ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

SentiView: A Lexicon Based Approach for Twitter Sentiment Analysis

Sandip D Mali, Dr. Sachin N Deshmukh, Ashish A Bhalerao

Student, Dept. of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MH), India

Professor, Dept. of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MH), India

Student, Dept. of CS & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MH), India

ABSTRACT: Today microblogging website are very popular like twitter on which user post their views, opinion etc. The information is generated either through computer or mobile by one user and many can view them. In this paper we focus on using twitter for sentiment analysis. Sentiment analysis is challenging task for this we can use various machine learning algorithm like Naive Bayes, SVM, maximum entropy etc. Sentiment analysis refers to predicting or telling the document or sentence text holds positive, negative or neutral opinion on some target. The aim of this paper is to develop a new system called SentiView which a lexicon based approach for sentiment analysis.

KEYWORDS: twitter, sentiment analysis, opinion mining, sentiment classification.

I. INTRODUCTION

As the users of microblogging platforms and services grow every day, data from these sources can be used in opinion mining and sentiment analysis tasks. For example, manufacturing / commercial companies may be interested in the following questions:

- What people think about our product (service, company etc.)?
- How positive (or negative) are people about our product?
- What would people prefer our product to be like?

Twitter is a popular real-time microblogging service that allows its users to share short pieces of information known as “tweets”, means tweet is the small text that would be generated by user related to certain things like product, his own opinion, his beliefs etc. The only problem with tweet is that its length should be less than or equal to 140 characters. First we will introduce various properties of messages that users post on Twitter. Some of the unique properties include the following:

1) Usernames: Users often include Twitter usernames in their tweets in order to direct their messages. A de facto standard is to include the @ symbol before the username (e.g @liang).

2) Hash Tags: Twitter allows users to tag their tweets with the help of a “hash tag”, which has the form of #<tagname>”. Users can use this to convey what their tweet is primarily about by using keywords that best represent the content of the tweet.

3) RT: If a tweet is compelling and interesting enough, users might republish that tweet, commonly known as retweet, and twitter employs “RT” to represent re-tweeting (e.g. “RT @RodyRoderos: I love Iphone 6 but I want Samsung note 2 :(”).

Tweets are also called as the microblog because of its short text. Microblogging websites have evolved to become a source of varied kind of information. This is due to nature of microblogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express their opinion for products they use in daily life. Due to this, Microblogging websites have evolved to become a major source of a diverse variety information, with millions of messages appearing daily on popular web-sites. Product reviewing has been rapidly growing in recent years because more and more products are selling on the Web. The large number of reviews allows customers to make informed decisions on product purchases. However, it is difficult for product manufacturers or businesses to keep track of customer opinions and sentiments on their products and services. In order to enhance the customer shopping experiences a system is needed to help people analyze the sentiment content of product reviews.In

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

this paper our main aim is the findings the tweets that contain opinion and based on them later determine their orientation that is the tweet is contain either positive or negative or neutral polarity. For that purpose we use a lexicon based or unigram model.

II. RELATED WORK

Po-Wei Liang and Bi-Ru Dai propose a new system architecture that can be automatically analyze the sentiment of microblogs or tweets. They combine this system with manually annotated data from twitter which is one of the most popular microblogging platforms for the task of sentiment analysis. In this system, machines can learn how to automatically extract the set of messages which contain opinions, filter out non-opinion messages and determine their sentiment directions. For this paper, they crawl tweets from twitter and perform some preprocessing on it. They retrieve tweets using twitter API. They crawl tweets of three distinct categories (camera, mobile phone, movies) as their training set from the time period November 1, 2012 to January 31, 2013. They perform some preprocessing task on that like eliminate tweets that are not in English, have too few words, have too few words apart from greeting words, have just URL. After all the remaining tweets are pre-processed as all words are transformed to lower case, extract emoticons with their sentiment polarity, targets are replaced with user, pos tagging, remove sequence of repeated characters and stop-words. According to the previous preprocessing step all words are transformed into a tuple structure (word, pos tag, English-word, stop-word). In the next stage filter-out tweets without opinion, to do this they use Naive Bays (NB). In this step, the system can classify the tweets into opinion and non-opinion class. Then the system passes the opinion part into the next step i.e. short text classification. In this part they observed that a word may have different meanings in different domains. For this they use two different algorithms like Mutual Information (MI), and X2 test. The final step of their work is to determine the orientation of the tweets i.e., positive or negative. In this paper they got accuracy about 67.58% for unigram and 70.39% for opinion miner. This result show that opinion miner give better result than unigram model.

Apoorv Agarwal, BoyiXie, Ilia Vovsha et.al build models for two classification tasks: a binary task of classifying sentiment into positive and negative classes and a 3-way task of classifying sentiment into positive, negative and neutral classes. They experiment with three types of models: unigram model, a feature based model and a tree kernel based model. They use manually annotated Twitter data for their experiments. One advantage of this data, over previously used data-sets, is that the tweets are collected in a streaming fashion and therefore represent a true sample of actual tweets in terms of language use and content. For all their experiments they use Support Vector Machines (SVM) and report averaged 5-fold cross-validation test results.

Go, A., Huang, L., Bhayani, R introduce a novel approach for automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative with respect to a query term. Their training data consists of Twitter messages with emoticons, which are used as noisy labels. They show that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) have accuracy above 80% when trained with emoticon data. In this paper they also describes the preprocessing steps needed in order to achieve high accuracy.

Alexander Pak and Patrick Paroubek, they focus on using Twitter, the most popular microblogging platform, for the task of sentiment analysis. They show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. They perform linguistic analysis of the collected corpus and explain discovered phenomena. Using the corpus, they build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document. They collected a corpus of 300000 text posts from Twitter evenly split automatically between three sets of texts i.e., positive emotions, negative emotions and no emotions. They perform statistical linguistic analysis of the collected corpus.

B. Pang, L. Lee, and S. Vaithyanathan they examine the effectiveness of applying machine learning techniques to the sentiment classification problem. They consider the problem of classifying documents not by topic, but by overall sentiment, e.g., determining whether a review is positive or negative. For their experiments, they choose to work with movie reviews. Their data source was the Internet Movie Database (IMDb) archive of the" www.rec.arts.movies.reviews" newsgroup. This dataset is available on-line at http://www.cs.cornell.edu/people/pabo/-movie-review-data/. Their aim in this work was to examine whether it suffices to treat sentiment classification simply as a special case of topic-based categorization. They experimented with three standard algorithms: Naive Bayes classification, maximum entropy classification, and support vector machines In terms of relative performance, Naive Bayes tends to do the worst and SVMs tend to do the best, although the difference are not large.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

P. Turney, presents a simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down). The classification of a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. A phrase has a positive semantic orientation when it has good associations (e.g., "subtle nuances") and a negative semantic orientation when it has bad associations (e.g., "very cavalier"). The first step is to use a part-of-speech tagger to identify phrases in the input text that contain adjectives or adverbs. The second step is to estimate the semantic orientation of each extracted phrase. A phrase has a positive semantic orientation when it has good associations (e.g., "romantic ambience") and a negative semantic orientation when it has bad associations (e.g., "horrific events"). The third step is to assign the given review to a class, recommended or not recommended, based on the average semantic orientation of the phrases extracted from the review. If the average is positive, the prediction is that the review recommends the item it discusses. Otherwise, the prediction is that the item is not recommended. The PMI-IR algorithm is employed to estimate the semantic orientation of a phrase. PMI-IR uses Pointwise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words or phrases. The semantic orientation of a given phrase is calculated by comparing its similarity to a positive reference word ("excellent") with its similarity to a negative reference word ("poor"). In experiments with 410 reviews from opinions, the algorithm attains an average accuracy of 74%.

Kunpeng Zhang, Yu Cheng, YushengXie et.al develop a sentiment identification system called SES which implements three different sentiment identification algorithms. They augment basic compositional semantic rules in the first algorithm. In the second algorithm, they think sentiment should not be simply classified as positive, negative, and objective but a continuous score to reflect sentiment degree. All word scores are calculated based on a large volume of customer reviews. Due to the special characteristics of social media texts, they propose a third algorithm which takes emoticons, negation word position, and domain-specific words into account. They build a web-based system called SES, which ensembles three algorithms we implemented and uses machine learning method to predict text sentiment. The system is aiming to predict sentiment on both sentence and document level. They conduct experiments on Facebook comments and twitter tweets using four different machine learning models: decision tree, neural network, logistic regression, and random forest. The experiment results show that random forest model reaches highest accuracy.

III. PROPOSED ALGORITHM

The proposed system is based on lexical approach. The following figure 1 shows the architecture and work flow of a SentiView system.

The SentiView systems architecture is divided into mainly six region.

1) Getting Data
2) Preprocessing on data
3) Generating tweet score
4) Extracting tweet containing opinion
5) Classification
6) Results

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

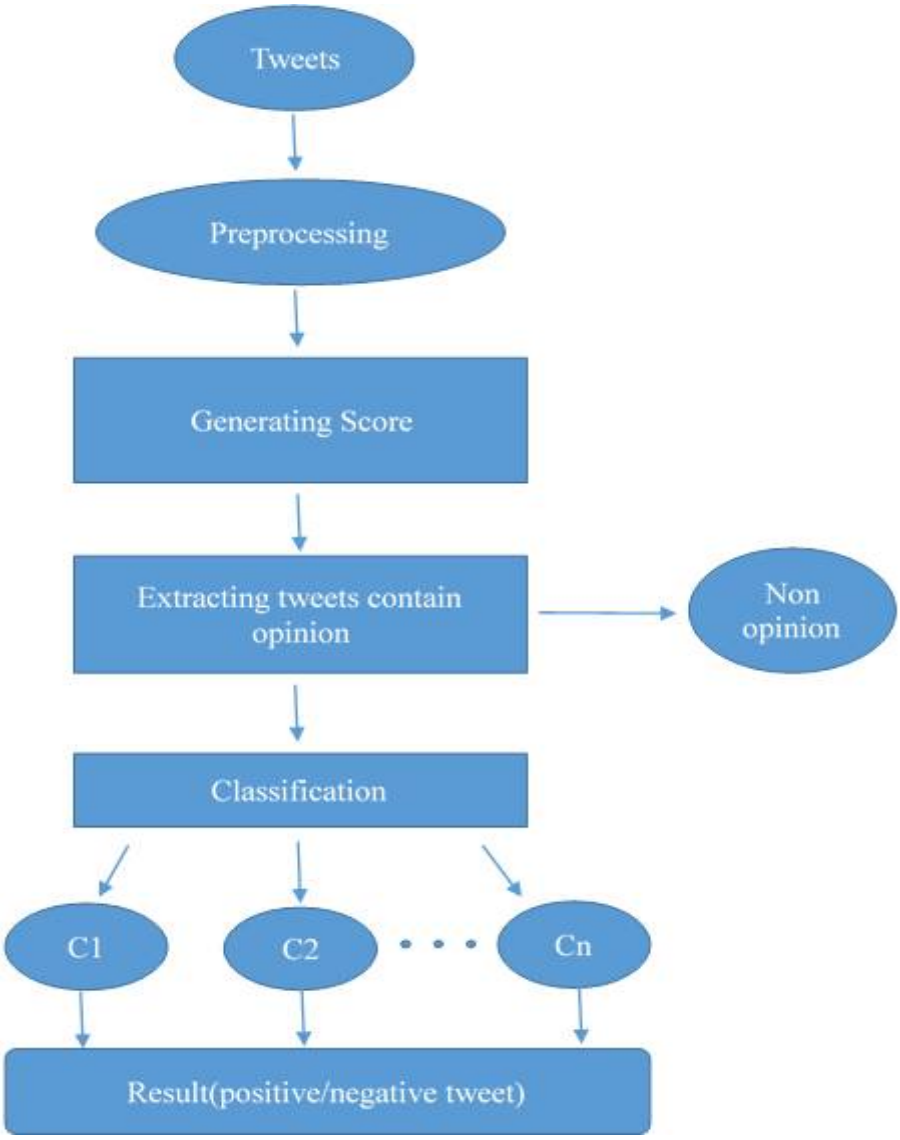


Figure1: System architecture of SentiView.

All these region are correlated with each other that is output of one region is input to the next region

1) Getting Data- It is the first step of SentiView, the aim is to get twitter data for this we use R Twitter API. By using this API we can extract tweets on any keywords or tag like fruits, market, #IPL, #IPhone etc. we can download tweet up to 30000 at a time we can download more tweet at time but some time session is timeout. But one main drawback of this API is we can download only up to 8-10 days only and by default we get recent tweets.
2) Preprocessing on Data- As the data is generated by the user so it is not perfect according to English language. It has so many impurities like spelling mistakes, repetition of characters, repetition of words, excess use of punctuations etc. In preprocessing we do the following tasks

- Remove retweets.
- Remove non-English tweets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

- All words are transformed into lower case.
- Remove Numbers.
- Remove punctuations.
- Remove stopwords.
- Word stemming.

3) Generating tweet score:-

In this stage we generate a single score for every tweet. The score is generated by how many positive or negative words are present in the tweet are counted for each positive word hit score is increased by 1 and for each negative word hit score is decreased by 1. The slandered positive and negative word list developed by Bing Liu is used for score generation. One important thing is that the place of positive word or negative word does not affect the score that is place of word is neglected.

4) Extracting tweet containing opinion:-

For this system we are only interested in the tweets that hold some positive or negative sentiment or opinion. Thus we eliminate the neutral tweets. Due to elimination of the neutral tweet the accuracy of SentiView system is increased compared to the previous system. For this task we need to set threshold to decide tweet contain any opinion or not. The threshold is set on score generated for every tweet. If the score of the tweet is non zero the tweet belongs to opinion class if score of the tweet is zero then tweet belongs to non-opinion class and thus removed from data or corpus.

5) Classification:-

In the classification stage we apply various machine learning classifier on data to classify tweet is either positive or negative. SentiView is a two class classifier system that is positive or negative class. For this purpose we use four algorithm that is Tree, Random Forest, Maximum Entropy and SVM.

6) Result:-

After the classification tweet the result of the classification is shown that is tweet is positive oriented or negative oriented

IV. EXPERIMENTAL WORK AND RESULT

The SentiView system uses lexicon based or unigram model for classification. The system is developed using the R platform. The data used in this system is crawled by an R twitter API under the package tweetR. The dataset contain approximately 5000 tweets downloaded using the tweeter API. The following figure shows the image of twitter API.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

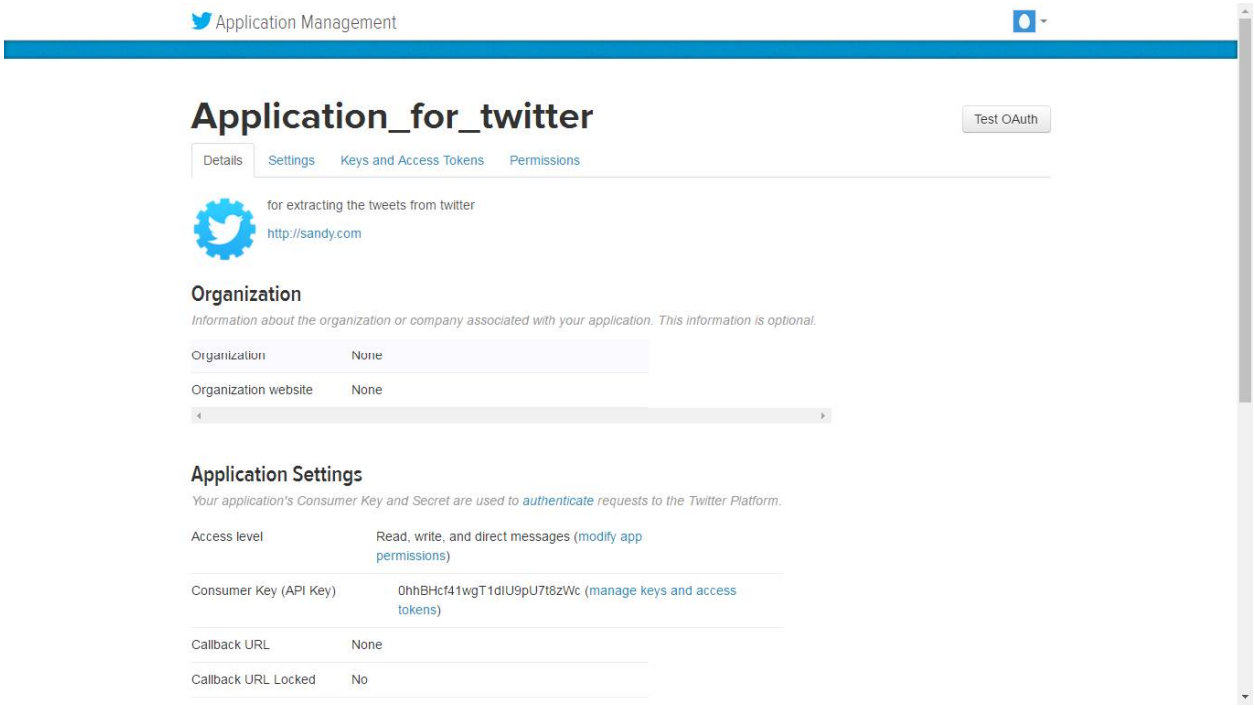


Figure 2- Screen shot of Twitter API

For creating Twitter API we need a twitter account with linked mobile number. Visit the website http://app.twitter.com and login using twitter username and password this will redirect you to a page where you can create, delete or manage you twitter API`s. After creating twitter API, this provide you access keys to this API for programming interface. The key include access key, access secrete key, consumer key and consumer secrete key. To access the tweet you have to create a session with twitter and download the tweet that you are interested.

We have downloaded the tweet for the #IPhone tag means this will return the tweet that has text in tweet #IPhone from recent to old fashion automatically. For our task we have downloaded 10000 tweets.

These tweets are written by the user so they are not correct and some tweet are also retweeted. Thus we remove the non-English tweet as well as tweet that are retweeted. We also remove the punctuations, numbers, stopwords because commonly they are not related with opinion and thus removed to minimize the search space. After removal of the stopword we apply steaming on the remaining that is transform the word to its base form. After steaming our preprocessing step is completed and we began to next step that is score generation.

A score is generated using unigram model or dictionary based model or lexical model.in score generation the word of each tweet are separated and then compared with the word list of positive and negative words provided by the Bing Lui. This word list contain the list of positive words and the list of negative words. If for positive words any hit occurs the score is increased by 1 and if hit is occurs for negative word then score is decreased by 1. In this step the score value is assigned to every tweet. It should be positive or negative depends upon the tweet. The maximum and minimum score we have got is +15 and -13 and all other values lies between them. The result of this step is stored in matrix form.

After generation of the score we have mark threshold such that below the threshold tweet are negative and above the threshold tweets are positive. Thus we set our threshold to 0 and we classify the tweet class based on score in either positive or negative. The tweet that has score exactly 0 are removed because they are either positive or negative or neutral. They are in conflict and thus they are removed.

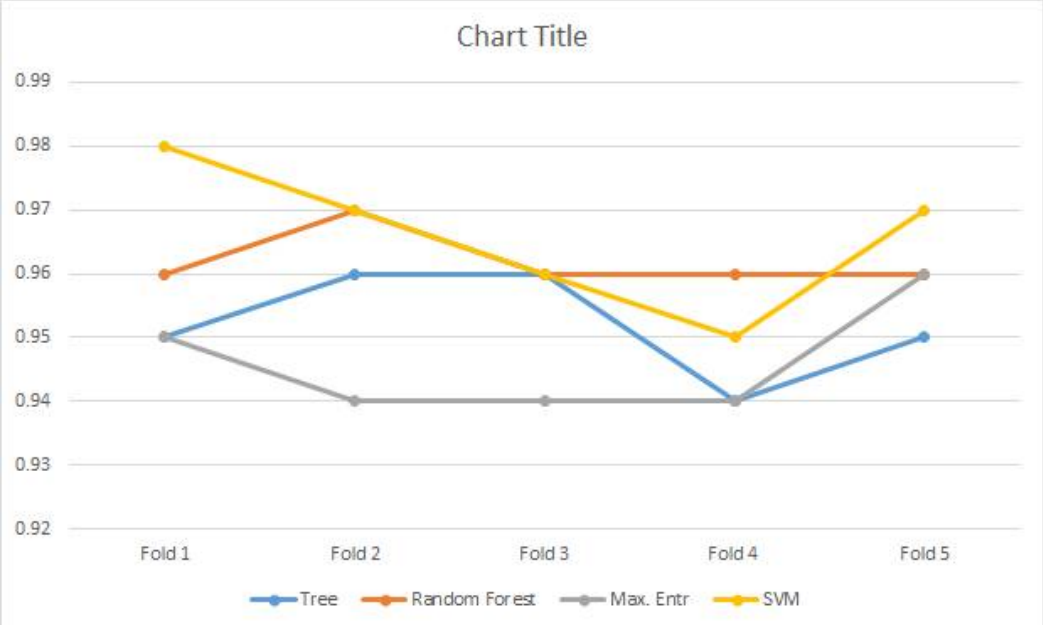
The following figure shows the graph of algorithm performance and the 5 fold cross validation.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

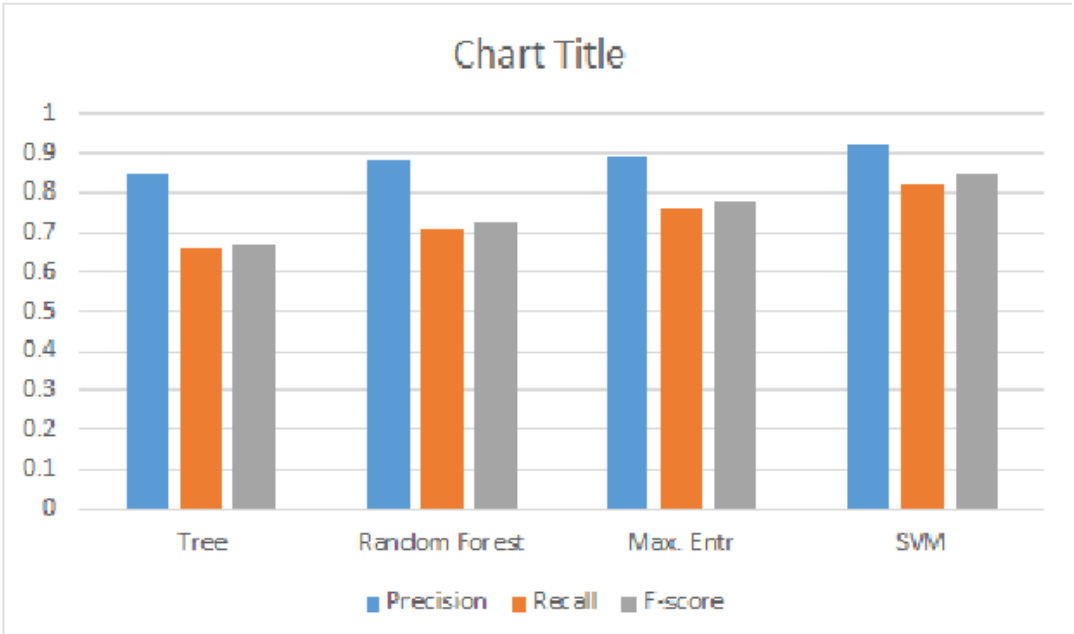
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016



Graph 1- 5 fold cross validation of algorithm.

The above graph 1 shows the result of 5 fold cross validation on each algorithm and in graph we can clearly see that SVM gives the better result as compare with others. The graph 2 shows the precision, recall and F-score values for each algorithm we have used and in that graph we can clearly see that again SVM gives the better result as compare with the others.



Graph 2- Algorithm performance.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

V. CONCLUSION AND FUTURE WORK

From the result we got using SentiView we can say that among the all other algorithm SVM gives the better result. The main reason of high accuracy comparing to previous is due to pre-processing and removal of non-opinion tweets from data, this reduce the search space.

The main drawback with this system is it lexicon based and domain independent. Due to lexicon based the heart of the system is word dictionary and it is limited. It consider only the word that have in the dictionary. Our future work is add as many word as possible to the dictionary and try other models like bigram model and create domain specific dictionary to get more result.

REFERENCES

1. Po-Wei Liang and Bi-Ru Dai, "Opinion Mining on Social Media Data", IEEE,2013.
2. Apoorv Agarwal, BoyiXie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau, "Sentiment analysis of twitter data",In Proceedings of the Workshop on Languages in Social Media, pages 30–38. Association for Computational Linguistics, 2011.
3. Go, A., Huang, L., BhayaniR, "Twitter sentiment classification using distant supervision", In: CS224N Project Report, Stanford 2009.
4. Alexander Pak and Patrick Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining",Proceedings of LREC, 2010.
5. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.
6. P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Appliedto Unsupervised Classification of Reviews", ACL, 2002.
7. S. Kim and E. Hovy, "Determining the Sentiment of Opinions", COLING, 2004.
8. Kunpeng Zhang, Yu Cheng, YushengXie et.al, "SES: Sentiment Elicitation System for Social Media Data", Proceedings of the IEEE 11th International Conference on Data Mining Workshops, p.129-136, December 11, 2011.
9. Diego ReforgiatoRecupero, Valentina Presutti, et.al. "Sentilo: Frame-Based Sentiment Analysis", Springer Science + Business Media New York 2014.
10. Luciano Barbosa, Junlan Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data" ColingPoster Volume, pages 36–44, Beijing, August 2010.
11. Dmitry Davidov, Oren Tsur, Ari Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys" Coling Poster Volume, pages 241–249, Beijing, August 2010
12. Bing Liu, "Sentiment Analysis and Opinion Mining", April 22, 2012.