REVIEW ON TWEETS SENTIMENT ANALYSIS BY MACHINE LEARNING APPROACHES

[1] Apurva Dadhwal [2]K.L Bansal
[1]M.Tech. Student, [2]Professor
Department of Computer Science
Himachal Pradesh University
Shimla, INDIA

Abstract-
The available data on social media has contributed to vast research using sentiment analysis. The twitter-based social media represents a gold-mine approach for analyzing the performance of the brand. Large opinions of the people are found over Twitter that are honest, informative, and casual as compared to the formal type of data-survey analysis using magazines or reports. Millions of people share and express their sentiments over the media discussing about the brands whom they interact with. When such type of sentiments are identified over the media, then the information gained from such sentiments represents fruitful results benefiting large companies or organizations Sentiment analysis has turned out one of the most significant tools in natural language processing because it opens up numerous possibilities to understand people's sentiments on different topics. The purpose of an aspect-based sentiment analysis is to understand this further and find out what someone is talking about, and whether he likes it or does not like it. There have been various ways to deal with handle this issue, utilizing machine learning. In this thesis, labeled data is used on the basis of polarity and Tweets preprocess and extract unigram features after preprocessing of the tweets.

***Keywords*— Sentiment analysis, Tweet Sentiment Analysis, PartsOfSpeech, Support Vector Machine, K Nearest Neighbour, Natural Language Processing, Long Short Term Memory**

I. INTRODUCTION

The great impact of social-media world wide has led to the discovery of sentiment analysis. The recent devlopments of smart technologies using mobile-based communication has entailed massive amount of data creation. The social media provides an ability to share thoughts, opinions, and emotions. The term sentiment analysis (SA) is popularly known as opinion mining which is a process of emotion classification usually conveyed by a text that may be positive, negative or neutral. The available data on social media has contributed to vast research using sentiment analysis. The twitter-based social media represents a gold-mine approach for analysing the performance of the brand. Large opinions of the people are found over Twitter that are honest, informative, and casual as compared to the formal type of data-survey analysis using magazines or reports. Millions of people share and express their sentiments over the media discussing about the brands whom they interact with. When such type of sentiments are identified over the media, then the information gained from such sentiments represents fruitful results benefiting large companies or organizations. This data is very helpful to monitor performance of different brands and to locate time periods and aspects receiving polar sentiments. The brands can be celebrities, political parties or events, products etc. Approximately more than 500 million Tweets are generated over daily basis which represents a huge/vast collection of data for the process of analysing the brand performance used by the members or teams of companies on manual basis [1, 4]. The tweets diversity cannot be captured probably by using constant or fixed rules. It is very difficult to measure tweet sentiment analysis due to its complex behaviour as compared to a well formatted documentary. The tweets do not rely over any formal type of language nor over any formal language word. The symbols and punctuations are basically used to express opinions such as emoticons, smileys, etc. So, the thesis work presents the supervised learning approaches and natural language processing techniques for understanding the concept of tweets based on its characteristics and patterns including sentiment-based queries.

1.1 Sentiment Analysis

The concept of sentiment analysis is understood by combining the terms "Senitiment" and "Analysis". The word sentiment represents feeling that can be joyful, confusing, irritating, distracting. The sentiments are the feelings based on certain attitudes and opinions rather than facts due to which sentiments are of subjective nature [2]. The sentiment implies an emotion usually motivated by opinion or perception of a person. The psychologists attempts to present multitude of emotions classified into six distinct classes: joy, love, fear, sadness, surprise and anger. The emotions based on sadness and joy are experienced on daily basis at different levels. We are mainly concerned about sentiment analysis detecting a positive or a negative response or opinion [2]. The major significance of sentiment analysis is that every emotion is linked to human perception forming an ingrained part of all humans which means that every human has the potential to generate different opinions acting as a tool for sentiment analaysis. Sentiment analysis refers to the analysis automation of a known text determining the distinct types of feelings conveyed. The term sentiment analysis and opinion mining can be used interchangeably [3]. Sentiment analysis as defined as an information extraction and natural language processing task with an aim to gain the feelings of writer expressed positively or negatively based on requests, comments or questions analysing large data-sets or documents. It basically intends to define writer's feeling

regarding a specific topic based on writer's own opinion [8]. It models a branch that can help in providing a judgement over distinct fields. The measurement of sentiments is a biased technique with it is really complex to achieve high accuracy of automated systems.

1.2 Twitter Sentiment Analysis

Sentiment analysis is a fairly growing field. Approximately 81 percent of the web-users usually 60 percent of the Americans have performed a research on a product online analysing that each year articles with different text domain forms are targeted over years. One such example experimented was comparison of consumer confidence based Gallup polls and Twitter sentiment [5]. The obtained results were positive and the value of correlation was 0.804, suggesting impeding that one can use Twitter for measuring different public opinions. So, the study user analysis of Tweets to extract distinct opinions, determining the polarity of the tweets on real time basis. A popular micro-blogging site named Twitter allows its users to write entries or texts up to 140 characters popularly known as Tweets [7]. Twitter has approximately 302 million users on monthly basis. Out of which 88% of the users have freely readable tweets and around 80% of the users have placed their location over the profiles. The Twitter created data is usually available through Twitter's API representing information on real-time basis as a stream opinionated data form. The tweets can be easily filtered using both the publishing time and the location. This has designed a new sentimental analysis sub-field named as Twitter sentiment analysis (TSA). The process of implementing natural language processing over textual form of data from the twitter media represents certain new challenges due to data informal nature [6, 9].

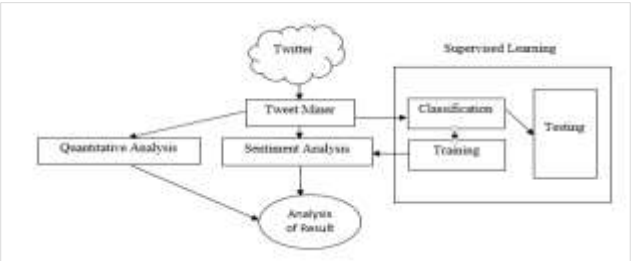

Figure.1 Twitter Sentiment Analysis

Tweets generally contains spelling mistakes and problem of character limitation resulting in abbreviation mistakes. Many unconventional methods on linguistic basis like words elongation or capitalization are used. In addition, tweets consists of unique features like hashtags and emoticons having an analytical value [10, 13]. The hashtags are used for categorization and mechanism of searching represented as "#" whereas the emoticons usually expresses different emotions expressed as ":-)". A replying form if tweet as directed to another user, a symbol "@" is mentioned in front of the person's that is to be replied [15].

II. SENTIMENT ANALYSIS: METHODOLOGY

The development of sentiment analysis was done using standard machine learning approach as discussed below in the following section.

2.1 Labeled Data

SemEval 2014 [11] organizers provided the data-sets for the purpose of research. Large Tweet categories like chat tweets, interactions of celebrities, and politics formed a part of it. It was reported that the Tweet generally represent a great part of specific twitter vocabulary like mentions and the hashtags. The dataset collected was annotated on Mechanical Turk i.e. a internet market place by hand [1, 13]. The organizers provided guidelines to annotators considering positive, negative or neutral behaviour in order to reduce the subjective nature at sentiment annotation. The publishers stated that the class distribution was reflective of the practical world tweets and highlighted that the set of data was cleansed from any kind of exception. One such example of tweet labelled sentiment was published over the website as represented below:

Table.1 Training data-set Class Distribution

Labelled Sentiment	Share
Positive	38%
Neutral	47%
Negative	15%

SemEval 2014 was also responsible for the labeled data-set development used for experiments internally and deep analysis. Such kind of data set contained a total number of 1,655 tweets with a distribution of class analogous to the data training set [10]. In order to identify better algorithm performance and the features often very useful, the developed data was used which forms a standard practice in the field of machine learning analysis. More often, this data was used by many researchers and the adjustment of the system to perform well over the data-set was done. The most important requirement was that the set of data does not perform overlapping operation over the trained set of data that was quite ensured by Rosenthal et al. (2014). The researchers concluded an analysis qualitatively in order to explore the developed form of data set. Some of the instances annotated were incorrect [4]. In the following example, the tweet sentiment analysis was marked neutral but the text conveys the sentiment in a positive way.

From the statement above, such an instance was marked negatively due to noise in data-set. But such a data-set is better significantly than other forms of data sets explained automatically using machine learning supervision methods [7].

2.2 Pre-processing

This process was done before the tweet-based usage of feature extractor in order to design or build the feature vector. The process is initialized using the following steps. Such steps convert the plain tweet text into the elements of processing nature with an additional information utilized by the feature extractor [22]. The tools of third party were used for all the steps specifically handling tweeted text unique nature.

> ***@_Nenaah oh cause my friend got something from china and they said it will take at least 6 to 8 weeks and it came in the 2nd week :-P***

Step1: Tokenization: It is a process of text conversion as a string into elements process-able known as tokens. In terms of tweets methodology, such elements can be emoticons, words, links, punctuations or hashtags. As shown in fig.3, "an insanely awsum time..." text was busted into "an", "insanely", "awsum", "time"….

The elements here get separated by some space whereas the sentence based punctuation ending such as full stop or exclamation mark get separated more often by a space. The hastags along with symbol "#" that precedes the tag is required to get retained as the symbol "#" may suggest distinct sentiments than the word to be used regularly in text. Therefore, the Twitter-based particular form of tokenizer helps to extract tokens.

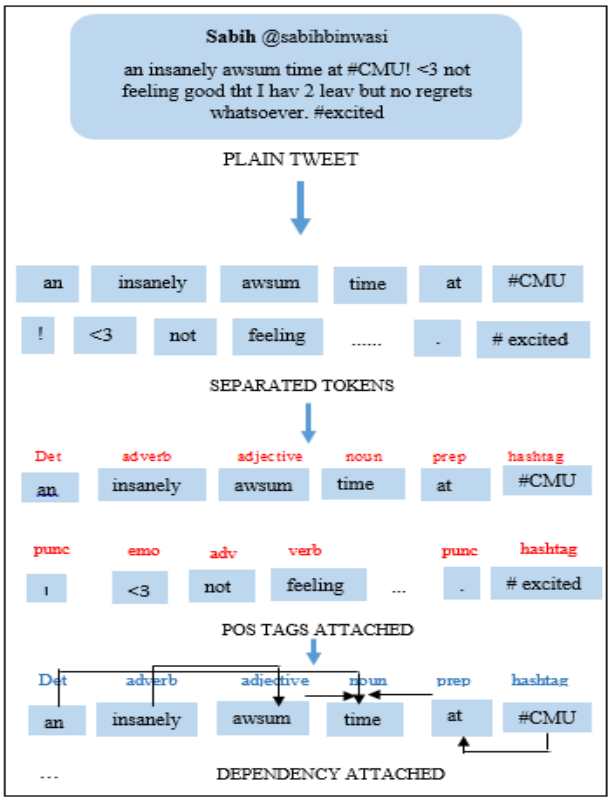

Figure.2 Pre-processing stages

Step 2: Parts of Speech Tags (POS)***:*** The POS tags are generally characterised by sentence based words dependent over the different categories of grammar in context to a word. Such an information or data is necessary for the process of sentiment analysis as words contain distinct values of sentiment that are basically POS dependent . Considering an example, the word "good" acting as a noun has no sentiment whereas "good" in its adjective form reflects a sentiment in a positive way. Fig.2 above shows that each token gets extracted in its last step and gets assigned a POS tag. The accuracy maintained by a POS tagger is about 93%.

Step 3: Dependency Parsing: It represents the relationship extraction among sentence-based words. Such a method is very useful for identification of relationship between "good" and "bad" in form of phrases like "not really good" where there is only adjacent word relationship. It explains the parent-child relation between tweet tokens as shown in fig.2. The accuracy maintained by dependency parsing is about 80%.

2.3 Feature Extraction

This is a process of designing a feature-vector from a known tweet. The entry (each) in case of feature vector is an integer that contributes on the attributing a class of sentiment to a tweet class. Such a contribution strong to negligible form. Here, the strong class represents feature based value entry influencing sentiment true class whereas the negligible class presents no such relation

between sentiment class and feature value. The algorithm here identifies the strength dependency between classes and the features using strong correlated form of features and preventing noisy-feature usage.

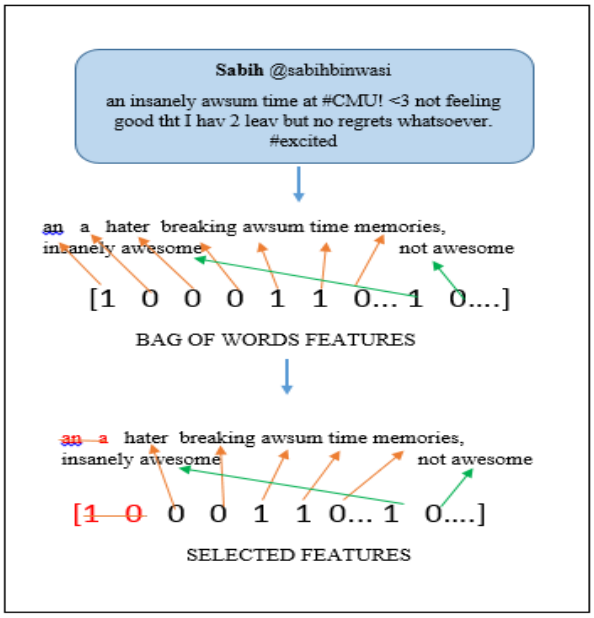

Figure.3 Feature Extraction

2.3.1 Feature Set-Bag of Words**:** These represents "bag or words" called unigrams as a set of features where the token frequency is considered to be a feature vector. This feature set was unanimously selected by the practitioners. The feature vector based entry is assigned to each of the specific tokens that were found in the trained labelled set. If such a token occurs in tweet, then it gets assigned by 1 as a binary value, otherwise it is considered to be zero [2]. As shown in figure above, the order of token-based sequence or the structure of grammar is not at all preserved. For instance, the token "awsum" forms a part of tweet hence it is marked or assigned as "1" whereas the token "hater" does not occur in tweet, hence it is labelled as "0". The token "hater" has a column for it in a feature vector. Hence, it would take place in some tweet in the trained data-set form. It was further analysed that the indication of only word presence yields good performance than the word-based frequency. In such case, an entry is also assigned for specific ordered token pairs termed as bi-grams. The token pair "insanely awsum" where it is assigned or labelled as "1" if it forms a part of tweet, otherwise it is considered to be "0". It indicates that the system is equipped not only to indicate the token presence but also indicate its context.

2.3.2 Feature Selection: It is expected to add unigrams, bigrams and trigrams adding large entries to a vector feature which can make the space of vector highly dimensional resulting in a more complex and hard task to identify the relation among each feature [9]. This represents the major issue of text-based classification and is popularly known as Curse of Dimensionality. The process notices that some of the features are not relevant for the operation of sentiment analysis, hence these un-necessary features are required to be removed. So, different researchers conducted the study of selecting the features as per the requirement. A feature attribution evaluation was conducted to analyse the impact of features. A very popular method named Chi-Squared Feature Selection was used which carried a classification algorithm evaluating the feature-based dependency value and the dependency of each of the class. Thus, if the feature has a high correlational dependence, then it is assigned with a very high rank. The method of Chi-Squared Feature Selection outperform well for text classification.

(c) Social Media Feature Set: Emoticons are used as symbols basically to express the gesture or feelings using language characters and punctuation. In fig. the emoticon "<3" depicts heart representing love symbol. These emoticons strongly analyse the positive or negative sentiment.

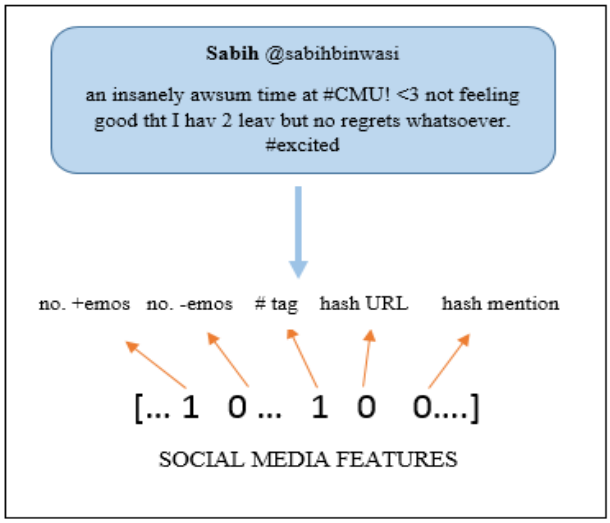

Figure.4 Social Media Feature Set

The figure above shows a tweet sample illustrating the social media feature set representing a number of positive emoticons labelled as "1" as per the dictionary while no such emoticons are found. In feature-based vector, the presence of hashtag, URL and mention ("@username") is also included. Such types of features are mostly used in the showing the differentiating impact between the polar i.e. positive and negative and non-polar i.e. neutral class. These type of indicators helps to identify the relation between sentiment class and the indicators itself. For, instance, a tweet with "@username1" is considered as a rare form of token in the trained data set, the algorithm of classification would fail in identifying a relation. With "@" considered to be as the binary feature, the chances for relation formation increases.

(d) Lexical Feature Set: These are basically driven by the use if lexicons. The task of sentiment lexicon analysis maps the n-grams or tokens to score-based polarity. These have been reported successfully with an ability to locate issues based on classification methods of sentiment analysis [1, 13]. After the process of bag-of-words features, Mohammad et al., 2013 analysed such a feature to be most successful one, solving the issue of identification of full-text sentiment, token sentiment using the lexicon-based feature set which makes the task very easy. We take an example of *AFINN* which is an affective lexicon discovered by Finn Årup Nielsen. Such type of lexicon was constructed manually for the mapping of the most frequently used words on Twitter having a ranges from -5 to +5, where, -5 is labelled as a negative type of token and +5 is considered as the most positive type of token. The tokens which does not contain any kind of sentiment are labelled as "0". Each of the tweet token is labelled independently in a polarity score. For instance, "time" is not labelled any type of score.

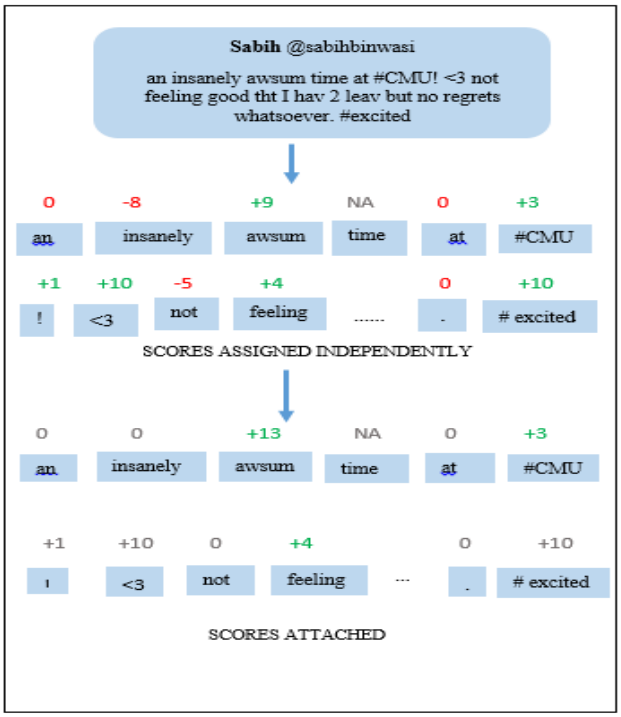

Figure.5 Lexical Feature Set

The lexical based feature set are subdivided with the following features:

(1) Handling Negations: As shown in the figure above "good" is labelled with positive type of score (+8). However, in case of "not feeling good", the "good" must be labelled or assigned a score i.e. negative. This creates an issue of negation context. The tokens "insanely" and "awsum" denotes a negative and a positive score as shown respectively. If one considers "insanely awsum", then one cannot used them in independent form but it intensifies the sentiment expressed by "awsum" which denotes intensifier context problem [6]. For solving such a problem, a list is used along with tree parsing dependency for readjusting the scores. Each time a word creates a negation form such as neither, never, not etc. where the system basically target its token head word . This negated token type is usually found above the level of negation. The score of such negated word gets readjusted.
(2) Score Combination: This method is helpful in combining the scores to improve the performance of the system. For instance, if there is existence of a most positive word having a +6 score and five negligible tokens of negative type with a score of +1, then the final score is considered as 0 which indicates that there is no type of sentiment that occurs in a given tweet. Considering an idea of having six buckets attributed to its strength and polarity i.e. positive, very positive, mildly positive, negative, very negative, and mildly negative. Thus the score range of such lexicons gets segmented arbitrarily into six type of categories [2]. In addition, such type of feature-sets contains information of raw type as compared to six bucket based feature set which enables classification algorithm to gain more relevant type of information.

2.4 Training Classifier

The above processes involved the use of bag-of-words, social media and lexical types of feature sets. As the feature vector is designed for each instance type in the trained form of data-set, such vectors are further fed to the training classifiers known as learning algorithm. The concept of Support Vector Machines (SVM) is used for classification of algorithm with binary classification process [12]. Such type of method helps in analysing different feature vectors with an assigned class in order to identify the relation dependency between a sentiment and each of the feature [13]. Here, each of the vector is considered as a point of data in vector dimensional space that equals to the size of feature-set. The SVM helps in identifying the vector dimension based hyperplane which divides the class into two types. One is the considered as "best" i.e. defined as a good type of separation gained by the hyperplane having the large distance to the point nearest to the training data type of any kind of class known as functional margin. In general, if the margin is large then the classifier error gets reduced. When the new form of tweet i.e. unlabelled is fed into the system, it helps in extracting the feature vector same as that of labelled tweets. Finally, the vector is fed to the learning model as an input. Such kind of hyperplane helps to check the new data-point location on the hyperplane side. Further, a class is assigned or labelled. Note, such a process accounts for only two type of classes whereas the problem of classifying a sentiment contains three type of classes. For handling this issue SVM uses all binary classification for all the type of classes. This technique as its first step involves testing process searching the instance class (one or two). If the output obtained is true then the process ends. If the process is not true, the test is done for other type of class to check whether the instance belongs to the same class or does not belong to that class. If this is also not true, then the instance gets assigned by the third type of class. The researchers reported the performance where firstly the SVM considered to be better with the features based on bag of words as compared to the Maximum Entropy and Naive Bayes. Secondly, the SVM with large number of bag of words, would help to boost more admissible features.

2.6 Impacts of Feature Sets

The feature sets have significantly improved the system's performance. The bag-of-words i.e. only unigrams were used as a baseline as such type of feature-set is considered to be a standard type of feature-set in the field of classifying sentiments. The use of bi-grams and tri-grams added enhancement/ improvement that was noticed when the context of tokens was verified which was very helpful in the prediction of sentiment whereas on the other hand, the social media features were not so helpful. It was considered that the lexicon based features enhanced the system performance significantly.

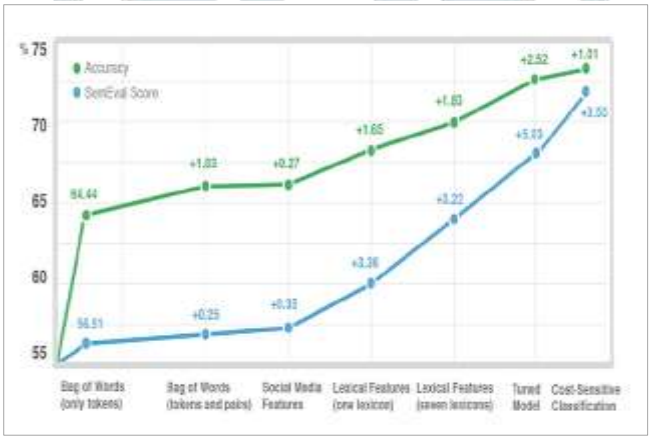

Figure.6 Impact of Feature Sets

The performance of the system was improved further when there was use of more than one lexicon which indicates that the addition of multiple type of lexicons provide text-data extra coverage and provides multiple perspectives on the basis of token polarities. Finally, the use of cost sensitive type of classification boosts the score of SemEval data set. As the polar type of classes gets more affected than the majority type class. So, such a technique was expected to boost the score of SemEval more than its level of accuracy [4].

3. SA: Tools, Levels, and Techniques

This section describes the sentiment analysis based on the tools, levels, and the type of techniques used.

3.1 Background Theory: The concept of sentiment analysis is comprised of common concepts of natural language processing in addition to many machine learning methods as discussed below:

3.1.1 Machine Learning: This forms the cornerstone in the field of sentiment analysis. This section is comprised of brief analysis of the supervised and unsupervised machine learning algorithms.

1) Support Vector Machines

- Considers the data points based on the type of spatial location which further attempts to split the space of feature into optimized segment class known as trained SVM.
- Trained SVM is basically used for new examples classification by labelling them a specific class on the basis of feature-space segment [5].
- This method is famous for linear classification which deals with two linear separating classes, where the space of feature gets divided into segments of class by the creation of hyperplane with maximum margin among the two classes.
- Maximum margin is an essential requirement for SVMs. In fig. the closed points of data of both the classes that are parallel to the hyperplane defining vector represents the support vectors.

2) Naïve Bayes Classifier

- Also referred to as Naïve Bayes Learners.
- Simple probabilistic classifiers based on Bayes' Theorem.

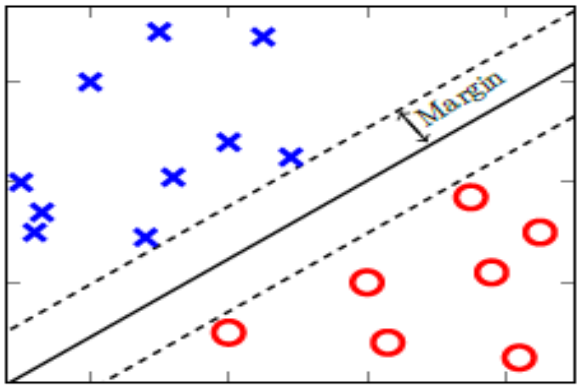

Fig.7 SVM linearly separating two classes

For instance, a document (d) is given. Here, the task of labelling a document of (c) class is done by generalizing probabilities of each of the class, and thereby finding the maximum a posteriori i.e. MAP estimate of probability.

$$cMAP = \underset{c \in C}{\operatorname{argmax}} P\left(\frac{d}{c}\right)$$

With the use of Bayes Theorem, the formula for each class is given as:

$$cMAP = \underset{c \in C}{\operatorname{argmax}} P\left(\frac{P\left(\frac{d}{c}\right) \times P(c)}{P(d)}\right)$$

As the each document probability is same all over the classes, then the class that maximises the numerator form is also same as the class which maximises the overall expression. Thus, the denominator of the above equation can be dropped.

$$cMAP = \underset{c \in C}{\operatorname{argmax}} P\left(\frac{d}{c}\right) \times P(c)$$

The document is expressed as a feature attribute based vector

$$d = a_1\, a_2\, a_3, \ldots\ldots, a_n$$

Thus, the approach is represented as:

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{a} P\left(\frac{a}{c}\right)$$

3.1.2 Natural Language Processing: It is basically concerned with the human natural languages for the computer-based communication process.

1) Linguistic Negation

- Grammatical concept used to reverse the truth of language-based propositions value.
- It defines two major forms. One is the syntactic negation and the other is the morphological negation [15]. In case of Syntactic negation the words set gets negated by a phrase or a word whereas in case of morphological negation, the individual word forms gets negated using an affix form.
- Syntactic negation splits the negation into two types: denials of assertions and rejections of suggestions.

2) Bag-of-Words Model

- It lists term occurrence as well as frequency of term occurrence, order of the term, and disregarding grammar

- The machine learning classifiers can be used directly resulting as feature vectors.
- Here, the used Term Frequency-Inverse Document Frequency (TF-IDF) is calculated as:

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

$\text{tf}(t, d) =$ Term frequency

$\text{idf}(t, D) =$ Inverse document frequency

- Part-of-Speech Tagging: It categorizes the sentence tokens into distinct parts of speech such as verbs, nouns, adverbs, and adjectives. It can also solve the word ambiguity problem.
- Syntactic Parsing: It involves the analysis of string of text using certain grammar rules sets, term ordering according to syntactic relation among each other. The result is observed in the form of parse trees due to human languages based substantial ambiguity. The statistical parsing and the methods of machine learning can be used for solving the ambiguity problem.
- Dependency Parsing: It helps in identifying the sentence-based main verb, and all other forms of syntactic units either dependent over the verb on direct or in-direct basis. These are generally dependent on syntactic rules [10].

NLP business applications in the current scenario includes the following methods:

- Text classification
- Chat Bot
- Customer service
- Text summarization
- Sentence segmentation
- Customer service
- Ad placement
- Market intelligence
- Reputation monitoring
- Regulatory compliance
- Machine translation

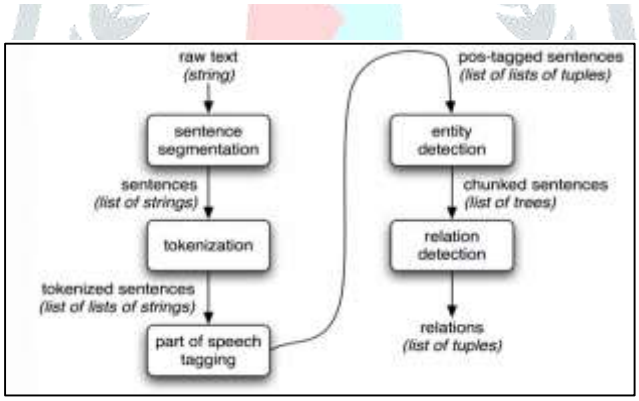

Figure.8 Natural Language Tool Kit (NLTK)

The mechanism of NLTK provides processing tools for certain languages. This kind of toolkit involves the processes like sentiment analysis, data mining, data scraping, machine learning, and many other tasks of language processing [12]. It provides a leading platform for modelling Python programs in order to work with human-based data language which provides an easy interface such as WordNet, along with libraries of text-processing for tokenization, semantic reasoning, stemming, parsing, tagging, and classification.

3.2 Sentiment Analysis: Tools

The tools used for the Sentiment Analysis are discussed in the section below:

3.2.1 Tweet NLP: It represents the collection of certain tools for performing NLP tasks in connection to the Twitter's conversational language. Tweet NLP includes hierarchical word clusters, a tokenizer, POS tagger, dependency parser. The POS tagger forms a significant tool for many application of NLP for modelling feature vectors for a specified classifier. The Tweet NLP based tagger presented an accuracy level maintained above 93%.

3.2.2 Scikit-Learn: This basically represents a framework for Python programming language offering large machine learning models as well as the tools for data analysis and pre-processing. It provide state-of-art applications for large machine learning models with its aim to pay particular attention to consistent API, good type of documentation, and good performance. Such type of documentation provides a simplified structure for both the experienced and in-experienced type of readers to search the deep information and to take an over-view of the topic, respectively [5, 8]. Some major significant aspects are:

1) Transformer: One such object is the necessary condition for the performance of machine learning methods along with Scikit-learn. The other objects may include expand, clean, or feature generation representations.
2) Pipeline: The tool Scikit-learn helps in providing a framework for pipelining methods of machine learning tools which makes the process to chain the tasks easily like feature extraction and pre-processing along with algorithm based on machine learning in a very tidy way. The framework of pipeline performs grid-based search for all the estimators on the basis of parameter. This estimator is any kind of object that basically learns from the data such as classifier, or Transformer object extracting or filtering impressive features from raw type of data in Pipeline methodology. A pipeline is trained in whole, thereby resulting in an object performing all the required steps such as classification, pre-processing, and feature extraction.
3) Grid search: It is very useful process for selection of vectorisation of n-grams combinations working well for some of the parameters of the classifier.

3.2.3 Pandas: It represents an open source library which provides high-level of data performance of data structures, and it involves the analysis of data for Python programming language. It involves different tools for effective writing and reading of data between distinct textual file formats and in-memory structures of data like comma-separated value-based files [7].

3.2.4. CRFSuite: It is an implementation based on a classifier named Conditional Random Fields (CRF) sequence. Such a classifier is presented in C++ programming language. The parameters C1 and C2 forms the input parameters for each of the CRF classifier for the settlement of L1 and L2 normalization levels coefficients, respectively.

3.2.5 Mechanical Turk: The Amazon Mechanical Turk represents a marketplace which requires human intelligence methodology for their work It basically helps in performing large tasks based on human intelligence providing large workforce (real people). It is a useful tool for annotating large tweet number manually that may be positive or negative one [1].

3.4 Sentiment Analysis: Techniques

In the field of sentiment analysis, the techniques of SA represents a major challenging research topic. The main aim of such techniques is to mainly classify the positive or negative opinions expressed by the document. The classification of SA is divided mainly into two distinct methods/approaches [7, 10]. One is the machine learning approach and the other is the lexicon- based approach. The machine learning methodology uses different algorithms of machine-learning whereas the lexicon- based approach is divided into dictionary based-approach and corpus-based approach.

3.4.1 Machine Learning Approach: It relies over sentiment analysis treatment as a problem of text classification. Here, classification of text is basically used in business automated decision that need proper text processing. It uses records (trained set) in order to train a design or model which is used in prediction of fresh records without any kind of label. Each and every record gets assigned to a specified class [9]. If a new form of unlabelled record is known (given), then the model helps in predicting its class label. Such classes may be positive, neutral, and negative. In this kind of approach, two types of its sub-approaches are distinguished based on the methods of learning. One is the supervised learning and the other is un-supervised learning.
1) Supervised Learning: It uses a classifier i.e. specifically supervised that learns from the assigned trained type of documents. The trained labelled documents contain words related to topic as key features. The opinion-based words expresses a positive or a negative opinion. The sub classes of supervised learning include Linear, Decision Tree, Probabilistic and Rule based classifier.

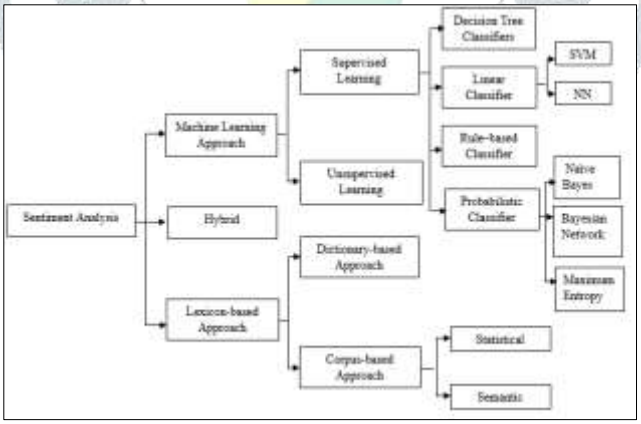

Figure.9 Sentiment Analysis Techniques

(a) Linear classifiers: They are simple in nature. The basic idea is to count the sentence-based negative and positive words and to compare the number of negative and positive words for the determination of sentence-based polarity [11]. These classifiers adds weight to the words. In such a case, the "most positive" words represents highest weight and the "most negative" words represents the lowest weight. The most popularly used linear classifier is SVM Classifiers which aims to pin-point the linear separator with class-based best separation method.
(b) Decision Tree: These methods are used for the purpose of prediction. They can be used easily for classification. Suppose, a record is known (given) with its unknown class label, such a record gets tested against decision tree, and the route gets traced from root to the node which when determines the prediction of class for such type of record [8]. Such methods are famous as their modelling does not need any expert domain or settings. The most widely used decision tree packages for the purpose of text classification implementation are C5 and ID3.

(c) Probabilistic classifiers: Also known as generative classifiers as they generate or form a model for each of the class. They basically use mixture of models assuming that each of the class represents the component of a model. The most famous type of Probabilistic classifier is the Naïve Bayes Classifier. This classifier is of very simple form and it easily gets coded in each and every language of programming as these involve simple mathematical analysis. Such a model works with bag-of-words (unordered-set). The frequency of each word is kept generally instead of its location. This classifier uses Bayes Theorem in order to determine the label from where the known type of feature-set belongs.
(d) Rule Based Classifier: This classifier is similar to decision tree classifier as both of the classifiers on its feature space encode the rules. The major difference is that in case of decision tree classifier, it uses hierarchical approach whereas the rule-based classifiers allows the mechanism of decision space overlapping. Various multiple research work suggest distinct ways to transform decision tree to a rule-based classifier. In case of rule-based classifier, the phase of training generates different criteria rule base. The most used popular forms are the confidence and the support.
3.4.2 Unsupervised Learning: This method is mainly used in creation or design of trained class-labelled documents which is considered to be the most complex method. So, in such cases, one would collect the un-assigned/labelled documents and further implements the unsupervised learning methodology. This technique is used for document based clustering analysis as it does not depend over pre-defined labelled training documents. The summary to be noted is that the supervised learning learns by examples whereas the unsupervised learning learns by the observation method [27, 15]. In other words, the unsupervised model, the learner do not get any information with solutions. But in supervised learning models, examples are presented to the model during the phase of training.
3.4.3 Lexicon-Based Approach: This method represents another type of unsupervised approach, but in such cases, a dictionary with synonyms and antonyms of opinionated phrases and words could be used along with their respective orientation of sentiments. Here, two methods named dictionary-based and the corpus-based methods are used commonly in its automated form. In case of dictionary-Based approach, the main strategy involves the manual collection of small opinion words sets and then the set is grown further by finding large text collections like WordNet. The new searched words are added to opinion set and the cycle gets repeated till no more words are found further. The major disadvantage of such a method is that it completely relies over corpora and there is no collection of opinion words largely each time with its available domain. Note, here all the lexicon-based words does not express the negative or positive opinion regarding an entity [12]. The Corpus-Based approach is used in two of the following conditions. The first is the discovery of new sentiment words from a corpus domain by using the list given of known opinionated words and secondly, the creation of sentiment lexicon from another one. Such an approach is not as effective as compared to dictionary-based approach as it needs a corpus with all words in English. The corpus-based method is divided in semantic and static approach depending over the used techniques.

3.4.4 Hybrid Approach: This type of approach uses both the approaches used above. It used a method that was based on Emotinet. It was proved that the method presented effectively identifies emotions from an expressed form of text either with or without less effective relating words present in it. For achieving the improved results, they used Support Vector Machine Algorithm relying over its major principle to search the linear type separator with its best possible separation among classes [11]. Chi-square test was followed by lexicon-based approach for identification of new tweets. The work involves the use of lexicon-based approach mixed with rule-based approach using manual labelled sets i.e. the case of supervised learning.

III. RELATED WORK

Onam Bharti, et.al [1] proposed an approach using KNN, Naïve Bayes, and the modified version of k-means clustering, and it found that the modified version is more accurate than the KNN techniques and Naïve Bayes individually. The researchers obtained classification accuracy of 91% on overall basis over the 500 mobile review test-set. The algorithm running time is O (n + V log V) for the process of training where n represents the word number in a document and the V represents the vocabulary reduced size. It runs faster than the algorithms of machine learning such as Support Vector Machines and Naïve Bayes classification that takes more time in converging optimally in regard to set of weights. The level of accuracy was comparable to the existing algorithms used for the classification of sentiments based on reviewing mobile. Sara Rosenthal, et.al [2] discussed the Sentiment Analysis in twitter task in the fourth year. SemEval-2016 Task 4 was comprised of five of its sub-tasks. Out of five, three of its tasks represent an important departure from its past editions. The first two forms of sub-tasks are the re-runs the data based on prior years and it was meant for predicting the overall sentiment and the tweet-based sentiment towards a topic. The three new sub-tasks focus over two of the basic variants of Twitter-sentiment-classification. The first type of variant involved a five-point scale conferring an ordinal type of character to the classified task whereas the second variant focussed on estimating the correct estimation of each class of internet prevalence i.e. the task to be called off as supervised literature learning quantification. Then the task continued to be very popular in attracting total 43 teams. Md Johirul Islam, et.al [3] showed the sentiment analysis applications and how the Twitter got connected and ran the queries of sentimental analysis. The experimental analysis was done on distinct queries from the region of politics to the humanity and has shown various interesting results. It was realized that the tweet based neutral sentiment was significantly high that clearly showed the current working lag process. Willian Becker, et.al [4] proposed a translation-free language-agnostic method for the analysis of Twitter sentiment which used the methodology of deep convolutional neural networks (DCNN) with embedding's of character-level for pointing the exact tweet polarity that might be written in distinct languages. The method proposed method was more accurate than various other architectures of deep neural while requiring substantially less learnable parameters. The resulting model was capable of latent features learning from all of the languages that were employed during the process of training in a straight-forward manner.

Kavita Pabreja, [5] discussed the twitter usage as a basic forum for sentiment understanding of the Indian citizens towards Goods and Services Tax was launched recently by the Govt. of India on 1st July 2017. The tweets that originated on 30th June and 1st July in India was analysed. The public emotion based on surprise, anger, joy, sadness, anticipation, were extracted based on distinct opinions. Disha Kohli, et.al [6] focused to analyze the expressed sentiments on Twitter demonetization such that the opinions of public and certain views were extracted, and analysed and further used to understand the positive and negative impact of such an impact on the Indian people. After analysing the sentiments of results, it was observed that many of the sentiments are of neutral type. The remaining tweets have shown that the positive type of sentiment remains over higher side i.e. about 50-55%. Bruno Lubascher, et,al [7] proposed and evaluated methods alternatively for predicting reactions to user-based posts on public pages of organizations or companies such as supermarket chains. For such a purpose, we collected various posts along with their reactions from the pages of Facebook of large supermarket chains and designed a dataset that was available for other kind of researches. For prediction of distribution reactions of distinct new post, neural network architectures were tested using a pre-trained word embedding's. Paolo Rosso, et.al [8] proposed a study based on the perspective of sentiment analysis, the presence of sarcasm and irony affected the task-based performance. The researchers has pointed out systems state-of-the-art generally providing good results they dealt with regular content, but when they evaluated with sarcastic or ironic type of content, their performance on overall basis got affected. Therefore, sentiment analysis systems (robust) needs to be understood when an individual communication in social media made the use of figurative kind of language devices such as sarcasm and irony. Preslav Nakov. [9] Expected the quest for different interesting type of formulations of general sentiment based analysis and its task to continue. We saw competitions such as those at SemEval depicted as the innovative engine as the researchers not only performed the comparisons on head-to-head basis, but also created tools and databases that enabled them to follow-up research for years afterward. Jawed Ahmed, et.al [10] proposed a study using an algorithm of Machine Learning based on Naïve Bayes performing the analysis of Sentiment. It worked well for the comments of negative type. The problems generally arose when the tweets are sarcastic or of ironic behavior, has own difficult context or reference. In order to improve the accuracy of evaluation, the researchers required something to take the references and context into consideration. They further tried to build a network i.e. LSTM network, and the results were benchmarked as compared to the NLTK machine learning implementation. Gerasimos Spanakis, et.al proposed and evaluated alternative methods for prediction these reactions to user posts on public pages of firms such as supermarket chains). For this purpose, we collected posts from Facebook pages of large supermarket chains and constructed a dataset which is available for other researches. In order to predict the distribution of reactions of a new post, neural network architectures (convolutional and recurrent neural networks) were tested using pre-trained word embedding's. Sara Rosenthal, et.al [11] described Twitter task of sentiment analysis based on the fifth year. SemEval-2017 Task 4 works continually with sub-task based re-run of SemEval-2016 Task 4 which involves the identification of tweet sentiment overall analysis, sentiment towards a topic with the classification over two-point and over five-point ordinal type scale, and the quantification of sentiment based distribution across the topic based on tweets; on two-point and five-point ordinal scale.

Table.2 Inferences Drawn

Author's Name	Year	Technology Used	Proposed Work
Onam Bharti, et.al	2016	KNN, Naïve Bayes	Proposed an approach using KNN, Naïve Bayes, and the modified version of k-means clustering, and it found that the modified version is more accurate than the KNN techniques and Naïve Bayes individually. The researchers obtained classification accuracy of 91% on overall basis over the 500 mobile review test-set.
Sara Rosenthal, et.al	2017	SemEval-2016	Discussed the Sentiment Analysis in twitter task in the fourth year. SemEval-2016 Task 4 was comprised of five of its sub-tasks. The first two forms of sub-tasks are the re-runs the data based on prior years and it was meant for predicting the overall sentiment and the tweet-based sentiment

			towards a topic.
Willian Becker, et.al	2017	Deep convolutional neural networks (DCNN)	Proposed a translation-free language-agnostic method for the analysis of Twitter sentiment which used the methodology of deep convolutional neural networks (DCNN) with embedding's of character-level for pointing the exact tweet polarity that might be written in distinct languages. The resulting model was capable of latent features learning from all of the languages that were employed during the process of training in a straight-forward manner.
Alan Ritter, et.al		Twitter demonetization	Focused to analyze the expressed sentiments on Twitter demonetization. After analysing the sentiments of results, it was observed that many of the sentiments are of neutral type. The remaining tweets have shown that the positive type of sentiment remains over higher side i.e. about 50-55%.
Sara Rosenthal, et.al	2017	SemEval-2017 Task 4 SemEval-2016 Task 4	Described Twitter task of sentiment analysis based on the fifth year. SemEval-2017 Task 4 As compared to 2016, there were two of the changes; the researchers firstly introduced new language i.e. Arabic for all types of sub-tasks, and secondly, the researchers helped to obtain the information (available) from twitter based user files. Such a task was considered to be very popular with participation of 48 teams that year.

As compared to 2016, there were two of the changes; the researchers firstly introduced new language i.e. Arabic for all types of sub-tasks, and secondly, the researchers helped to obtain the information (available) from twitter based user files. Such a task was considered to be very popular with participation of 48 teams that year. Ronald Kolcsar, et.al [12] presented an improved strategy using Twitter as a social media in order to extract the temporal or spatial patterns of the park-based visits for the purpose of urban planning, along with tweet sentiment analysis that focused over the Twitter-based frequent users. It further analyzed the spatiotemporal park based visiting behavior of more than approximately 4000 uses for almost 1700 of the parks, that examined near around 78,000 tweets in UK, London. The research novelty represents the combination of temporal and spatial aspect of twitter-based data while implementing the emotion and sentiment for the park visits in the overall city. Naresh Sharma, et.al [13] presented an approach for sentiment classification using a classifier named k-nn classifier with the use of bag of words method

taken as feature selector. The result obtained indicates that k-nn approach provides a good high form of accuracy compared to sentiment classification based polarity. The results obtained presents that k-nn classifier outperforms well for analyzing sentiments. The results have also shown that the classifier along with bag of words based feature selection outperforms well dependent over polarity based sentiment classification.

VI. Conclusion

In this study, the concept of Support Vector Machines (SVM) is used for classification of algorithm with binary classification process. Such type of method helps in analysing different feature vectors with an assigned class in order to identify the relation dependency between a sentiment and each of the feature. Here, each of the vector is considered as a point of data in vector dimensional space that equals to the size of feature-set. The SVM helps in identifying the vector dimension based hyperplane which divides the class into two types. One is the considered as "best" i.e. defined as a good type of separation gained by the hyperplane having the large distance to the point nearest to the training data type of any kind of class known as functional margin

REFERENCES

[1] Bharti, O. and Malhotra, M. 2016. Sentiment Analysis on Twitter Data. International Journal of Computer Science and Mobile Computing, 5(6): 601-609.
[2] Preslav, N. Ritter, A. Rosenthal, S. Sebastiani, F. and Stoyanov, V. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In Proceedings of the 10th international workshop on semantic evaluation (semeval-2016): 1-18.
[3] Hamid, B. and Islam, MJ. 2017. Sentiment analysis of twitter data. Department of Computer Science, Iowa State University.
[4] Joonatas, W. Becker, W. Cagnini, HEL. and Barros, RC. 2017. A character-based convolutional neural network for language-agnostic Twitter sentiment analysis. Neural Networks (IJCNN), International Joint Conference. 2384-2391.
[5] Pabreja, K. 2017. GST sentiment analysis using twitter data. International Journal of Applied Research. 3(7): 660-662.
[6] Roy, K. Kohli, D. Kumar, R. Kumar KS. Sahgal, R. and Yu, W. Sentiment Analysis of Twitter Data for Demonetization in India–A Text Mining Approach. Issues in Information Systems, 18(4): 9-15.
[7] Krebs, F. Lubascher, B. Moers, T. Schaap, P. and Spanakis, G. 2017. Social Emotion Mining Techniques for Facebook Posts Reaction Prediction. Department of Data Science and Knowledge Engineering, Maastricht University.
[8] Farias, DIH. and Rosso, P. 2017. Irony, sarcasm, and sentiment analysis. Technical University of Valencia, Spain. University of Turin, Italy, 113-128.
[9] Nakov, P. 2017. Semantic Sentiment Analysis of Twitter Data. Qatar Computing Research Institue, HBKU, Doha, Qatar
[10] Mohammad, S and Ahmed, J. 2017. Sentiment Analysis and Classification of Tweets Using Data Mining. International Research Journal of Engineering and Technology 4(12): 1471-1474
[11] Rosenthal, S. Farra, N. and Nakov, P. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017): 502-518.
[12] Anna, KG. Ristea, A. Kolcsar, R. Resch, B. Crivellari, A. and Blaschke, T. Beyond Spatial Proximity—Classifying Parks and Their Visitors in London Based on Spatiotemporal and Sentiment Analysis of Twitter Data. ISPRS International Journal of Geo-Information, 7(9):378.
[13] Tyagi, A. and Sharma, N. 2018. Sentiments Analysis of Twitter Data using K-Nearest Neighbour Classifier. International Journal of Engineering Science 8(4):17258-17260.