(IJUSE) 2018, Vol. No. 4, Jan-Dec e-**ISSN: 2454-759X, p-ISSN: 2454-7581**

LEVERAGING THE IMPROVED PRE-PROCESSING OF TEXT AND SVM (SUPPORT VECTOR MACHINE) WITH KERNEL TO DEVELOP INNOVATIVE APPROACH IN TWEET SENTIMENT ANALYSIS.

Hardik Chaudhary

ABSTRACT

Tweet sentiment analysis is effective problem now days, because much information comes from tweets by different users and analysis of opinion of product, event, movie etc. In this paper analysis of tweets by it's features and classes by support vector machine with kernel approach. Our result show proposed pre-processing of text and classifier show significance improve from naïve bays and normal SVM.

I. INTRODUCTION

The platform like social media platforms has become an essential site for conversations related to politics in the entire world. In November 2012, the presidential election of United States, a microblogging service was developed to analyse the real time sentiments expressed in twitter in context of the President, Barack Obama, and the other nine challengers of republican as of this writing, four of them remains in the running. With these analyses, it has been explored that insights is provided by Twitter in context of the campaigns unfolding concept and public shifted opinion indications. Today, microblogging has become an essential communication tool for the users of internet. In popular web-sites, daily millions messages are appearing which provides microblogging services such as Tumblr, Facebook, and Twitter. Those messages writers (authors) gives their life details, opinions are shared regarding various topics and current issues are discussed. Due to the message free format and microblogging platforms easy accessibility, shifting of Internet users from traditional communication tools (like mailing lists or traditional blogs) to services of microblogging. The increasing posts by users about the services and products they utilized, or expressed their religious and political views, web-sites of microblogging which has become people's valuable sources for expressing their sentiments and opinions. This data can be utilized efficiently for social studies or marketing. A dataset is utilized which is in collected messages form from Twitter. The microblogging platform users create very short messages of larger number contained in Twitter. As the microblogging services and platforms audience growth is increasing day-by-day, these sources data can be utilized in sentiment analysis and opinion mining tasks.

In reviews, the sentiment analysis is the product reviews exploration process on the internet for determining the opinion overall or feeling regarding a product use. The content generated by users is the representation of reviews, and this involves increasing attention and for marketing team's rich resources, psychologists and sociologists and concerned others along with views, opinions,

personal or general attitudes and public mood [1]. The web having reviews of larger number on the web represents the feedback of user's current form. It is difficult for companies or humans to get the newest trends and the general or state opinions summarizes about products owing for the biggest size diversity and social media data size, which creates the requirement of real time and automated opinion extraction and mining. The opinion sentimental senitiment decision is a challenging problem because of its factor subjectivity that is what people are thinking essentially. Sentiment analysis is considered to be classification of task as the text orientation is classified into either negative or positive. The widely utilized approach is machine learning towards classification of sentiments to the methods based on lexicon and linguistic methods in addition [2]. It has been claimed that these techniques perform average in classification of sentiment as they perform in categorization of topic owing to its opinionated text nature in which more text understanding is required while occurrence of few keywords could be the main key to the classification accuracy [3]. The classifiers in machine learning such as maximum entropy, support vector machine (SVM) and Naive Bayes are utilized in [3] for classification of sentiments for achieving accuracies.

II. LITERATURE REVIEW

Analysis of sentiments has been handled as a task of Natural Language Processing at many granularity levels. A classification task of document level from which it starts [1, 2], which has been handled at the sentence level [3, 4] and at the phrase level more recently [5, 6]. The data microblog such as Twitter, on which real time reactions are posted by the users to and "everything" opinions, poses fresher and various challenges. Some recent and earlier results on sentiment analysis of sentiments on Twitter dataset are by [7], [8] and [9]. Distant learning is used in [7] for acquiring the data sentiments. Tweets ending are utilized in negative emoticons like ":-(""":(" as negative and positive emoticons such as ":-)" ":)"as positive. Models are built utilizing MaxEnt, Support Vector Machines (SVM), and Naive Bayes, and as reported that other classifiers are outperformed by SVM. Bigram, Unigram model is tried in terms of feature space in conjunction along with POS (parts-of-speech) features. All other models are outperformed by unigram as analyzed by them. POS features and bigrams do not help specifically. In [9], data is collected which follows same distant learning model. A different task classification performed by them though: objective versus subjective. The tweets ending are collected with emoticons for the subjective data in a similar manner as [7]. Popular newspapers twitter accounts are crawled by them for objective data such as "Hindustan Times", "Washington Posts", "New York Times", etc. They reported that both bigrams and POS and bigrams help as contrary to presented results in [7]. However, both these approaches are based primarily on the ngram models. Furthermore, the utilized data for testing and training is collected from the search queries and is biased therefore. In contrast, the features are presented which achieves gain significantly over a baseline unigram. Various data representation method is explored additionally and significant improvement is reported over the unigram models.

Another contribution is that manually reported results annotated data in which there is no suffering from known biases of any kind. Random sample data of tweets streaming unlikely collected data by utilizing specific queries. The hand-labeled data size allows performing experiments in terms of cross validation and variance checking in classifiers performance across folds. Another

(IJUSE) 2018, Vol. No. 4, Jan-Dec e-**ISSN: 2454-759X, p-ISSN: 2454-7581**

significant effort for Twitter data sentiment classification is given in [10]. They use Polarity predictions are used by them from three different websites for training a model as noisy labels and utilized 1000 labeled manually tweets to tune and another 1000 labeled manually tweets to test. However, their test data collection is not mentioned. The syntax tweets features use is proposed by them such as hash-tags, re-tweet, punctuation, link, and marks of exclamation in conjunction having features such as prior words polarity and words POS. Their approach is extended with the use of prior polarity based on real value, and with POS and prior polarity combination. The obtained results showed that the features which enhance the classifiers performance of the most features combining prior words polarity with their speech parts. The syntax features of tweet help but only marginally. In [11], sentiment analysis is performed on data feedback from Global Services Support survey. The linguistic features role is analyzed such as POS tags. The analysis of extensive feature is performed by them and selection of feature and demonstrated the abstracted linguistic features analysis contributes to the accuracy of classifier. An extensive feature analysis is performed which shows that the performance of the abstracted100 linguistic features is similar to as hard baseline unigram.

The sentiment analysis (SA) task, a.k.a. opinion mining, has been a popular research topic in communities for years. Sentiment analysis previous research [12, 18] focused mainly on reviews of movies or product, which are convenient experimentally and evaluation is easy. For other types of documents including news and webpages, efforts are made for exploring the similar task [28]. While such work bulk has been focused on the level of the document, few others [17, 16, 15] addressing the analysis of sentiment in the phrase and level of sentences that regards sentences as samples classification. The sentiment polarity is obtained for a given text (sentences, snippets, or documents). This thoroughly studied problem is compared; investigated rarely for topic like sentimental analyses. Though the attempt of few work for incorporating the sentiment factor in the topic models like PLSI (probabilistic latent semantic indexing) or LDA (latent Dirichlet allocation) for giving the opinion generation description [14, 13], still it is hard for reaching an agreement to the definitions related to topics and how the sentiment classification (negative/positive) meaning is given for them. The problem is in the definition to sentiment polarity topic utilizing one-bit representation (negative or positive) only is not well-modeled.

III. MACHINE LERANING METHODS

We test various classifiers: Naïve Bayes and SVM and comparing with the proposed method on the basis of parameter: accuracy, precision, recall and F-measure.

Baseline

Twitter is social media website on which tweets are utilized for sentimental analyses. The basic approach is to utilize a list of negative and positive tweets. As a baseline, publically available list of twitter keyword is utilized. For every tweet, the appearing number of positive and negative keywords is counted. Higher polarity count is returned to the classifier.

Support Vector Machine

(IJUSE) 2018, Vol. No. 4, Jan-Dec **e-ISSN: 2454-759X, p-ISSN: 2454-7581**

Another popular classifier technique is Support Vector Machine (SVM) and we utilize SVM with linear kernel. Input data is in the form of vectors set. Each vector entry corresponds to feature presence. As an example, every single word found in the tweet represent individual feature. It is assigned value 1, when feature is present and 0 is the assigned value while the feature is absent. The presence of feature utilized which opposes the count, so that input data does not need to be scaled, which increases the overall process speed. The SVM approach is built on the tweet level classifier. For estimating negative or positive probability, we utilized average distribution of tweets polarity:

$$P_t(x_i|y_i) = \frac{\sum_{\tau \in T_i} P_{t_{x_i}}(\tau)}{\sum_{\tau \in T_i} P_{t_p}(\tau) + \sum_{\tau \in T_i} P_{t_n}(\tau)} \qquad x = \arg \max_{x_i \in (p,n)} P_t(x_i|y_i)$$

IV. SYSTEM MODEL

Step 1: API extracts objects from Twitter with higher accuracy.
Step 2: Pre-processing tweets collected and data for enhancing accuracy and removing noisy features.
Step 3: The system extracts subjective features based on Information Gain, Bigram and use Alchemy API data set to extract the object-oriented feature. Extract object words of tweets which have sentiment polarity and named object-oriented feature.
Step 4: The system puts tweets extracted features into the Support Vector Machine classifier and the classifier classify tweets into the negative class or the positive class.

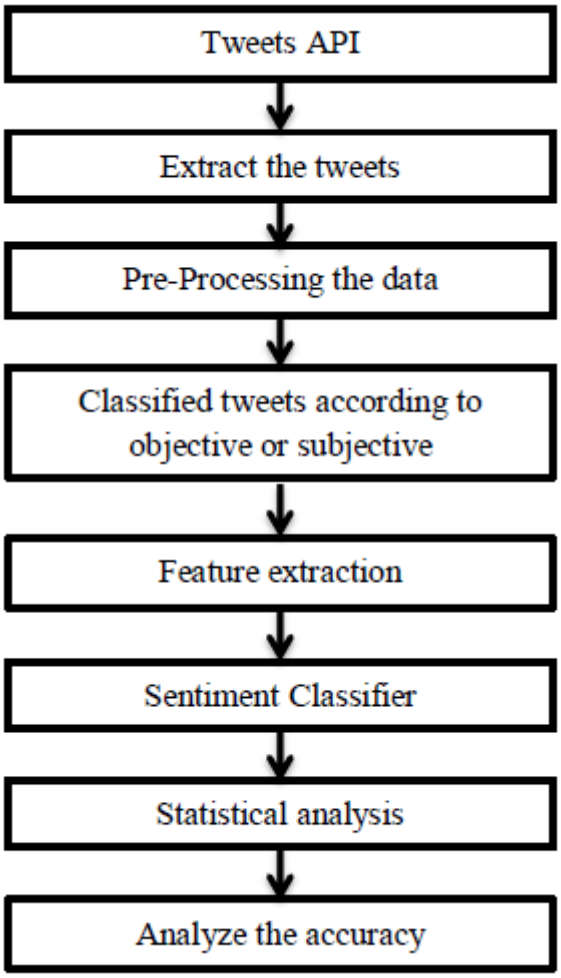

Fig 1: System model

V. EXPERIMENTAL RESULTS

Table 1: Comparison table of different classifier based on various parameters (accuracy, precision, recall and F-measure).

	Naïve Bayes classifier	SVM classifier	SVM extended
Accuracy	59.5	69.9	84.05
Precision	44.44	57	100
Recall	90	90	72.5
F-measure	60	50	72.5

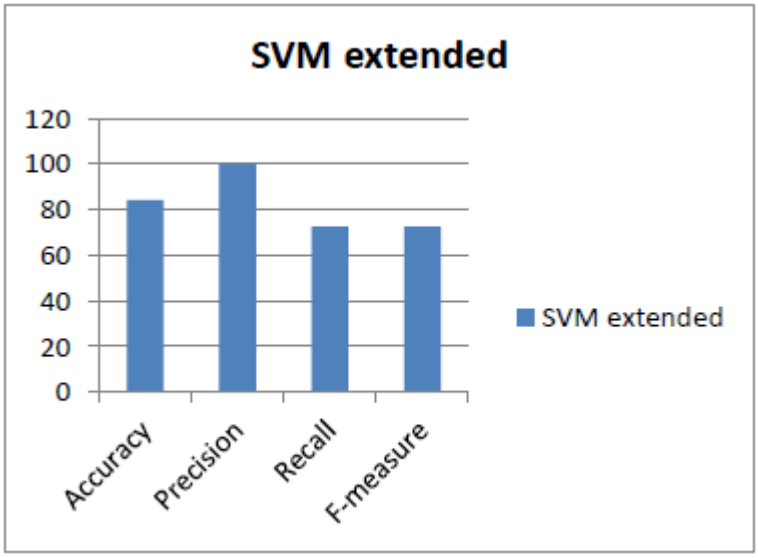

Fig.2: Performance analysis of SVM extended classifier on the bases of various parameters

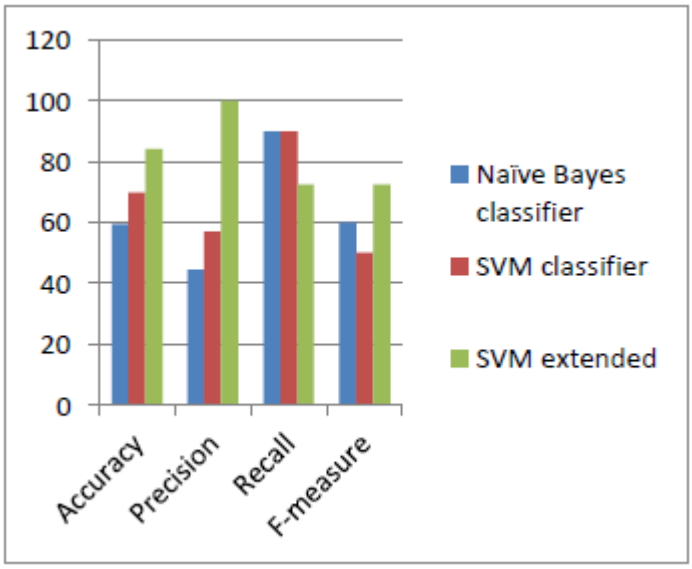

Fig.3: Comparison graph between different classifier based on various parameters (accuracy, precision, Recall and F-measure)

VI. CONCLUSION

Above given experiment analysis of tweet text by classifier Naïve Bayes, SVM and SVM – extended with kernel approach. Our results show SVM extended show improvement in Accuracy, precision, recall and F-measure.