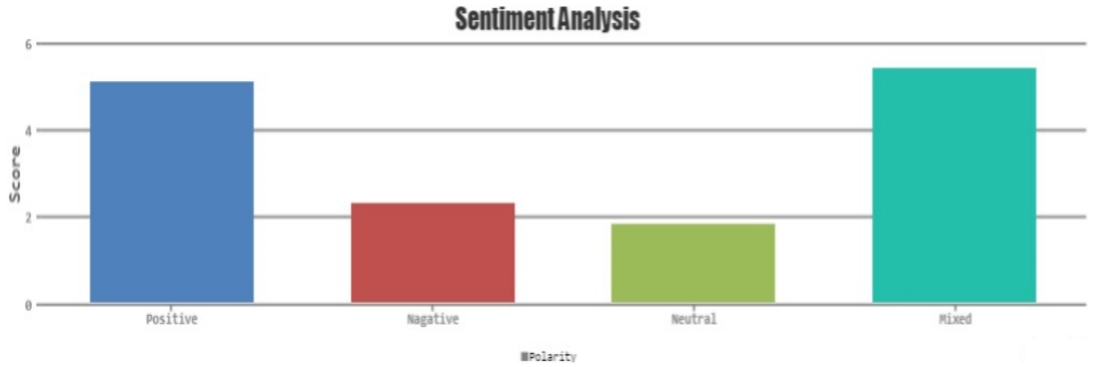

Fig. 3: The relation between the score point and the polarity of the tweets discussed about the current online streaming product topic #iphone10

5. Discussion

Figure2 represents the score point value about the current trending topic #iphone10. The results are been taken based on the tweets discussed by various users related to a product review #iphone10. This table proves that the overall tweets discussed about the topic contains more mixed polarity than the positive, negative and neutral. Based on the calculations, it clearly infers that the mixed polarity contains 5.436573%, positive polarity contains 5.107084%, negative polarity contains 2.306425 % and the neutral polarity contains 1.8121911%. The more mixed (combination of both positive and negative) comments related to the topic is been discussed by the users than the positive and negative comments and only certain comments is been discussed based on the neutral.
Figure3 represents the relation between the score point value and the polarity of the tweets discussed about the topic. This graph indicates that the overall tweets contains more mixed (combination of positive and negative) comments and the various users have discussed more about the mixed feelings related to the topic. The positive and the negative comments is also been discussed about the topic. And the graph also represents that there are less neutral comments discussed by the users about the topic in the tweets. Hence these results prove that the online streaming data is been taken for better up-to-date analysis and also achieves better results for various streaming topics. Similar sentiment analysis can be done by taking various other trending topics. Similarly with these kind of datasets, the streaming twitter data is been analyzed for both the trending and non-trending topics. The trending topic named, ban Sterlite is also been analyzed and had the results of 5.1344743% of positive polarity, 4.8899755% of mixed polarity, 2.9339855% of negative polarity and 2.9339855% of neutral polarity and the other topic related to the aadhar card issue is also been analyzed and found that it has 5.0% of negative polarity, 4.1304345 of mixed polarity, 1.9565217 of neutral polarity and 1.7391304% of positive polarity.

6. Implication

From the implications of the results, it is proved that for better up-to-date analysis the streaming data can be used. It gives better results when compared to static data. The result proves that the sentiment analysis is been done and can predict the overall polarity of the tweets, the user who wishes to see the review can infer from the analysis of the data using the resultant graph and it gives the user an efficient result about that topic. In this paper, the streaming data is used for the analyzing the tweets and finding its polarity. This paper uses streaming data for social media analysis. For extracting the streaming data, Twitter 4j is used. Twitter 4j: Twitter 4j is an unauthorized java library for the twitter API and has android platform and google app engine. Using this api, integrating the java application can be done using the twitter support. Using this, the streaming data can be easily collected from the Twitter Database (TDB). Extract tweets: Here the tweets are been extracted from the twitter.
Extracting the tweets are necessary to do the sentimental analysis. The tweets extracted is based on the users posting the tweets according to the topic. According to the input given, the tweets are collected from the twitter service. Streaming data related to the topic are collected. Extracting the trending tweets based on the latest news in the current social media any amount of tweet can be extracted and can be modified into any form. Our work deals with extracting the streaming data according to the input given. The input by the user can be based on a trending topic or also based on a regular topic. The tweets which are given by the user in the social media using hashtags etc. in the twitter. The input is been represented as a hashtag followed by the topic. Our project has the ability to collect the tweets of 50 users for all topic.

7. Conclusion, Limitations and Future Works

This work proposes a system for categorizing the results based on the polarity analysis of the data. The data which are extracted from the twitter database will be a streaming data. The prediction gives more accurate results and quick response using the social network based behavioral analysis. The limitation of this work is that the languages other than English are not considered for the sentiment analysis. This work can be further extended by improvising the sentiment analysis for various other social media like Instagram, Face-book. In future, the sentiment analysis can also be done in the languages other than English. In this work estimating the polarity and the calculation of the sentiment cannot be done if it is in the language other than English. In our work, the sentiment analysis can be done only in English and the future work can be extended by using other languages also.

References

[1] Martinez-Camara, E., Martin-Valdivia, M.T., Urena Lopez, & Montejo-Raez, A. (2014). Sentiment analysis in Twitter, Natural Language Engineering. 2014, vol. 20, pp. 1-28, 1.
[2] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. Stanford University, Stanford, USA, Tech. Rep. CS224N, 2009.
[3] Martinez-Camara, E., Montejo-Raez, A., Martin-Valdivia, M.T., & Urena-Lopez, A. (2013). SINAI: Machine Learning and Emotion of the Crowd for Sentiment Analysis in Microblogs. Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) pp. 402-407.
[4] Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2011). Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis. Tech. Rep. HPL-2011-89, 06 2011.
[5] Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter sentiment classification, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. vol. 1. Association for Computational Linguistics, 2011, pp. 151-160.

[6] Perez-Rosas, V., Banea, C., & Mihalcea, R. (2012). Learning Sentiment Lexicons in Spanish. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).
[7] Trivedi, R.S., & Eisenstein, J. (2013). Discourse Connectors for Latent Subjectivity in Sentiment Analysis. HLT-NAACL, 2013, pp. 808-813.
[8] Polanyi, L. & Zaenen, A. (2004). Contextual Lexical Valence Shifters. Proceedings of the (AAAI) Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004.
[9] Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis, in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. USA: Association for Computational Linguistics, 2005, pp. 347-191.
[10] Jia, L., Yu, C., & Meng, W. (2009). The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness. Proceedings of the 18th ACM Conference on Information and Knowledge Management, ser. CIKM 09. New York, NY, USA: ACM, 2009, pp. 1827-1830.
[11] Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. Proc. 38th Int. ACMSIGIR Conf. Res. Develop. Inf. Retrieval, 2015, pp. 959-962.
[12] Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Comput. Intell. Mag., vol. 9, no. 2, pp. 48-57, May 2014.
[13] Cambria, E. (2016). Affective computing and sentiment analysis. IEEE Intell. Syst., vol. 31, no. 2, pp. 102-107, Mar./Apr. 2016.
[14] Chen, Z., Liu, & B., Abdullah, R. (2016). Developing conceptual governance model for collaborative knowledge management system in public sector organisation. Journal of Information & Communication Technology (JICT), 15, 171-191.
[15] Gimpel, K. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. J Proc. 49th Annu. Meet. Assoc. Comput. Linguistics: Short Papers, 2011, pp. 42-47.