International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

# Sentiment Analysis of Top Colleges in India Using Twitter Data

**Pallavi. S, Ramya K.V, Rachana C, Vidyashree, Gangadhar Immadi**

ISE Department, New Horizon College of Engineering, Bengaluru, Karnataka, India

## ABSTRACT

Sentiment analysis is used for identifying and classifying opinions or sentiments expressed in source text. Social media is generating a vast amount of sentiment rich data in the form of tweets, status updates, blog posts etc. Sentiment analysis of this user generated data is very useful in knowing the opinion of the crowd [6]. Due to the presence of slang words and misspellings, twitter sentiment analysis is difficult compared to general sentiment analysis. Sentiments from the source text will be analyzed by using a machine learning approach. Mining opinions and analyzing sentiments from social network data which will help in several fields such as even prediction, analyzing overall mood of public on a particular social issue. The accuracy of classification can be increased by using Natural Language Processing (NLP) Techniques. We present a new feature vector for classifying the tweets as positive, negative, neutral and undefined. The mined text information is subjected to Ensemble classification to analyze the sentiment. Ensemble classification involves combining the effect of various independent classifiers on a particular classification problem [1]. Multi-Layer Perceptron (MLP) is used to classify the features extracted from the reviews. A Decision Tree-based Feature Ranking is used for feature selection. Based on Manhattan Hierarchical Cluster Criterion the ranking will be done [5].

**Keywords :** Sentiment Analysis, Machine Learning, Opinion Mining, Natural Language Processing Twitter, Multilayer perceptron(MLP).

## I. INTRODUCTION

The evolution of Internet has changed the way people express their views. It is now done through blog posts, online discussion forums, product review websites etc. This user generated content (review) is used in a large extent by the people. Online reviews play a major role for taking the decision before buying any product. The data generated in a large extent and these data will be difficult to analyze for a normal user. Various sentiment analysis techniques will be used to automate this. The two main techniques used in sentiment analysis are Symbolic techniques or Knowledge base approach and Machine learning techniques. Knowledge base approach requires a large database of predefined emotions and an efficient knowledge representation for identifying sentiments. Machine learning approach makes use of a training set to develop a sentiment classifier that classifies sentiments. For machine learning approach a predefined database of entire emotions is not required, So it is simpler than Knowledge base approach [6]. For classifying the tweets, we use different machine learning techniques. Sentiment Analysis in twitter is quite difficult due to its short length. Presence of emoticons, slang words and misspellings in tweets forced to have a pre-processing step before feature extraction. There are different feature extraction methods for collecting relevant features from text which can be applied to tweets also. But the feature extraction can be done in two phases to extract relevant features [4]. In the first phase, twitter specific features are extracted. Then these features are removed from the tweets to create normal text. After that, to get more features again feature extraction is done. This method is used to generate an efficient feature vector for analyzing twitter sentiment. Since no standard dataset is available for twitter posts of electronic devices, so we created a dataset by collecting tweets for a certain period of time [6].

Sentiment analysis (SA) refers to identifying and extracting subjective information from natural language text(source text). The problem of automatic sentiment analysis has received significant attention in recent years, largely due to the explosion of online social-oriented content (e.g., user reviews, blogs, etc)[4].

Opinion Mining (OM) identifies author’s viewpoint on a subject instead of identifying subject itself. OM’s ultimate goal is to extract customer opinions on products and to present it in an effective way to serve certain objectives. Based on presentation of the summarized information The steps and techniques will differ. For a product the negative and positive reviews will be provided, and classifying each review based on its polarity (positive/negative) is required. But, if we were to show customer feedback on a product’s features, it is necessary to extract product features and analyze each feature’s sentiment[5] .

## II. RELATED WORK

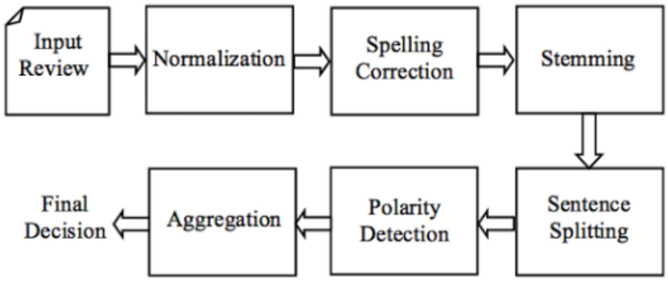
**I PREPROCESSING TECHNIQUE**:

- Data Cleaning: This method is used to remove the repeated letters and words. Each and every keyword in the sentence, Word Sense tagging is done. URLs, hashed words, names are removed from the tweets. The cleaned tweets now undergo Parts Of Speech tagging[1].
- Synset Finding: This method is used To capture the semantic similarities among the tweets .For this we use the synsets of WordNet. Synset contains the set of words that are semantically related to the word of interest. For every key word in tweet the synset of the word is retrieved from the WordNet database. By synset accuracy of classification is increased by covering all the semantically related data items. Before finding synset the original words in the tweets are stemmed up to the root word by user Stemmer. Stemming reduces the feature vector size while preserving the key terms[1].
- Feature Vector Formation: After synset findings the data is subjected to feature vector formation. The feature vector consists of key terms of the tweets along with the synset words. The feature vectors thus obtained are subjected to classification by traditional classifiers and Ensemble classifiers and the results are presented to user stating the sentiment polarity of the public on the topic[1].

## III. SPELLING CORRECTION

Many a times, people may unknowingly misspell the words which will complete change the meaning of the sentence. The ever-present approach of replacing more than two occurrences of a letter with two occurrences of the same letter is not a complete solution. Misspellings may occur from the user’s finger slipping to a nearby letter or the user’s spelling the word phonetically. By using probabilistic model based on Baye’s theorem the words will be corrected to the best possible effort which shows a better accuracy[2].

Further, stop-words which are the common words in the English language and do not contribute towards the sentiment of a sentence will be removed with referral from a corpus of stop-words and also from the dictionary meant to test for ambiguous[2].



## IV. LATENT TOPIC/CONCEPT MODELS :

There are number of word-level embedding methods that are able to capture semantic similarity between word pairs. One of the earliest but widely used approaches is Latent Semantic Index- ing (LSI). LSI applies SVD(singular value decomposition) to term-document co-occurrence matrix. It produces a low-dimensional representation for both doc- uments and words, and enables efficient computation of semantic similarity between them. LSI is considered the pioneering work that inspired methods such as probabilistic LSI (pLSI) and Latent Dirichlet Allocation (LDA) using a generative probabilistic framework[4] . There are also supervised variants that try to compute embedding by fitting to the labeled data. Such approaches have been applied to a variety of information retrieval tasks such as link prediction, cross-lingual retrieval, and image annotations . [4]

### 4) N-GRAM:

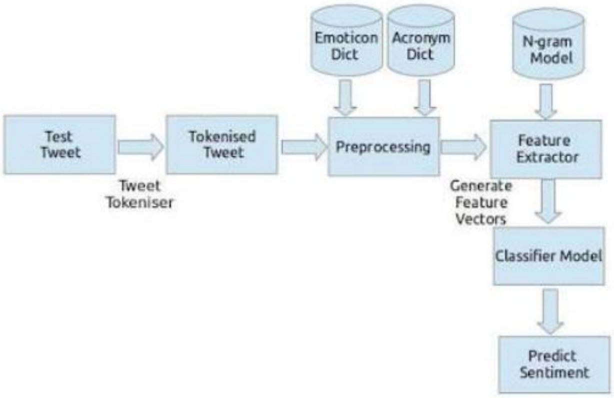
An n-gram is a contiguous sequence of n items from a given sequence of text or speech. The phonemes, syllables, letters, words or base pairs are considered as items according to the application. The n-grams typically are collected from a text or speech corpus. An n-gram model is a type of probabilistic language model. This model is used for predicting the next item in such a sequence in the form of a (n − 1)–order Markov model. An n-gram model models sequences, notably natural languages, using the statistical properties of n-grams[4].

In a simple *n*-gram language model, the probability of a word, conditioned on some number of previous words (one word in a bigram model, two words in a trigram model, etc.) can be described as following a categorical distribution (often imprecisely called a "multinomial distribution"). In practice, the probability distributions can be smoothed by assigning non-zero probabilities to unseen words or *n*-grams; see smoothing techniques[4].

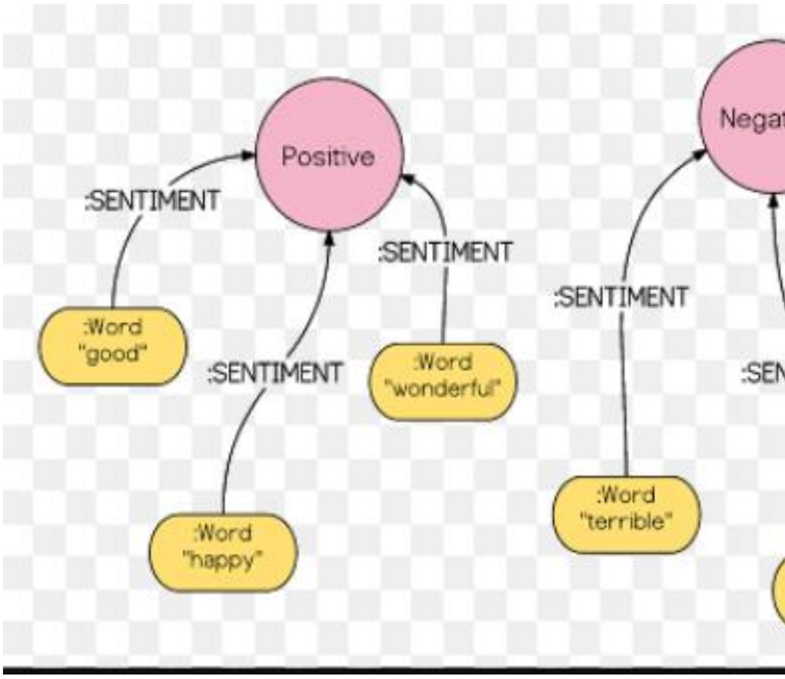
### DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING:

Lately, “deep learning” research grows to bring in attentions. From recent results it is possible to learn the kind of complicated functions that can represent high level abstractions and one would need deep architectures. Each layer in the architecture represents features at a different level of abstraction, defined as a composition of lower-level features. Statistical language modelling is a key topic in natural language processing (NLP), where the difficulty is the curse of dimensionality, especially when modeling joint distribution between many discrete random variables[4]. Language model which is based on the multi-layered neural networks, that tries to model a distributed representation for each and every word and the probability function for word sequences, simultaneously. Later by utilizing a single multi-layered convolutional neural network architecture to handle multiple classic NLP tasks at the same time[4]. The framework provides an end-to-end system that, given a sentence, outputs a host of language processing predictions. This method is motivated by the above approach and uses a multi-layer “deep” neural network[4].

### VI SENTIMENT ANALYSIS:



Sentiment analysis (SA) means to identify and extract subjective information from natural language text. In recent years, The problem of automatic sentiment analysis has received significant attention, largely due to the explosion of online social-oriented content (e.g., user reviews, blogs, etc). Latent semantic analysis has been used in to to calculate the semantic orientation of the extracted words according to their co-occurrences with the seed words, such as “excellent” and “poor”. The polarity of the arti- cle is then determined through averaging the sentimental orientation of its corresponding words. Instead of limiting the sentiment analysis at the word level, the mainstream research community per- forms sentiment classification at the article level. Different methods are based on this principle have been proposed. These methods can be contrasted in terms of features they use: utilizing either uni- gram features and/or filtered bigrams.



## V. MULTILAYER PERCEPTRON

Neural Networks (NN) are parallel computing systems and it consists of large number of simple processors with interconnections. NN models use organizational principles in a weighted and directed graphs network. Where nodes are artificial neurons and directed edges connections between neuron outputs and inputs. NN contains many interconnected processing elements which operate simultaneously. Pattern recognition data processing is bulky and recognition in conventional NN is slow as propagation takes place in multiplication and addition calculation required for data processing[5].

A Multilayer Perceptron (MLP) is a feed forward Artificial Neural Network (ANN) model that maps input data sets onto appropriate output sets. An MLP has many node layers in a directed graph, each layer being connected to the next. Other than the input layer each node is a neuron (processing element) with nonlinear activation function in layers. Networks are trained using Supervised learning technique. This learning technique is called back propagation. MLP is modified as a standard linear perceptron that differentiates data not linearly separable.

## VI. PROPOSED DECISION TREE-BASED FEATURE RANKING

The decision trees are used as embedded method of feature selection. In this decision tree-based feature ranking, a Decision tree induction will select relevant features and ranks these features. Decision tree induction is decision tree classifiers learning, which will construct a tree structure with internal nodes (non-leaf node) denoting an attribute test. Each branch represents test outcome and external node (leaf node) denotes class prediction. The algorithm at each node will always choose the best attribute to partition data into individual classes. Information gain measure is used to choose the best partitioning attribute by attribute selection. Attribute with highest information gain splits the attribute[5].

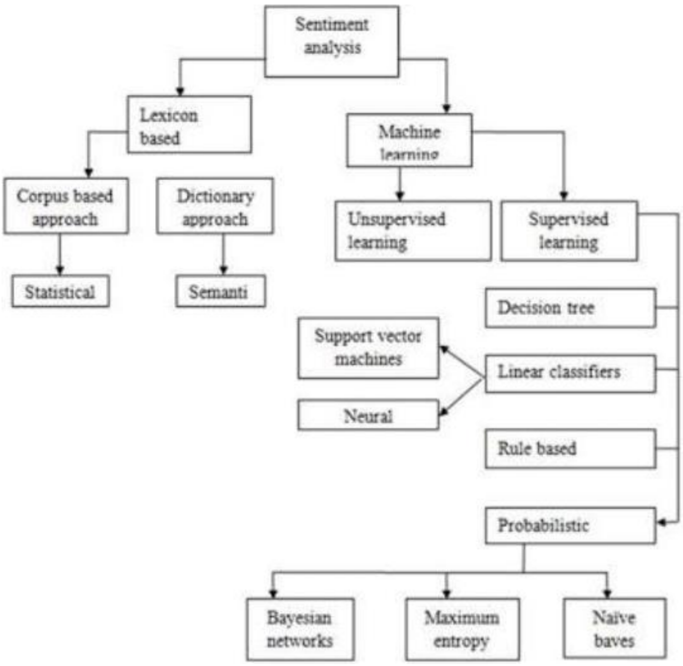
Before constructing trees, following base cases are considered :

- A leaf node is created if all samples belong to same class.
- When no features provide information gain, it creates a decision node higher up the tree using the expected class value[5].

## VII. MACHINE LEARNING TECHNIQUES:

Training set and a test set are used for classification in MLP. Input feature vectors and their corresponding class labels are present in the training set. By using this training set, a classification model has been developed which tries to classify the input feature vectors into corresponding class labels. By predicting the class labels of unseen feature vectors the model is validated by using test set. There are various number of machine learning techniques like Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify reviews. Term Presence, Term Frequency, negation, n-grams and Part-of-Speech are some of the features that can be used for sentiment classification. These features are used to find out the semantic orientation of words, phrases, sentences and documents[6].

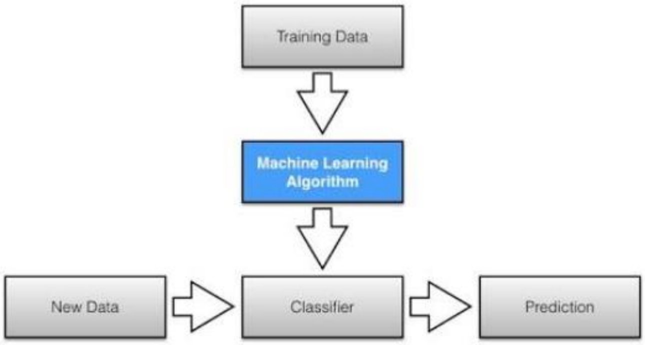
Semantic orientation is the polarity which may be either positive or negative. Highly dependent features work well with Naive Bayes[6]. This is surprising because the basic assumption of Naive Bayes is that the features are independent[6].



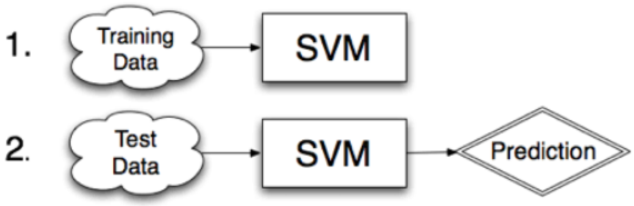
## VIII. NAIVE BAYES CLASSIFIER

In machine learning, Naive Baye’s classifier are a family of simple probabilistic classifier that is based on Baye’s Theorem with a strong indepdence assumption between the features that is presence of a particular feature in a class which is unrelated to the presence of any other feature.

Sentiment analysis using Naive Bayes classifier is based on bag-of-words model. Using the bag-of-words model we can check which particular word of the text document belongs to positive word list or negative word list. If the word belongs to positive word list then the total score of the text is updated with +1 and vice versa. At the end , if the total score of positive is more then the total score of the negative then the text is classified as positive or vice versa.

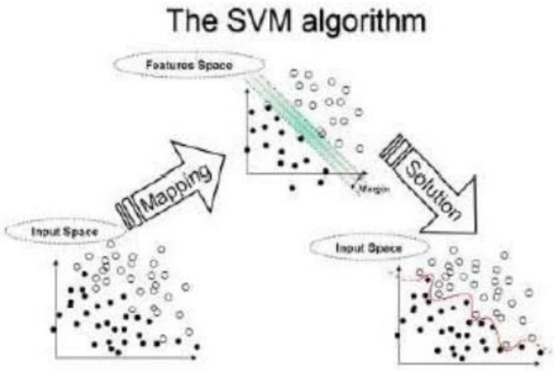


## IX. SUPPORT VECTOR MACHINE(SVM):



SVM is perhaps one of the most popular and talked about machine learning algorithm. SVM are supervised learning models that consists associated learning algorithms that will analyze the data used for classification and regression analysis.

Give a set of training examples, SVM will assign each example to one or the other of two categories. SVM training algorithm will help to build a model that assigns newly encountered words to one category or the other, making it a non-probabilistic binary linear classifier.



## X. CONCLUSION

There are different machine learning and symbolic techniques to identify sentiments from text. It has been observed that Machine learning techniques are efficient than symbolic techniques. Twitter sentiment analysis can be done by using these techniques .An efficient feature vector is created by doing feature extraction to deal with misspellings and slang words. Machine learning algorithms like Naive Bayes and support vector machine(SVM) and Artificial Neural Network(ANN) model like Multilayer Perceptron yield promisingly accurate predictions on unseen data.

## XI. REFERENCES

[1]. Kanakaraj M, Guddeti R M.R.performance analysis of ensemble methods on twitter sentiment analysis using NLP techniques published by IEEE in the year 2015 .

[2]. Bespalov D, Bai B, Qi Y. Performance analysis of ensemble methods on twitter sentiment analysis using NLP techniques published by IEEE in the year 2011.

[3]. Gaurav Bhat , Ankush mittal.Sentiment analysis of top colleges in india using twitter data published by IEEE in the year 2016.

[4]. Bahrainian S.A, Dengel A. Sentiment classification based on supervised latent n-gram analysis published in the year 2013.

[5]. Jeevanandam Jotheeswarn, S.Koteeswaran.Decision tree based feature selection and multi layer perceptron for sentiment analysis published by arpnjournal of engineering and applied sciences in the year 2015.

[6]. Rajasree R, Neethu M.S. Sentiment Analysis in Twitter using Machine Learning Techniques published by IEEE in the year 2013.

[7]. Z. Niu, Z. Yin, and X. Kong, “Sentiment classification for microblog by machine learning,” in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286–289, IEEE, 2012.

[8]. B. Pang and L. Lee, Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135, 2008.

[9]. Gayatri N, Nickolas S. and Reddy A, V. 2010. Feature selection using decision tree induction in class level metrics dataset for software defect predictions. In Proceedings of the World congress on engineering and computer science.