. A review on Sentiment Analysis Using Machine Learning Techniques

S.Biron Gifty[1], Ancy John J[2]
Department of Computer Science and Engineering,
1.Bethlahem Institute of Engineering
2.Arunachala College of Engineering for Women

Abstract

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral.. A sentiment analysis system for text analysis combines natural language processing and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.Sentiment analysis helps data analysts within large enterprises gauge public opinion, conduct nuanced market research, monitor brand and product reputation, and understand customer experiences. The best businesses understand the sentiment of their customers—what people are saying, how they're saying it, and what they mean. Customer sentiment can be found in tweets, comments, reviews, or other places where people mention your brand. Sentiment Analysis is the domain of understanding these emotions with software, and it's a must-understand for developers and business leaders in a modern workplace.In the proposed work various machine learning algorithms for sentiment analysis was compared to find the accurate method.This minimal improvement in the accuracy is expected to get improved when applied to a larger corpus of big data where it will show its significance. Key Words: SVM, Machine learning, Naïve Bayes, sentiment analysis.

1.Introduction

Sentiment analysis is the process of using natural language processing, text analysis, and statistics to analyze customer sentiment. Sentiment analysis also known as Opinion Mining is an interesting way to find the opinions of a user and to effectively categorize then to be positive, negative or neutral. Now-a-days sentiment analysis has shown its significance in almost all the fields of media. Natural language processing is deeply tied with Sentiment analysis. When a user expresses his views, it is important for the organization to correctly identify the requirements of the user to make him stay longer as their customer. For that a deep understanding of their customer's opinion[1][3] is important. By the analysis of product reviews by the customer, it is easier for the company to decide about the future of that product. In the same way, it is very important to analyze the comments given in social media.[1] Twitter Analytics has become a separate field by itself, where even studies show the impact of tweets over the sensitive fields like [9] market prediction.The words in a Sentiment analysis is classified on the basis of semantic orientation (SO), that is the word is basically classified using its weight, polarity, and its strength. Semantic Orientation is extremely helpful in determining marketing reviews, compiling reviews etc. In general semantic orientation always refers to the strength of the words, phrases or texts in addition to the sentiment analysis which is the main goal of our process[16]. Semantic Orientation involves adjectives, phrases, words, texts, adverbs, verbs and noun. In the case of our project, we are going to perform the following steps:

Convert tweets to lowercase using .lower() function, in order to bring all tweets to a consistent form. By performing this, we can assure that further transformations and classification tasks will not suffer from non-consistency or case sensitive issues in our data.

Remove 'RT', UserMentions and links: In the tweet text, we can usually see that every sentence contains a reference that is is a retweet ('RT'), a User mention or a URL. Because it is repeated through a lot of tweets and it doesn't give us any useful information about sentiment, we can remove them.

Remove numbers: Likewise, numbers do not contain any sentiment, so it is also common practice to remove them from the tweet text.

Remove punctuation marks and special characters: Because this will generate tokens with a high frequency that will cloud our analysis, it is important to remove them.

Replace elongated words: an elongated word is defined as a word that contains a repeating character more than two times, for example, 'Awesoooome'. Replacing those words is very important since the classifier will treat them as different words from the source words lowering their frequency. Though, there are some English words 1 that contain repeated characters, mostly consonants, so we will use the wordnet from NLTK to compare to the English lexicon.

Removing stopwords: Stopwords are function words that are high frequently present across all tweets. There is no need for analyzing them because they do not provide useful information. We can obtain a list of these words from NLTK stopwords function.

Handling negation with antonyms: One of the problems that come out when analyzing sentiment is handling negation and its effect on subsequent words. Let's take an example: Say that we find the tweet "I didn't like the movie" and we discard the stopwords, we will get rid of "I" and "didn't" words. So finally, we will get the tokens "like" and "movie", which is the opposite sense that the original tweet had.

In machine learning, there is a term called Classification which falls under the category of Supervised Learning (i.e It requires a training set). Classification is nothing but simply identifying which object falls into which category. And there are many approaches for classification like Naive Bayes Classifier, Neural Network Classifier, Nearest Neighbor Classifier, Support Vector Machines, etc.Sentiment analysis is simply an application of classification. We have our pre-defined sentiments like positive, negative, etc. So, a machine learning model simply classifies the enter text into our pre-defined categories.

2.Algorithms used in Sentiment Analysis

Machine learning algorithms play an important role in sentiment analysis. Specifically speaking, lots of works in sentimental analysis uses classification algorithms like Support Vector Machine (SVM),LSTM, Naïve Baise Classifier, Maximum Entropy classifier, Decision tree, KNN (K-Nearest Neighbor) to detect positive, negative or neutral sentiments.

A. Support Vector Machine

Support Vector Machine SVM is a discriminative classifier considered as the best text classification method. what a SVM does is finding the best hyperplane that separates the data points of two different classes. For sake of simplicity and because we are creating a sentiment tool, we call these classes 'positive' and 'negative', and denote them by respectively and . Both sides of the plane represent a different class and this plane can thus be seen as a decision boundary for any new point, because we can easily classify this new point based on which side of the plane it lies. Therefore, after a decision boundary is obtained, a SVM is very useful for making accurate predictions of new data points. The entire process of constructing a SVM can thus be divided into two parts. In the first part, we train a 'machine' by providing it a classified dataset. For the sentiment analysis tool, this data consisted of news articles which were labeled as one of six possible sentiments. Recall that, to keep things simple, we only consider the positive and negative sentiment for this practical report. The training is done by finding the best hyperplane that separates the different classes. Training the machine thus implies finding the best values.In the second part, we are now able to use the trained machine to make predictions about the classification of any data point. For sake of simplicity, we state that positive points lie on the side of the plane for which it holds that and vice versa for negative points. Moreover, we define the variable , which denotes the sentiment classification of message . Using this notation we can state that if data point lies on the correct side of the plane. Note that this can be replaced by , because and can be scaled up such that this equation holds. It is obvious that, once the best hyperplane has been found, it is fairly easy to make predictions, based on which side of the plane the data point is located. The difficult and tricky part is therefore how to find this best separating hyperplane, that is, how to find the values that uniquely determines this hyperplane.

B.LSTM

Next we are discussing about Long Short-Term Memory Recurrent Neural Network model. It has a very interesting architecture to process natural language. It works exactly as we do. It reads the sentence from the first word to the last one. And it tries to figure out the sentiment after each step. For example, for the sentence "The food sucks, the wine was worse.". It will read "The", then "food", then "sucks", "the" and "wine". It will keep in mind both a vector that represents what came before (memory) and a partial output. For instance, it will already think that the sentence is negative halfway through. Then it will continue to update as it processes more data.

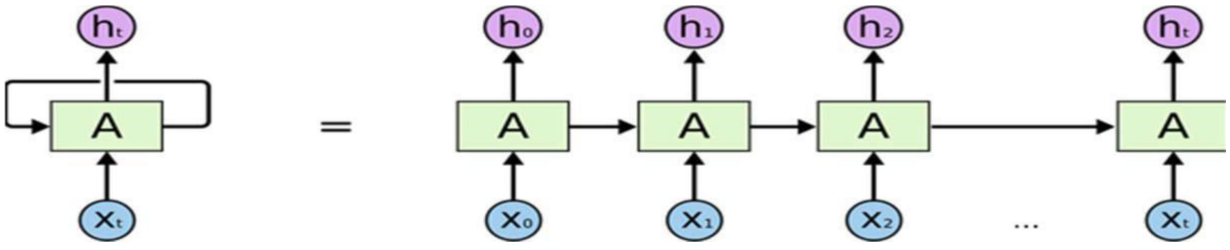

Fig :1 Long short-term memory (LSTM) networks

This is the general idea, but the implementation of these networks is much more complex because it is easy to keep recent information in mind, but very difficult to have a model that captures most of the useful long-term dependencies while avoiding the problems linked to vanishing gradient.This RNN structure looks very accurate for sentiment analysis tasks. It performs well for speech recognition and for translation. However, it slows down the evaluation process considerably and doesn't improve accuracy that much in our application so should be implemented with care.

C.Naïve Bay's Classifier

We know that the Naive Bayes Classifier[21][22] is based on the bag-of-words model.With the bag-of-words model we check which word of the text-document appears in a positive-words-list or a negative-words-list. If the word appears in a positive-words-list the total score of the text is updated with +1 and vice versa. If at the end the total score is positive, the text is classified as positive and if it is negative, the text is classified as negative. Simple enough. With the Naive Bayes model, we do not take only a small set of positive and negative words into account, but all words the NB Classifier was trained with, i.e. all words presents in the training set. If a word has not appeared in the training set, we have no data available and apply Laplacian smoothing.

D. K Nearest Neighbor (KNN)

K-NN[6] is the simplest of all machine learning algorithms. The principle behind this method is to find a predefined number of training samples closest in distance to the new point and predict the label from these. The number of samples can be a user-defined constant or vary based on the local density of points. The distance can be any metric measure. Standard Euclidean distance is the most common choice for calculating the distance between two points. The Nearest Neighbours have been successful in a large number of classification and regression problems, including handwritten digits or satellite.

image processing and so on.

E.Random Forest

Random Forests[21] are the learning method for classification and regression. It construct a number of decision trees at training time. To classify new case it sends the new case to each of the trees. Each tree perform classification and output a class. The output class is chosen based on majority voting that is the maximum number of similar class generated by various trees is considered as the output of the Random Forest. Random Forests are easy to learn and use for both professionals and laypeople with little research and programming required. It can easily be used by persons that don't have a strong statistical background

F.Maximum Entropy classifier

The Max Entropy classifier[21] is a probabilistic classifier which belongs to the class of exponential models. The MaxEnt is based on the Principle of Maximum Entropy and from all the models that fit our training data, selects the one which has the largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.

3. Existing Method

The data extraction process is generally performed within the source system. It's common to perform data extraction using one of the following methods: (1)Full extraction. Data is completely extracted from the source, and there is no need to track changes. The logic is simpler, but the system load is greater.(2)Incremental extraction. Changes in the source data are tracked since the last successful extraction so that you do not go through the process of extracting all the data each time there is a change.Feature extraction starts from an initial set of measured data and builds derived values intended to be informative and non-redundant, facilitating the subsequent learning and generalization step. The extracted features are classified using various classification techniques and the evaluated.In this method the accuracy is less.

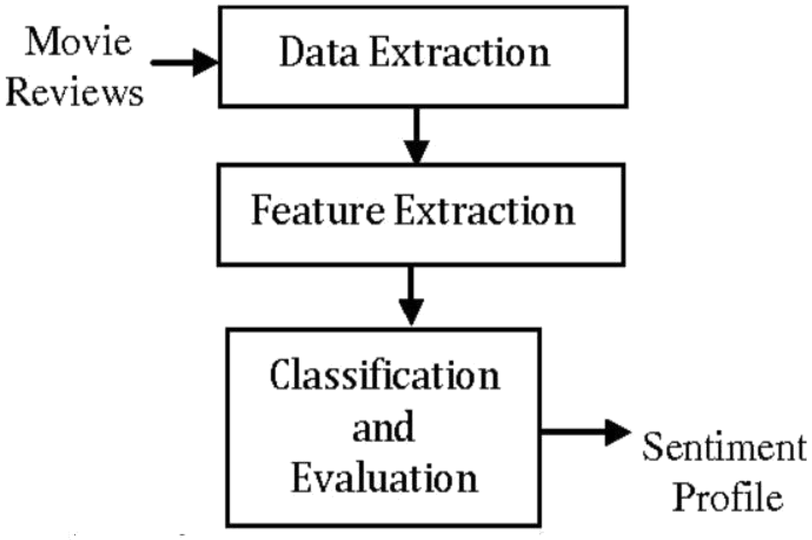

Figure 2: Existing sentiment analysis

4. Proposed Method

There are various supervised learning algorithms such as knn, svm , naïve baye's classifier, maximun entropy classifier, etc. As a deviation from the existing work,experiment carried out from the comparison of knn, svm , naïve baye's classifier, maximun entropy classifier was done. The output of each classification algorithm is taken and a comparison was done. Input tweets are given in the form of keywords. The tweets are retrievedfrom archive or current database.Then its classified using various machine learning algorithms. The classified tweets specifies whether it is positive or negative.The the outputs of all used algorithms are compared to determine its accuracy.

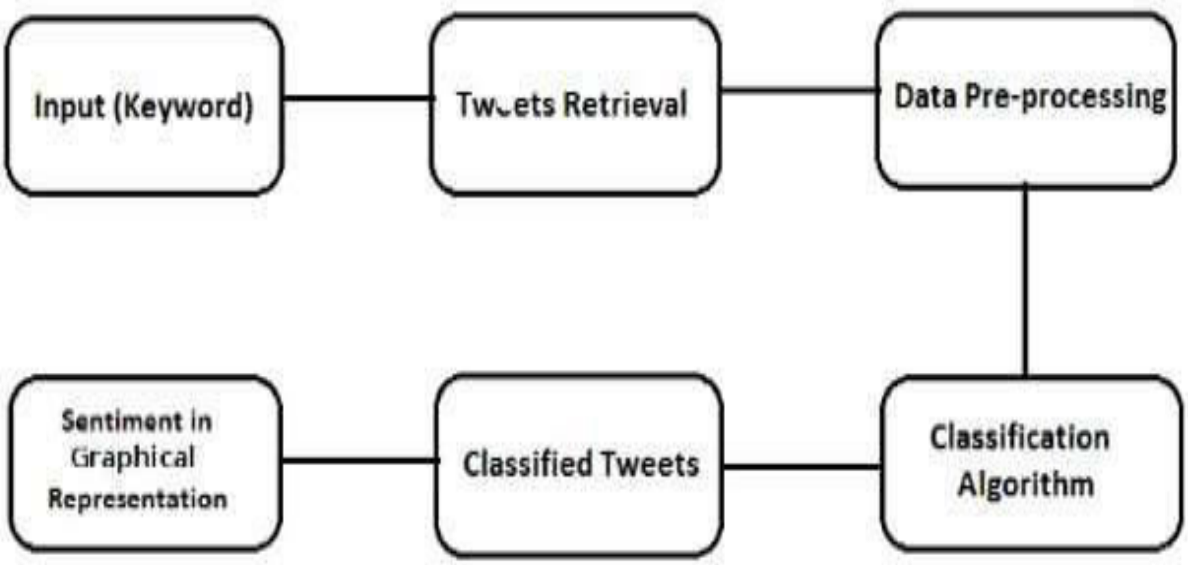

Figure 3:Proposed method

5.Experiments and Results

In this paper, a comparative study of supervised learning, was done to find the accurate method for sentiment analysis.. This is evaluated using the metrics accuracy and precision. As this is supervised learning, accuracy is highly superior for smaller datasets. As all the positive and negative sentiments are trained, learned and labeled. It correctly classifies the reviews as positive and negative sentiments with less error rate.[15]

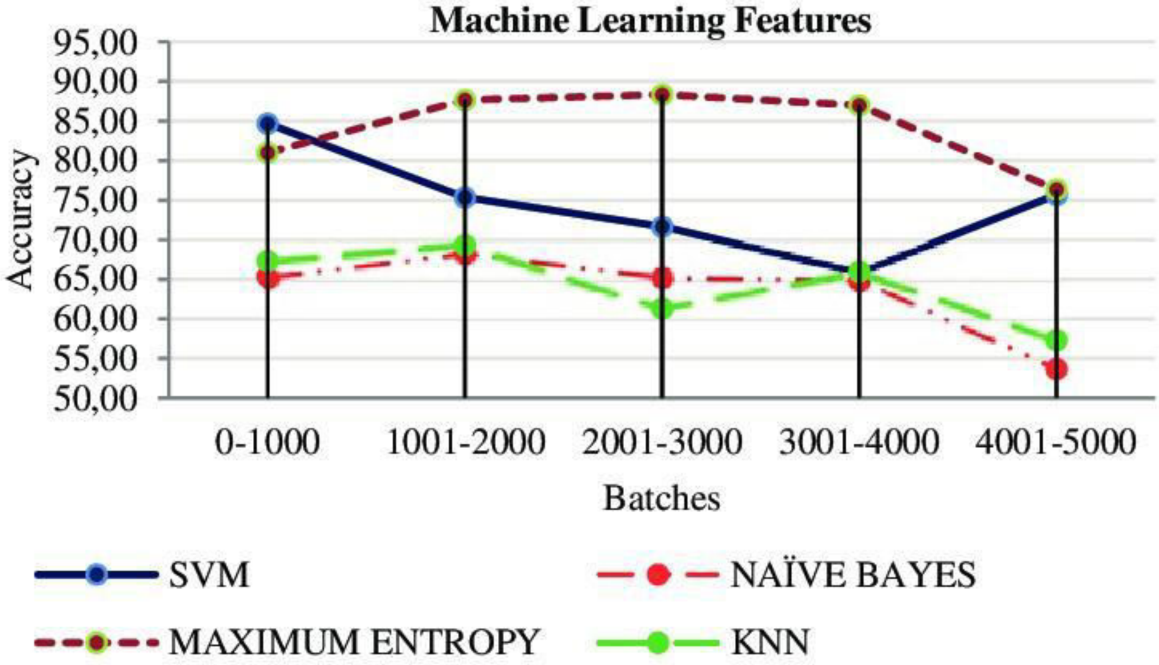

Figure 4: Sentiment analysis for different movies

6. Conclusion

Sentiment Analysis is very essential in our daily routine. It has its diverse specification in the areas of social media such as analysis of twitter data, through the help of Support Vector Machine. Through Sentimental Analysis marketing strategy, campaign success, improving product messaging and other areas.. Sentiment Analysis has been effective in all its cases in which it has been implemented. Filters like CNN, using deep learning techniques, is also used as a part of Sentimental Analysis. All these factors make an impact in the difference of learning, in order to increase the proposed work.. The factors which can also be applicable for larger datasets, which improves the efficiency and accuracy. In future big data analysis technique can be used to classify all emotions for large volume of tweets.

References

[1]. Varghese R., Jayasree M., A survey on sentiment analysis and opinion mining, International Journal of Research in Engineering and Technology 2(11) (2013), 312-317.

[2]. Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R, Sentiment analysis of twitter data, Proceedings of the workshop on languages in social media, Association for Computational Linguistics (2011), 30-38.

Vinita Sharma, Literature Survey (2014).

[3]. Sahayak V., Shete V., Pathan A, Sentiment Analysis on Twitter Data, International Journal of Innovative Research in Advanced Engineering (IJIRAE) 2(1) (2015), 178-183.

[4]. Singh R., Kaur, R, Sentiment Analysis on Social Media and Online Review, International Journal of Computer Applications 121(20) (2015).

[5]. Medhat W., Hassan A., Korashy H., Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5(4) (2014), 1093-1113.

[6]. Sources from Wikipedia, Kernel Methods.

[7]. Sindhwani V., Melville P., Document-word co-regularization for semi-supervised sentiment analysis, Eighth IEEE International Conference on Data Mining (2008), 1025-1030.

[8]. Nair B.B., Mohandas V.P., Sakthivel N.R., A genetic algorithm optimized decision tree-SVM based stock market trend prediction system, International Journal on Computer Science and Engineering 2(9) (2010), 2981-2988.

[9]. Nanli Z., Ping Z., Weiguo L., Meng C., Sentiment analysis: A literature review, International Symposium on Management of Technology (ISMOT) (2012), 572-576.

[10]. Taboada M., Brooke J., Tofiloski M., Voll K., Stede, M, Lexicon- based methods for sentiment analysis, Computational linguistics 37(2) (2011), 267-307.

[11]. Vaitheeswaran G., Arockiam, L, A Novel Lexicon Based Approach to Enhance the Accuracy of Sentiment Analysis on Big Data, International Journal of Emerging Research in Management and Technology (IJERMT) 5(2) (2016).

[12]. Sivakumar P.B., Mohandas V.P., Sobh T, Evaluating the predictability of financial time series, A case study on SENSEX data, Innovations and Advanced Techniques in Computer and Information Sciences and Engineering (2007), 99–104.

[13]. Padmavathi S., Rajalaxmi C., Soman K.P, Texel identification using K-Means clustering method, Advances in Computer Science, Engineering & Applications (2012), 285-294.

[14].Abarna K., Rajamani M., Vasudevan S.K, Big data analytics: A detailed gaze and a technical review, International Journal of Applied Engineering Research 9(9) (2014).

[15]. Geethan P., Jithin P., Naveen T., Padminy K.V., Shruthi Krithika J., Vasudevan S.K, Augmented reality X-ray vision with gesture interaction, Indian Journal of Science and Technology 8 (2015), 43-47.

[16]. Sankar A., Suresh A., Varun Babu P., Baskar A., Vasudevan S.K, An in-depth analysis of applications of object recognition, Research Journal of Applied Sciences, Engineering and Technology 10(1) (2015), 1-14.

[17]. Rajendran A., Kiran M.V.K., Vasudevan S.K., Baskar A, An exhaustive survey on human computer interaction's past, present and future, International Journal of Applied Engineering Research 10(2) (2015), 5091-5105.

[18]. Gaurangi Patil, Varsha Galande, Vedant Kekan, Kalpana Dange, Sentiment Analysis Using Support Vector Machine, International Journal of Innovative Research in Computer and Communication Engineering 2(1), (2014).

[19].Yong Yang, Chun Xu, Ge Ren, Sentiment Analysis of Text Using SVM, Electrical, Information Engineering and Mechatronics of the series Lecture Notes in Electrical Engineering 138 (2012), 1133- 1139.

[20]. Jaspreet Singh, Gurvinder Singh & Rajinder Singh ,Optimization of sentiment analysis using machine

learning classifiers,Human-centric Computing and Information Sciences volume 7, Article number: 32 (2017)

[21]. Munir Ahmad, Shabib Aftab, Syed Shah Muhammad and Sarfraz Ahmad, Machine Learning Techniques for Sentiment Analysis: A Review , International Journal Of Multidisciplinary Sciences And Engineering, Vol.
8, No. 3, ApriL 2017 [issn: 2045-7057

[22]. Dipak Kawade, Kavita Oza, Sentiment Analysis: Machine Learning Approach Article in International Journal of Engineering and Technology 9(3):2183-2186 · June 2017 DOI: 10.21817/ijet/2017/v9i3/1709030151