An Integrated Approach to Sentiment Analysis using Machine Learning Algorithms

Abstract

Sentiment Analysis (SA) is nothing but mining the emotion from many sources. Some of them include texts, audio, video, etc. Every individual has their own opinion, hence own reviews and ratings. Based on these reviews if we classify the sentiment of the opinion of the public, the profit or loss calculations on the product or application is directly found. Our algorithm takes Naïve Bayes (NB) as a foundation of classification of textual data taken from the public and categorizes the tweets accordingly. To this, we are adding an organic emotion factor called Average Impact Factor (AMF). In the market, there were several algorithms which can be used in mining the sentiment from the given textual data. But this sentiment has flaws as it cannot detect the true emotion from the text or it overfits the opinion of the public. Based on this idea, we integrated the AMF on the public tweets and reviews to evaluate the true sentiment and to improve the time factor too of the opinion mining. We used data of tweets related to Demonetization that happened in India, 2016. When compared to NB Classifier and Support Vector Machine (SVM) algorithms, there is an improvement in time constraint and accuracy too in our Integrated Sentimental Analysis (ISA) classifier.

Keywords

Average Impact Factor, Naive Bayes Classifier, Tweets, Sentiment, True sentiment, Sentiment analysis, Twitter data, Support Vector Machine.

1. Introduction

Sentiment Analysis is the computational studying and identifying people's attitudes, emotions, and opinions towards the individuals or events. The individual or entity is also called as entity [1]. It is the classification of text strategy that classifies text. This process is based on the individual opinions which has some sentiment of the given context. Consumers have the trust on the opinion of other consumers which are written as comments in various social media platforms. Specifically, on those comments which have greater experience or which show a keen insight of a product or service, rather than various marketing strategies of the company or any other sources. Social Media is the main platform which influences consumer preferences or opinions towards a service or a product by molding their thinking abilities, attitudes and behaviors. Social networking has a greater influence on Internet. It also influences the purchasing behavior or opinions on any product of the people, is growing over the years. Retailers, understood that the usage of Social Media can extend their scope towards the improvement of their strategies which can increase their sales gradually. Social network or chat box can provide a virtual interaction which can improve the company's brand and also help them to gather extremely useful and data about people opinions or demand trends [2]. SA plays a prominent role in a branch of Artificial Intelligence (AI) i.e., Natural Language Processing (NLP). With the help of SA in NLP, we could get the sentiment of the text as we get the summary of the paragraphs using NLP [3]. Using this we can extract information about sentiment from product reviews and comments in social networking sites and various platforms where there is a huge interaction between different individuals used by both producers and consumers to know the trend of the sales and demand of the product, thus it helps the people to take necessary action on the product sales [4]. As this happens very often, it plays a prominent role in the improvement of sales of the product and even to classify the category of the products. SA involves insight or a clear view of an opinion of the individual from the tweets or texts in product reviews or movie reviews to analyze those opinions [5].

The structure of the data in reviews are almost in the text format which is in the form of unstructured in nature. For example, movie reviews corpus in NLP will be in text and unstructured format for which we have to

do preprocessing to make it understandable to the machine as well as humans. Thus, the stop words like are, a, an, etc. & also the other unwanted information like symbols, special characters are removed from the raw text for further analysis. The data which has undergone the preprocessing stage will undergo through an another process called vectorization, where the text data is converted into a matrix of numbers. During this process these matrices are then given as the input for a selected optimal Machine Learning (ML) algorithms for the classification of the reviews from the given data. The classification accuracy depends on the parameters used. There will be various parameters used to evaluate the performance or the accuracy of the ML algorithms [6]. Social networking sites will be used as a medium or a means in which the users may post their opinions of a product or service or movie review and this data will be collected from those blogs or sites, and used for classification [3]. The tremendous increase of usage of the social media or the social networking sites like Facebook, Twitter, Instagram has raised the possibility of using web. There are used to explore the insights and track the opinions of the individuals from their reviews or comments about any particular product or topic in the different social interaction sites [7]. SA is chosen as a best topic for research work by many researchers due to its usage in the application marketing strategies for better result and also due to the changing needs of the people [3]. In this paper, we have used twitter dataset to apply the SA. It can be used to solve many problems of broad range that are of attentive to human and computer interaction. This is the practitioners and researchers, and also the people in the fields such as sociology, marketing, and advertising, psychology, economics, and political science are focused on SA. There are some serious problems for the practical applications which are occurred due to the nature of content in Twitter or Facebook - Social Networking Sites or Social Media [8]. The main task of SA is the classification of the sentiments for the given input text, where the text in the document will be classified as a positive, negative & neutral polarities of a target object [9]. It is observed that this process is performed in three different approaches or levels such as document level, sentence level, and aspect level. In document-level approach, the classification process is undergone for the whole document at once and document is classified into different polarities. In sentence-level approach, the classification process is undergone for each statement or sentence for the given data or document. In aspect-level approach, the expressions of sentiments which is present within the given document and the area to which it refers [6]. In our study, sentence-level approach is taken into consideration as we perform the analysis of the sentiment on Twitter data.

The ML techniques are mainly of two types, which are very often used in SA. Those techniques are supervised and unsupervised learning which are been classified on the bases of the nature of the input data. The supervised learning is a technique which applies on supervised data – labelled data, which is processed to get the best output from the given input with a high accuracy which helps for the process of proper decision making. Unlike supervised learning technique, unsupervised learning technique is the process that uses unsupervised data – that does not need any label data; which will be not an ease procedure to process the data. To solve the problem occurred with processing of unlabeled data which produces lesser accuracy or doesn't give a clear output values, clustering algorithms or techniques are used like K-Means, Nearest neighbor, etc. [6]. In our paper, we have presented the usage of a supervised learning method on labeled data. In majority of the cases, the use of statistical distribution like normal distribution, poison distribution or ML algorithms such as NB, SVM has proven to be successful. The method or technique which is chosen, used to calculate the sentiment polarity of the given sentence and then classify it into a positive or a negative or a neutral class. Thereafter, SA can be used for the conclusion of the factors or aspects expressed in the opinions. With the help of processing with SA on reviews, ratings, recommendations, online opinion and other forms of online expression has turned into an important evidence for businesses looking to market their products, increase the sales of the products, which will help them to find out the new opportunities and increase their fame or reputations. As the business companies are looking forward to automate the process of filtering out the noise, understanding the conversations, identifying the relevant content and action it appropriately, many researchers are now looking at the field of SA for necessary improvements. It is the most basic level attempt to derive the emotion or 'feeling' for the body of text by using the ML techniques like SVM, NB, etc. The SVM can be described as the set of classifiers which are used for SA with several univariate and multivariate methods for feature selection [10]. NB algorithm can be described as, it is simple probabilistic classification technique, in which Bayes' theorem is applied with strong independent assumptions [5].

We have observed an aspect from the previous methodologies, that the previous models used to classify the data by calculating the polarities from the text using ML models, but there are some cases like some people post their tweets in a sarcastic way and those tweets will be overlooked by the ML model. In this paper, we explore the tweets in twitter dataset and find the polarities of each tweet to classify into different polarities. After finding

polarities, we use our integrated approach to find the celebrity factor and calculate the weight of the polarities. Celebrity factor is that tweets which contain the '@mention' are processed to find the polarities and their weights. We add these both weights from which our integrated model results out the genuine sentiment. We have compared our results with the outcomes of the previous methodologies where they give only 80% accuracy and our model gives 82% accuracy, from which we can conclude that our integrated approach gives a very good result. Remaining sections of the paper covers Related Work in Section 2, Section 3 consists of Problem Statement and Proposed Methodology, Section 4 will contain Result Discussion, Section 5 has Future work, and Section 6 is the Conclusions.

2. Related Work

Tan et al. [11] carried out an experiment, in which an attempt for adaptation of domain in the topic SA was proposed. There are three main contributions. They are: First, using the maximum old data of the domain, they have proposed a best technique or method, i.e., Frequently Co-occurring Entropy (FCE), which processes the picking out of the common features, it is also to use those figures as a link like a bridge from an old domain to a new domain. As long as the iteration occurs, they have used only the common features for the old domain data. Secondly, to get the knowledge from both of the domains i.e., old and new domains, they have proposed Adapted Naïve Bayes classifier. The exact idea of their proposed methodology was to deploy or use a weighted Expectation Maximization (EM) algorithm which can combine both old domain and the new domain data, after that it gradually increase the weights for the data in the new domain with decreasing the weights for the data of new domain in each iteration, with the hope to fit the new-domain data as well to the proposed model. Thirdly, they have conducted the many experiments on adaption of tasks for the six domains which has improved the classifier's performance dramatically, it also provides a best performance than semi-supervised and transfer learning techniques. It was concluded that their proposed methodology has improved the classification accuracy for the classification of text for the domain adaption.

In the Mubarok et al. [4], the classification using NB was done on two different variables. They are, one was aspects and the other was sentiments on some aspects or features (food, service, price, ambiance, and miscellaneous). To present the system in a generative model, they have told both the above mentioned variables influences the use of words in sentences. They mentioned as the probabilistic distribution of words will depend on the value of each variable. Thus, their process made a conclusive statements based on the sentiment polarities of certain aspects or features. This process or method involves in counting the polarities – positive, negative, neutral and conflict, each aspects or feature percentage of the data is calculated. To facilitate the users, to view results of aspect based SA on the all reviews they have made a summarized form of data for all the restaurants domain in the form of rating. The rating of each aspect was calculated based on the number of each polarity on the corresponding aspect or feature. In their paper, they concluded that NB classifier has showed the best result for aspect based SA which has the F1-Measure of 78.12%. They have mentioned various measure of their Model. They are: That the aspect classification best F1-Measure was 88.13%, and for sentiment classification the best F1-Measure was 75%. They have also mentioned that the Part of Speech (POS) tagging method and also about the Chi-Square (CS) method are also can be used for feature selection which further was used for the process of classification in NB classifier. The CS also has been proved by them which was speed up its computation time in the process of classification of NB although it degraded the performance of the system in their work.

The Comparison study performed by Devika et al. [3] describes that SA has many approaches to classify the sentiment. The ML strategies or algorithms will work when the algorithm is trained with training dataset and it produces the output when it is tested with testing dataset. It is known that an ML technique or algorithm, first get trained with some part given data with the outputs which are known, so that in future it can work with new data with unknown outputs. Some of the strategies are SVM, NB, Maximum Entropy (ME) Classifier, K-Nearest Neighbor (KNN) and more. It was mentioned on their paper that, the SVM method gives the advantages of having input space with high-dimensional, few irrelevant features & document vectors can be sparse, but the disadvantages are a large amount of data is required. NB algorithm advantages are it is very simple method and both the accuracy and the performance or accuracy is combined and disadvantages are can be used when the size of training data is less. By considering the merits and demerits of each strategy, they have used an integrated approach using NB to get high accuracy results.

They have mentioned the following are the ML algorithms for SA are as follows:

It was mentioned that SVM requires a large amount of training set as it is a non-probabilistic classifier and will be performed or processed by classifying points which uses a (d-1) dimensional hyperplane. It will choose a hyperplane having the biggest margin or area and it defines decision boundaries which will use the concept called decision planes where the separation between a set of objects having different class membership is done by using the decision plane.

N-gram is a SA concept and it was mentioned that it is used in both linguistics and probability fields, where the contiguous sequence of n items is n-grams, which the sequence was taken from a given text or speech sequences. To the application the items which can be phonemes or syllables or letters or words or base pairs. Text or the speech corpus are the data places where the n-grams can be collected. Shingles are the items or words which can be also referred as n-grams.

In that paper it was mentioned about the Maximum Entropy (ME) classifier which is also called as Conditional Exponential classifier, where they are parameterized using a set of weights. They are used to add up the joint features which are generated from an encoded set of features. The maps which are encoded for each feature pair will set labels to a vector. ME classifiers are the exponential or log-linear classifiers, it is because they will work using this sum as an exponent where the sum is obtained by set of features which are extracted from the input and combining them linearly. It was also mentioned that if this method is performed in an unsupervised manner, then the Point Wise Mutual Information (PMI) was used in order to get the co-occurrence of a word with both the positive and negative words. Independent features are not assumed in ME classifier. The uncertainty of a uniform distribution was maximum. Entropy is the measure of uncertainty.

Multilingual Sentiment Analysis is also a topic mentioned in their paper where it is described as, different language should be used to express the individual's views which helps the researchers to get the date for their research on it. Natural Language Tool Kits are used to perform or process it. In this method, language models first define the language. The sentiment classification is performed using the given language by changing it to English and process it to get the output.

Multimodal Sentiment Analysis framework which was in a paper by Poria et al. [12], in which the sets of required elements for text and visual data are included in it. From the different modalities a simple method for combining them was also explained in it. Sentic-computing-based features are used to enrich the textual SA module in their method or model. These features are the reasons to improve the performance of the model.

The experiment performed by Tripathy et al. [6] is done using different supervised learning ML models which is for the classification of movie reviews dataset. The IMDb movie dataset is used for the purpose of their proposed model where the ML algorithms are further processed using n-grams model. It was observed in their paper that as the accuracy of the classification decreases the 'n' in n-gram increases. They have concluded that Term Frequency - Inverse Document Frequency and CountVectorizer are combined together to produce a model with high accuracy with the usage of ML algorithms.

The SA performed by Mertiya and Singh [5] using NB and Adjacent analysis. Their model will work on the set of tweets dataset which are related to movie reviews. First, they preprocessed the data (tweets) and then they performed SA using NB and determined polarities of the tweets, then they performed Adjective analysis which helped to determine the intensity of the tweet. From the determined results they have concluded that their model is 88.5% accurate. Temporal SA, using phrases of sentiment their method or technique produces two types of graphs. They are: a topic graph and sentiment graph which are the expressions of sentiment patterns such as "happy" or "delighted at".

Martin et al. [12] have performed a SA of political communication which uses a dictionary approach along with crowd coding. The analysis of large text corpora was enabled by the dictionary which will be as a resource for the intensive hand-coding struggles. From the dictionary of words, the tonality of sentences is calculated and they are validated with results from manual coding. Finally, an efficient and valid measurement of sentiment will be provided by the results of their method shown which have shown that the crowd based dictionary.

Hanhoon et al. [13] have performed SA of restaurant reviews using Senti - lexicon and improved NB algorithm. In their paper, the improved NB algorithm was proposed, in their thesis the proposed model effectiveness along with the improved accuracy and balanced way of classification were proved.

Malcolm et al. [14] have published a paper on the topic called Investor Sentiment in the Stock Market. Investor sentiment plays a crucial role in stock market where generally the stocks are fluctuated using the sentiment of the investors. They have proposed a top-down approach to analyze investor sentiment on stock market. They have proved that the measure of investor sentiment will impact the sentiment which will be a clearly discernible, important, and regular effects on individual companies or on stock market as a whole.

The Survey on SA was done by Mohamed [15]. In it he has done a comparison between the review structure of sentiment and challenges in SA, based on a relationship. He also specified that the theoretical type in sentiment challenges.

The paper written by Tomohiro et al. [16] tells the understanding of the sentiment by people from news articles. Their paper specifies about the method that takes input as a texts along with respective timestamps such as news articles or web-blogs and produces two kinds of graphs. The two graphs are described as: The topical trends which are associated with a specific sentiment are called as the topic graph, and the trends which is made up of the sentiments which are associated with a topic is called as sentiment graph.

Bessi and Ferrara [17] have investigated about the role and the effects of social bots or automatic accounts and written in their paper. They have written that the social bots or automatic accounts are mainly used for the manipulation of conversations in online. Especially, they have written that those bots or automatic applications are pervasively present and was active in the year 2016 during United States presidential election, in the online political discussion.

A paper which was published by Anshul et al. [18] on topic of stock prediction by using twitter sentiment analysis. Using the dataset collected from Twitter feeds and DJIA values over a period of June 2009 to December 2009, they have proposed a method which is a new cross-validation technique for financial data and using Self Organizing Fuzzy Neural Networks (SOFNN) the proposed model has obtained an accuracy of 75.56%. Based on the predicted or obtained values from their proposed model, they have also implemented a naive portfolio management strategy.

Johan et al. [19] have published a paper on which the prediction of the stock market was described by Twitter mood. In that paper they have specified about their investigation on public mood. On the basis of correlated on even predictive values of DJIA the large-scale collection of tweets is measured.

In the paper published by Xian Fan et al. [20] it was mentioned that on application of word vectors for analysis of sentiment of APP, the specification of reviews was done as per their investigation on it which specifies the effectiveness of the representations as vectors over different kinds of text data and the quality of domain-dependent vector was evaluated. They have concluded that the huge amount of meaningful researches will be done using the sentiment which are produced from the aspects.

3. Problem Statement and Proposed Methodology

A. Problem Statement

Our approach developed three problem statements, which after solving makes the future approaches in SA task bit easier. They are as follows:

(i) When someone approached the twitter data and starts the classification, he runs into an enormous amount of data processing. Our model proves that data from the optimum number of

tweets with some specific constraints are enough to find overall polarity, hence reduced time complexity and increased efficiency.

(ii) An improved model for classification can be build which focuses on the time constraint.

(iii) An algorithm for further classification of tweets and to find true sentiment out of the tweets need to be constructed so that more accurate calculation of sentiment from the given tweets.

B. Proposed Methodology and *Building model for finding the proportionality ratio*

Before knowing what, we should know why. The reason to develop this approach to develop an algorithm which proves there is a relationship in terms of proportionality ratio between retweet count of the tweets of a user (celebrity for example) and sentiment map which gives overall sentiment on the topic that has been retweeted. Our approach is purely to make things very simple, i.e., to reduce the time consumption of the data processing. When we make glance in the previous algorithms defined [1,2,3,4], the time factor is not considered for their analysis. But in our approach, we have taken time factor into consideration. We have trimmed the data taken or extracted from the twitter and trimmed dramatically more than 50% of overall data collected. But we left with no loss of accuracy in prediction of sentiment taken out of the data. Figure 1 gives the overall idea of data trimming with the help of the retweet count.

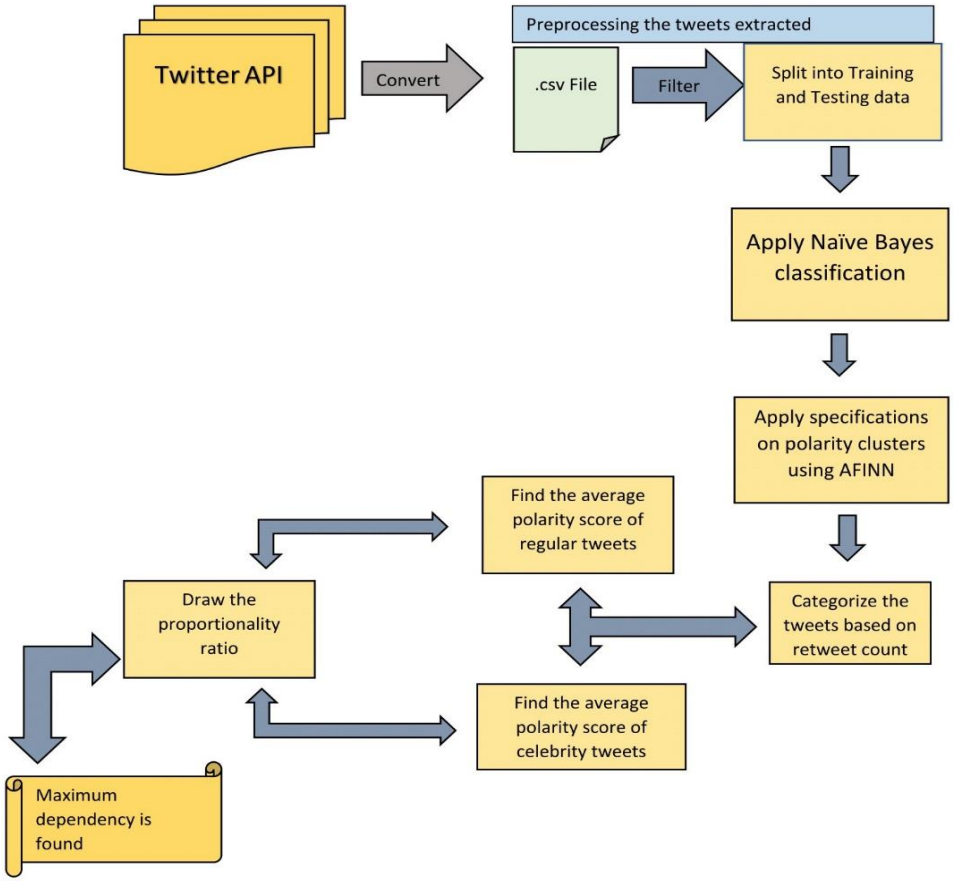

Figure 1: Model for Proportionality Ratio

Each step mentioned above in the Figure-1 is clearly detailed below with every possible step that is noticed in our work is depicted well in the below elaboration of steps.

C. Data Extraction and preprocessing:

This world is full of data. Everyone needs data for many uses, say for example commercial uses. Equal to the requirements, there is a much greater number of sources available in the market to provide the data. One of the finest sources used by many people (even the researchers from reputed universities) is Twitter. The twitter provides the facility to create a developer account so that we can extract tweets with required hashtags. This method is secured because every developer has unique tokens of access and secret keys to request the extraction. We are interested to work on the tweets related to Demonetization that happened way back in 2016 in India. Thus, we applied #demonetisation, #demonetization, #Demonetisation, #DEMONETISATION in our python algorithm of regular expression and extracted the tweets related to our requirement. We used file handling technique to store all the tweets which include the specifications like user name, created date & time, tweet text, retweet count, etc.

The file containing the tweets related to Demonetization is extracted with the help of twitter API is made into required format (CSV, JSON, etc.…). We were interested to work with CSV files, hence formatted into the former one. After the extraction, tweets are checked for features. Whether every feature is useful for training the model or not. There are some features which are felt is not required for the model we proposed. The features like data & time of creation of tweet, Screen name, favorite count, reply to SNN, Status Source are selected for filtering. Hence the features like these are irrelevant and are filtered. We also removed the data whose value is marked as NAN. In the abstract, you can say we have performed normalization of the data, by filling the missing or NAN marked values. Along with this redundant data is also eliminated. The NB algorithm is exclusively for the features which are assumed to be independent of each other. Hence to make this rationale, we have considered only the features which are wisely independent of each other.

Apart from these steps, there is one more thing left to be done. That is formatting the tweets in the file. When we speak about the format, it means every text should have a similar structure defined in a proper way. Tough we use only the numbers in the first proposed model, there is a need for crunching text of tweets too for extraction of sentiment from them. Format all the tweets so that no capitalization, punctuation, or non-ASCII characters are present, as well as splitting the tweet into an array holding each word in a separate holder is done.

D. Apply the Naive Bayes Classification:

One of the finest probabilistic technique used in algorithmic domain is Naïve Bayes algorithm, which is the flavor or updated version of Bayes theorem. These two are very similar at core but differs by the term Naïve (which introduces the concept of conditional independency). The meaning of conditional independence means that whenever a data object with certain number of features is taken, every feature is independent to each other in terms of cardinality.

The given objects are classified into various number of categories, with the help of respective features. This process of classification/ categorization becomes difficult when one need to handle enormous amount of data, provided dependent features of data object. The principle of NB classifier is nothing but, one feature of the data object is not dependent of another. This assumption is called class conditional independence. The need of this assumption arose to reduce the computation complexity when enormous amount of data is given. The assumption is strong and also sometimes not applicable, where anomaly existed in data given. But studies found thumping need of NB classifier when it comes to formation of Artificial Neural Network (ANN). In addition to this, high accuracy along with efficiency is also a factor luring to adopt this technique.

Using Bayes theorem, we can write,

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\,p(F1,\ldots,Fn|C)}{p(F1,\ldots,Fn)} \text{ ----------(1)}$$

Here, class C can be positive or negative.

The theorem explained above is a simple mathematical model of NB classifier. If we dive into deep, NB is a family of classifiers. Every classifier is based on the popular Bayes probability theorem. Tough we assumed that the features selected to be the dataset are independent, NB can still perform well under dependent features too. This is a wall break for the unrealistic assumption of independence. There are wide ranges of applications where NB classifiers are used. Some of their applications are on diagnosis of diseases and decision making process about the treatment. The taxonomic studies which consists of the RNA sequences classification, and filtering of spam messages in e-mail clients.

E. Specification of the classified tweets

Before going into actual classification, we have some steps to follow to create a better classification algorithm. The first most sub-task of the pattern classification is feature extraction and selection. There are three mal criteria of good features are listed below:

- Salient: features are important and meaningful
- Invariant: this is related to image classification
- Discriminatory: the selected features are to be capable to distinguish between patterns

Prior to fit the model using ML algorithms we need to have an idea about this: how can we generate feature vector from the best representation of the tweets from the text document. There is a commonly used model called a *bag of words model.* The theme of this model is simple. In the early stages, we create a vocabulary – the collection of all different words (which are mostly used in the tweets). The vocabulary is then used to construct the d-dimensional feature vectors for the individual texts where the dimensionality is equal to the number of different words in the vocabulary. This is termed as *Vectorization.* We combine aspects such as opinion strength, emotion and polarity indicators, generated by existing SA methods and resources and found the polarity scores of the tweets. Hence, obtained the opinion of the twitter users.

	critical	**question**	**was**	**about**	**sudden**	**decision**
$\mathbf{X}_{T1}$	1	1	1	0	0	0
$\mathbf{X}_{T2}$	0	0	1	1	0	1
$\sum$	1	1	2	1	0	1

Table-1: *Bag of words representation of two sample texts T1 and T2*

F. Categorizing according to retweet count:

All the tweets are sorted according to Algorithm-1 in the order of the retweet count so that we can find the influence of the popular users of Twitter. All the tweets are filtered so that only the tweets which have more than 1000 retweets are taken into consideration, remaining are left out for further steps.

Algorithm-1: tweetClassifier(tweet)

1. columnFinder(tweet) //finding the retweetCount column data to be checked
2. Extract retweetCount
3. If retweetCount > 1000, insert into popularSlice
4. Otherwise into regularSlice
5. Return popularSlice and regularSlice

G. Calculating the averages of tweets and finding proportionality ratio:

Soon after categorizing the tweets based on the number of retweets or retweet count, we calculated the average of polarity scores on two halves.

The first half contains the tweets of 17 (whose retweet count is more than 1000), and the second half contains tweets of 5144 users (whose retweet count is less than 1000).

- Polarity sum of tweets with retweetCount > 1000 = -2
- Polarity sum of tweets with retweetCount < 1000 = -565

The average polarity of first half = $\frac{-2}{17}$

The average polarity of second-half = $\frac{-565}{5144}$

The averages of both halves are divided with their respective number of users. The results obtained are just amazing. The first half has the value of overall polarity as -0.112 and second has -0.109. Of course, the overall polarity of the Twitter data we have taken is negative. But it is proven that tweets of the users with more retweets are influencing the overall sentiment of the twitter data. The above is illustrated in Figure-2.

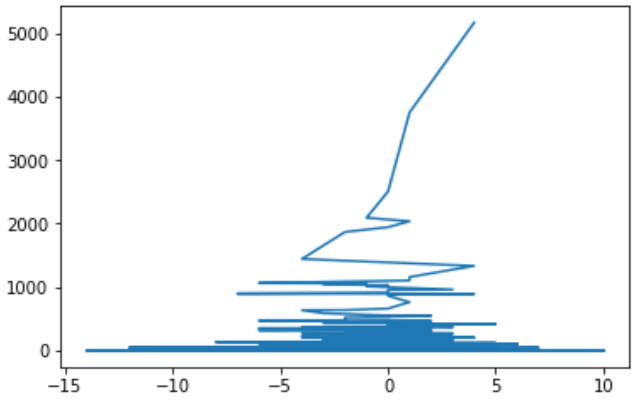

Figure 2: Polarity Score v/s Retweet Count

Above graph made the scenario hidden behind very clear. The scenario embedded is this. The users with more popularity i.e., the user who normally got the more than the appreciable number of tweets count with retweets, he/she has more impact on the factor that is been nominated or being retweeted by that user. This influence can be either positive or negative, depends on the type or kind of reputation that user has in the community of social media and other as well, which includes the field in which he is working. This can be seen in this way too. Whenever the user with more followers made some tweets on the topic, it will become trending in very less time. Then it is also possible that the followers can be fans or anti-fans or critics. Whatever the response is being generated in the form of retweets is simply because of the one who tweeted on the topic first (the user with more followers). Hence the source of opinion generator in social media is the users with a greater number of followers. Hence, they are the game changers, we can say. This is officially given by no one since now. It is once said that in the Wall Street Journal that, a celebrity who entertains the public and having a greater number of followers are the one who can influence the market in any possible way.

This statement is released when there is a cold war between two famous YouTubers. When Your Tube didn't include one of the famous celebrity in their farewell video, it got most dislikes ever before for a farewell video of YouTube.

Thus, the user having more popularity is the one who is influencing the sentiment on a scenario in daily life in some way (either positive or negative). Taking this as a foundation, we planned a strategy which can improve the time efficiency of the algorithms on opinion mining.

H. Improved model for Sentiment Analysis

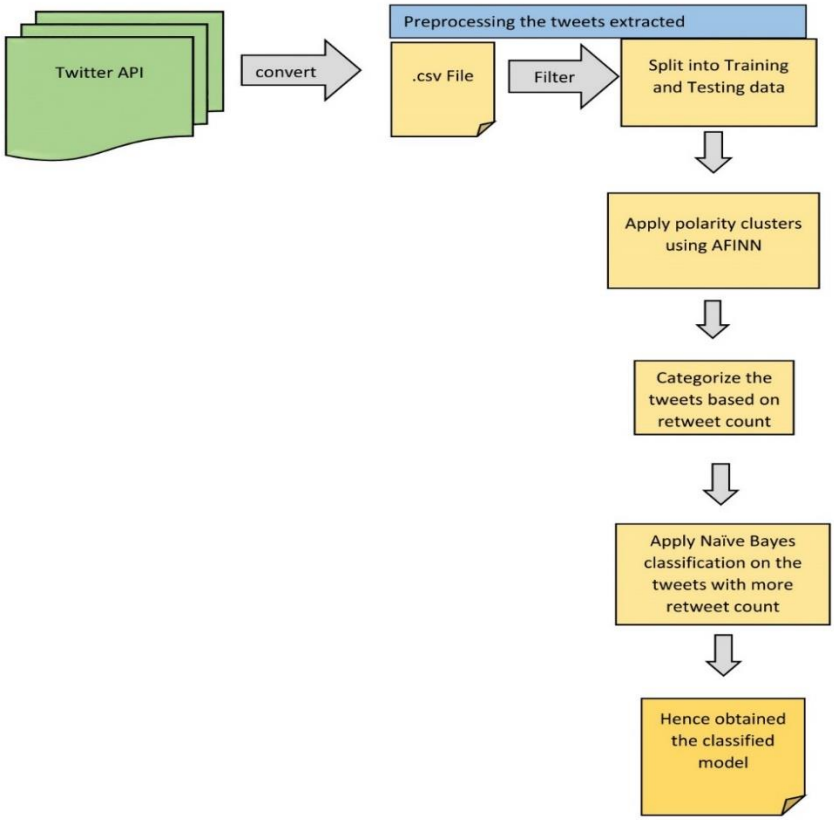

Figure 3: Improved Model

Figure-3 depicts the improved version of opinion mining on the data extracted from Twitter.

Algoritm-2:

1. For (i=0; i <= length(tweets) ; i++)
2. begin,
3. Array polarity[i] = polarityScore(tweet[i]) //using AFINN corpus
4. return Array_polariy,
5. end.

polarityScore(x) is defined in Algorithm-3.

Using the Algorithm-2, tweets of the twitter are assigned with the polarity scores. After finding the polarity scores of every tweet, now it's time to look after the retweet count. This can be found using Algorithm-1. After finding the retweet count, we mapped it with the polarity scores. We plotted this relation. Hence obtained the graph depicted in figure-2.

Now tweets are categorized based on the count of retweets. We took the tweets only whose retweets are above 1000. This is because we assumed that popular people get a minimum of 1000 retweets on average for any tweet they do.

Then after NB classification is applied on the tweets with more retweet count. The overall sentiment is evaluated from these tweets. We have found that overall sentiment of this minute number of tweets (tweets with more retweet count) gave the overall sentiment, which is as same as the overall sentiment of the tweets (tweets with less retweet count included). The Table-1 illustrates the idea in an abstract way.

Tweets with retweet count>1000	**Tweets with retweet count<1000**
Count=17	Count=5144
Overall polarity = -2	Overall polarity = -565
Average sentiment per tweet = -0.11	Average sentiment per tweet = -0.109

Table-2: *Illustration of sentiment from popular tweets v/s regular tweets*

As we have seen in the above table, tweets are categorized based on the retweet count. Then, the polarity score for each tweet is calculated from Algorithm-3

Algorithm-3:

1. Import the afinn module
2. initialize afinn = Afinn()
3. calculate score = afinn.score(text)
4. return score

AFINN is the library which works on the word-rating principle. This AFINN module can be taken as the corpus which is used to mark the valence of the words in integers. These integers, of course, contain the values ranging from minus values to plus values.

I. Classification of new tweets

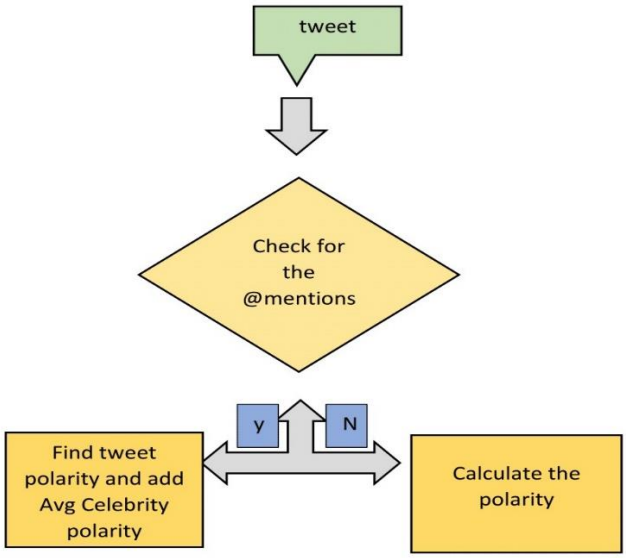

Figure 4: Classification of an individual tweet

Above block diagram illustrates the scenario, how a tweet apart from the dataset is evaluated for sentiment. This is required because, whenever a new tweet with specific # (for example: #bahubali) comes into the algorithm it is classified according to the @mentions in the text. The next time whenever you want to find the polarity or overall sentiment of the public on a topic, this @ and # technique is very useful in terms of both time and accuracy.

Example:

Tweet given: RT @RNTata2000: The government's bold implementation of the #demonetization program needs the nation's support.

In the given tweet @RNTata2000 and #demonetization if found. This means the above tweet is retweeted from @RNTata2000 and topic that is tweeted is demonetization.

Our aim is to build a model to find the true polarity for easy classification of texts in the future. The model is built as mentioned in the block diagram of Figure-4.

Initially, the tweet is checked for @mentions. If yes, the individual tweet polarity, as well as @mention's tweet on the #topic polarity, is calculated. And the average of both is taken as true polarity. The reasons for calculating true polarity is as below:

(a) The text may have positive words, but with sarcastic context
(b) The user may be fan or anti-fan of the @mention he's referring to

In both cases, we are mining true sentiment as stated above through the procedure that is mentioned in the block diagram in Figure-4.

4. Results Discussion

As of now, we are testing three approaches: (i) SVM (ii) NB classifier (iii) Integrated SA classifier. We have used the time taken to classify and accuracy of the classification as the metrics to compare the models mentioned above. We have performed the test over 5000 texts extracted from the Twitter API. Twitter API is a good source for the data to extract from. Twitter provides flexibility to extract the tweets based on the hashtags (which they normally refer to indicate topic on trend).

However, as per our knowledge, we have found that our approach is performing better in terms of time and accuracy than the other two which we are taken into consideration. The time taken by the NB is much more than the SVM.

The same is depicted in the below graph

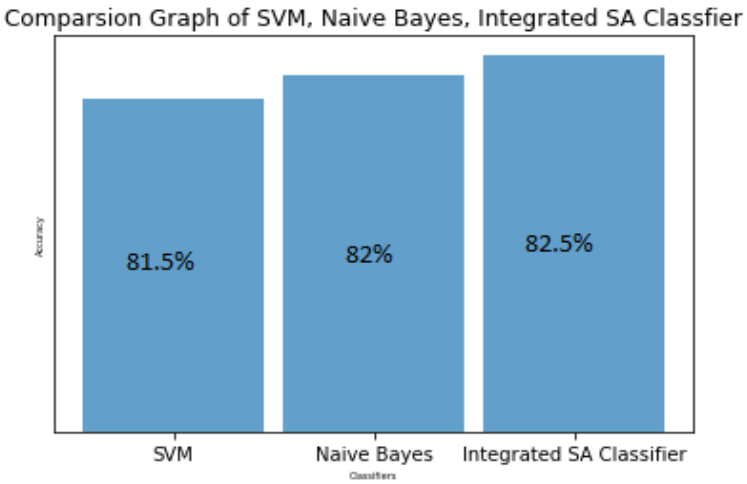

Figure 5: Comparison of three approaches

Approach	**Accuracy percentage (%)**
SVM	81.5
NB classification	82.0
Integrated SA classifier	82.5

Table-3: Accuracy comparison

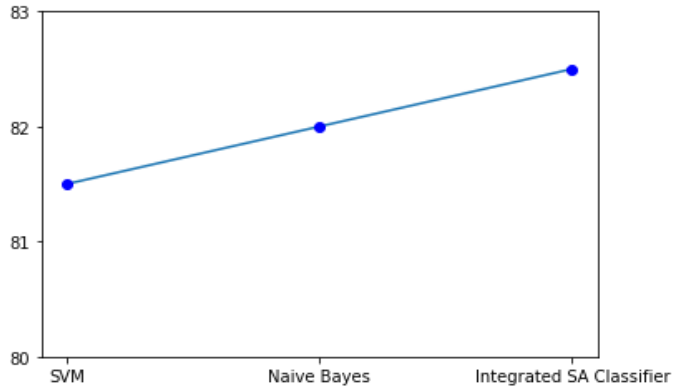

Figure 6: Dot and Line Graph for comparison

Since the regular tweets have a large amount of data, we sometimes find the uncertain data (with special symbols, irrelevant texts, retweets). By filtering them with the popular tweets and replacing with the overall polarity or overall sentiment on the data given, our approach has given the best performance among the above-mentioned approaches.

5. Future work

Till now, we have covered a miniscule of the data available in the world, i.e., every person who in some way connected for the market need not use the twitter as his platform of sharing his views. Even if he shares, it is not necessary that he shared what he is truly feeling. Our further research will be processing of videos and images from the public and mine the opinion of the public from the processed information. We are planning to conduct an experiment on any one of the topics and collect texts, videos, and audios through which, we will build a hybrid model to mine the sentiment on the topic. Along with this, we are also making a tool using ML which acts as the counter sentiment analytical tool. It means, it makes the text into unbiased or uncertain without loss of its context.

6. Conclusion

As discussed above, we consider the NB classification as the foundation to classify the data, then we applied calculations to find retweet count, this improved the time efficiency of the algorithm. And when we take accuracy into consideration, it is also increased too. Because the number of tweets is optimum for finding overall sentiment. This avoids the unconventional, uncertain and useless data to be taken into the classification algorithm.

References

[1]. Walaa Medhat, Ahmed Hassan, Hoda Korashy (2014) 'Sentiment Analysis algorithms and applications: A Survey', Ain Shams University Journal, Vol 5, No. 4, pp. 1093-1113

[2]. Federico Neri Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas (2012), 'Sentiment Analysis on Social Media', *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,* Istanbul, 2012, pp. 919-926.

[3]. Devika M D, Sunitha, Amal Ganesh (2016) 'Sentiment Analysis: A Comparative Study On Different Approaches', Fourth International Conference on Recent Trends in Computer Science & Engineering, Procedia Computer Science, Elsevier, ScienceDirect, Vol 87, pp.44-49

[4]. Mohamad Syahrul Mubarok, Adiwijaya, Muhammad Dwi Aldhi (2017) 'Aspect-based Sentiment Analysis to Review Products Using Naïve Bayes', AIP Conference Proceedings, Vol 1867, No. 1.

[5]. Mohit Mertiya and Ashima Singh 'Combining Naive Bayes and Adjective Analysis for Sentiment Detection on Twitter',International Conference on Inventive Computation Technologies (ICICT),IEEE,2016,pp.1-6

[6]. Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath (2016), 'Classification of Sentiment Reviews using N-gram Machine Learning Approach', Expert Systems with Applications, Elsevier, ScienceDirect, Vol 57, pp. 117-126

[7]. Andrea Ceron, Luigi Curini, Stefano M Iacus, Giuseppe Porro (2014) 'Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France', New Media & Society, Sage Journals, Vol 16, Issue 2, pp. 340-358

[8]. C.J. Hutto, Eric Gilbert (2015) 'VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text', Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM, Association for the Advancement of Artificial Intelligence.

[9]. Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu (2009) 'Adapting Naive Bayes to Domain Adaptation for Sentiment Analysis', European Conference on Information Retrieval, Springer, Vol 5478, pp. 337-349

[10]. Emma Haddi, Xiaohui Liu, Yong Shi (2013) 'The Role of Text Pre-processing in Sentiment Analysis', Procedia Computer Science, Elsevier, ScienceDirect, Vol 17, pp. 26-32

[11]. Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, Amir Hussain (2016) 'Fusing Audio, Visual and Textual Clues for Sentiment Analysis from Multimodal Content', Neurocomputing, Elsevier, ScienceDirect, Vol 174, Part A, pp. 50-59

[12]. Martin Haselmayer, Marcelo Jenny (2017), 'Sentiment analysis of political communication: combining a dictionary approach with crowdcoding', Qual Quant, Springer, Vol 51, Issue 6, pp 2623–2646

[13]. Hanhoon Kang, Seong Joon Yoo, Dongil Haan (2012), 'Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews', Expert Systems with Applications, Elsevier, ScienceDirect, Vol 39, Isuue 5, pp. 6000-6010

[14]. Malcolm Baker, Jeffrey Wurgler (2007) 'Investor Sentiment in the Stock Market', Journal of Economic Perspectives, Vol 21, No 2, pp. 129–151

[15]. Doaa Mohey El-Din Mohamed Hussein (2016) 'A survey on sentiment analysis challenges', Journal of King Saud University – Engineering Sciences, ScienceDirect, Vol 30, No. 4, pp. 330-338

[16]. Tomohiro FUKUHARA, Hiroshi NAKAGAWA, Toyoaki NISHIDA (2007) 'Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events', Proceedings of ICWSM'2007

[17]. Alesandro Bessi, Emilio Ferrara (2017), 'Social bots distort the 2016 U.S. Presidential election online discussion', First Monday Peer- Reviewed Journal, Volume 21, No 11.

[18]. Anshul Mittal, Arpit Goel (2012) 'Stock Prediction Using Twitter Sentiment Analysis', Standford University, Volume 15, pp.1-5.

[19]. Johan Bollen, Huina Mao, Xiaojun Zeng (2011) 'Twitter mood predicts the stock market', Journal of Computational Science, Elsevier, ScienceDirect, Vol 2, Issue 1, pp. 1-8

[20]. Xian Fan, Xiaoge Li, Feihong Du, Xin Li School, Main Wei (2016) 'Apply Word Vectors for Sentiment Analysis of APP Reviews', IEEE/3rd International Conference on Systems and Informatics (ICSAI), Shanghai, pp. 1062-1066.