Twitter Sentiment Analysis of Mobile Reviews using kernelized SVM

Driyani A [a], J.L. Walter Jeyakumar[b]

[a]Research Scholar, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu
[b]Associate Professor, Dept. of Computer Science, St.Xavier's College, Tirunelveli, Tamilnadu
Email:[a] driyanirajesh@yahoo.in, [b] walterjeya@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Sentiment analysis is a technique of analysing the opinions commented in social media on various topics. There are few ways the sentiment analysis can be done. Machine learning plays a crucial role for analysing the opinions and reviews. Mobile related tweets have been scraped from twitter. Noise on the tweets has been removed using pre-processing and feature vectors were created. Support Vector Machine has been used to classify the reviews as either positive or negative. 4 cross validation technique were used to bring out the better accuracy. 3 different sizes of dataset with iphone11 reviews have been used for training and testing with different kernels of SVM. RBF kernel is found to be working better for classifying the tweets but at the same time it has been found that the accuracy decreases when the data grow bigger.

Keywords: Social Media, SVM, Kernels and Cross validation

1. Introduction

There are various social media apps like twitter, Facebook, Instagram etc. and e commerce sites like flipkart, amazon which contain huge amount of opinions and reviews about different products which people use in day today life. These sentiments will be very useful to the companies when the sentiments give some meaningful information about the status of their product or brand in the market. Companies will use those reviews to upgrade their product to a better version. Social media users use those sentiments to know about different opinions about products and do purchase accordingly.

Machine learning algorithms and lexicon based approach can be used for sentiment analysis. Lexicon based approaches are too much dependent on the dictionaries. Machine learning approach has been proved to be better than the dictionary based. In machine learning, supervised algorithms are showing better classification results comparing to unsupervised or semi supervised algorithms.

Support vector machine is a supervised machine learning algorithm which proves to be a better one for doing sentiment analysis than the other algorithms. Live tweets have been scraped, pre-processed and then classified using SVM either as positive or negative. Different kernel functions of SVM also have been used to know which kernel works better for classifying tweets

2. Literature Survey

This paper [1] talks about aspect based sentiment classification problem. Hat crime twitter sentiment and benchmark Stanford twitter sentiment datasets were used for training and testing. The combination between lexicon-based method using Sentiwordnet and PCA feature selection method helped to improve sentiment classification accuracy. Feature selection was done by PCA thus achieved good classification accuracy using SVM. SVM [2] has been used for learning and classifying the dataset. Benchmark datasets Pang Corpus and Taboada Corpus were used for training the classifier. The main aim of this work is to apply different n gram techniques and compare the influence of using different n grams. Three different weighting schemes TFIDF, Binary Occurrence (BO) and Term Occurrence (TO) were used to generate the word vectors. SVM achieved higher accuracy when the number of features selected is fewer using CHI2 method and found that unigram model outperforms all n gram schemes. This paper [3] used improved SVM for classification. The hyper parameters C and gamma of RBF kernel is modified to yield the better accuracy. Twitter dataset and Gold set from Amazon were used as the dataset. POS tagging has been used to tag each word with its appropriate POS. Stanford tagger was used to tag the words in the sentences. SentiwordNet dictionary has been used to calculate the score of tagged words and SVM is used for training and classification. Optimal value has been found for the hyper parameters thus the improved SVM gives good accuracy comparing to the existing one.

Collection of reviews [4] of different laptop companies like HP, APPLE, DELL, LENOVO etc. have been taken from e-commerce sites like Amazon, eBay, flipkart. Objectivity and subjectivity of sentences have been found using POS tagging. SentiwordNet has been used to remove not opinionated sentences from the dataset. Feature which is specified high times in the reviews will be considered to improve the accuracy of the SVM classifier. This paper proposed a novel way of resolving the problem of negation that usually appears in any review. Sentiment Vector Space Model (s-VSM) was used [5] for text representation to solve data sparseness. TF-IDF scheme is used to weight the features and vector space model is used for representing the reviews and finally Linear SVM classifier is built to classify the test corpus using SVM. Based on CHI square and document frequency (DI) a new feature selection algorithm CHI square difference between positive and negative categories (CDPNC) has been proposed. The experimental results illustrate that the classification performance is superior to the other feature selection method.

Authors used [6] 3 different datasets to test SVM. Pang and Lee movie review, Taboda and Grieve computer reviews and one dataset is crawled from Amazon about digital cameras. VSM model was used to generate the bag of words for each document. Main goal is to compare different n gram schemes and different approaches such as TFIDF, BO and TO have been used to generate word vector. 3 cross and 10 cross validations have been applied on the dataset. It’s found that TO be the worst weighting scheme for the corpus selected and found unigram is the worst option for their system. Mobile data reviews [7] have been collected as dataset. First the product reviews are transferred to the part of speech tagger. TF-IDF scheme was used for extracting the features and clustering techniques for validating the feature and SVM as the classifier. As the clustering technique is used for feature validation the accuracy of the classifier is improved.

The corpus used [8] was gathered from Twitter, tweets about top 10 automotive brands. Feature vector is the most important concept in implementing a classifier. Unigram model outperforms all other models. The choice of kernel and proper tuning of SVM hyper parameters are core factors, contributing to SVM accuracy. RBF kernel is used and the SVM type used is C-SVC which is the multiclass classification. The dataset consisted [9] Epinions.com movie reviews half positive and half negative. SVM was used to bring several favourability measures (OS good, Turney) for phrases and adjectives which is combined with unigram models and lemmatized versions of unigram models. . Linear kernel SVM Used 3 and 10 cross fold validation. Lemmas outperform unigrams in all experiments.

The inventory [10] dataset from the internet have been used for this work. The customers are classified based on their purchase behaviour to predict the online purchase turn over. The SVM model classifier was built using this dataset and predicted the customer behaviour with good accuracy. The tweets are crawled [11] from the twitter and used for the work. There are foreground and background tweets (noisy tweets). These noisy tweets were removed by the FB-LDA models. FB-LDA removes the noisy tweets and gets the most optimised tweets related to the target. RCB-LDA model was used to find out the more relevant tweets for sentiment variation. Sentiment label is assigned to the tweets using Sentistrength and twitter sentiment analysis tools and then tweets were classified by the SVM. This paper [12] deals with text valence detection. The solution has been based on SVM classifier. Datasets were obtained from real user feedback on products from different web pages. The proposed solution has been evaluated with English, German, Czech and Spanish languages. SVM with big data approach yielded better accuracy

3. Methodology

3.1. Scraping mobile tweets from Twitter:

Twitter is a social media page where people used to discuss about various topics happens around the world. In order to get tweets from twitter one should create their own Consumer_key, Consumer_secret, Access_key, Access_secret. The Apple iphone11 mobile data reviews were scraped through twitter API mentioning the product name in the search query and the scraped tweets will be saved as .CSV file. The tweets, tweet id, date of the tweet have been stored in the CSV file.

3.2. Pre-processing:

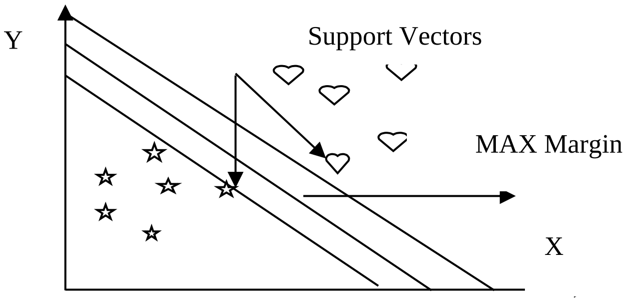
Scraped tweets will have some noisiness which will be irrelevant to the sentiment. That noisiness will be removed using pre-processing techniques. Converting the whole tweets as lowercase, removing punctuation, white spaces, single spaces, new line characters, words with one or two characters, stop words will bring the new look to the tweets. Then the tweets will be tokenized, stemmed and lemmatized.

3.3. Feature Vectors:

Feature vectors are used to represent numeric values of the text in the reviews. Feature vector is an n-dimensional vector of numerical values which represent some object. SVM algorithm needs numeric representation of reviews to do processing and analysis. Thus feature vectors are created to do the analysis using SVM algorithm.

3.4. Support Vector Machine:

SVM is a supervised learning algorithm used for classification, SVM algorithm learns from the training data in order to find the optimal hyper plane to classify properly whenever the new data occurs. In SVM, Each data is a vector. The dimension of the space will be defined based on the features in the dataset. SVM choose support vectors between the classes. The support vectors are the data points from the training dataset which lies close to the optimal hyper plane. If the SVM wants to classify the test dataset, support vectors will be sufficient for the better classification.



3.4.1. SVM Kernels:

SVM has a technique called kernel trick. Kernels are the mathematical functions used to transform the non-separable problem into separable problem. Some important SVM kernel functions are linear kernel, Gaussian kernel, and Radical Basis kernel (RBF) and Polynomial kernel.

4. Results and Discussions

Live tweets about Apple iPhone 11 were scraped from Twitter. 3 sets of training data were taken from twitter. One dataset contains 18000 tweets and the second one contains 34000 and the third one contains 66000 tweets. Tweets were pre-processed and converted into feature vectors. Cross validation techniques were used to do the analysis of reviews. In cross validation, the original data set will be partitioned into training set to train the classifier and test set to evaluate the classifier. 4 cross validation technique was applied while doing classification where one fourth of the dataset will be considered as test data and the cross validation will be repeated 4 times in which the test sample changes every time.

The analysis has been done using various kernel functions of SVM. Results have been taken using RBF kernel, linear kernel and polynomial kernel. The accuracy rate is compared between the different kernels of SVM using 3 different sizes of datasets contains iphone mobile reviews.

The below Table and the Figure summarizes the Accuracy rates of various kernels based on different dataset sizes.

SVM Kernels	**Dataset Size (Tweets)**	**Accuracy (%)**
RBF	18000	91.87
	34000	75.45
	66000	71.06
Linear	18000	84.05
	34000	62.80
	66000	56.23
Polynomial	18000	84.69
	34000	61.69
	66000	56.75

Table 1: Accuracy based on various kernels and datasets size

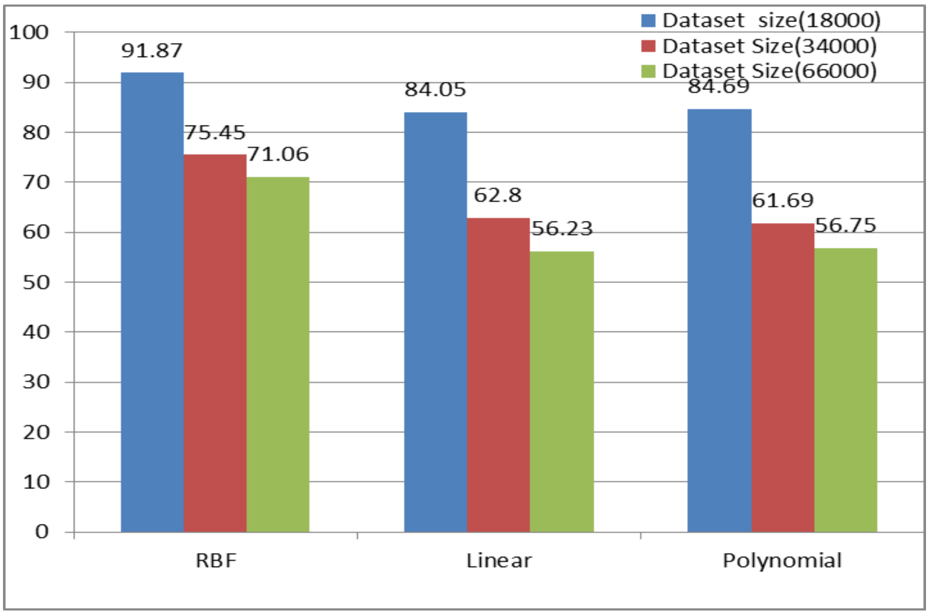


Figure 1: Accuracy based on various kernels and datasets size

5. Conclusion and Future work

Tweets about mobile phone have been scraped from Twitter. 3 different sizes of dataset have been used to do the analysis using Support vector machine as a classifier. Different kernels of SVM have been used while doing analysis and 4 cross validation techniques were also used to bring out the better accuracy. This analysis have found that RBF kernels works well doing sentiment analysis among all the kernels and at the same time the accuracy gets lower once the data grow bigger. As a future work, The SVM can be integrated with other classification algorithms or can be integrated with deep learning techniques to achieve better accuracy regardless of the dataset size

References

1. Nurulhuda Zainuddin, Ali Selamat, “Twitter Feature Selection and Classification Using Support Vector Machine for Aspect-Based Sentiment Analysis”, Springer International Publishing Switzerland, pp. 269–279, 2016
2. Nurulhuda Zainuddin, Ali Selamat, ”Sentiment Analysis Using Support Vector Machine”, IEEE International Conference, 2014.
3. Bhumika M. Jadav,Vimalkumar B,” Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis”, International Journal of Computer Applications, Vol. 146 ,No.13, July 2016
4. D.v.Nagarjuna Devi, Siriki Prasad,” A feature based approach for sentiment analysis by using support vector machine”, IEEE international conference, 2016
5. Fang Luo, Cheng Li, “Affective-Feature-based sentiment analysis using SVM classifier, IEEE international conference, 2016
6. M. R. Saleh, M. Martın-Valdivia, A. Montejo-Raez, and L. Urena-Lopez, “Experiments with svm to classify opinions in different domains” , Expert Systems with Applications, vol. 38, no. 12, pp. 14 799– 14 804, 2011
7. Upma kumara, Aravind K Sharma, “Sentiment Analysis of Smart Phone Product Review using SVM classification Technique”, ICECDS, 2017
8. Jao Allen, Kurt junshean, “Optimizing Support Vector Machine in Classifying Sentiments on Product Brands from Twitter”, IEEE Explore, 2014
9. Tony Mullen , Nigel Collier, “Sentiment analysis using support vector machines with diverse information sources”, Conference proceedings on Empirical Methods in Natural Language Processing, 2004
10. K.Maheswari, P.Amutha Priya, “Predicting Customer Behaviour in online shopping Using SVM Classifier”, IEEE International Conference, 2017.
11. Bholane Savita, Deipali Gore, “Sentiment Analysis on Twitter Data Using Support Vector Machine”, International Journal of Computer Science Trends and Technology, Vol.4, Issue 3, 2016
12. Lukas Povoda, Radim Burget, “Sentiment Analysis based on Support vector machine and Big data”, IEEE International Conference, 2016