**Sentiment Analysis Using Features Based Support Vector Machine – A Proposed Approach**

**Prakash P. Rokade, Dr. Aruna Kumari D**

Departement of Computer Science & Engineering, KLEF Deemed UniversityVaddeswaram, Guntur, A.P., India

Departement of Computer Science & Engineering, KLEF Deemed UniversityVaddeswaram, Guntur, A.P., India

|  |  |  |
| --- | --- | --- |
| **Article Info** |  | **ABSTRACT** |
| ***Article history:***  Received Jun 12th, 201x  Revised Aug 20th, 201x  Accepted Aug 26th, 201x |  | Business decisions for any service or product depend on sentiments by the people. The mood of people towards any event, service and product are expressed in sentiments. The text sentiment contains different linguistic features of sentence. A sentiment sentence also contains other features which are playing a vital role in deciding the polarity of sentiments.The features like duplication of sentiment, unknown emotics may change the polarity of sentiment.If features selection is proper one can extract better sentiments for decision making. A directed preprocessing will feed filtered input to any machine learning approach. Support vector machine is one of the tools of machine learning for better sentiment analysis.Better use of parts os speech (POS) folled by guided preprocessing and evaluation will provide less errorus polarity of sentiments. |
| ***Keyword:***  Feature Extraction,  Machine Learning,  Sentiment Analysis,  Support Vector Machine |
| ***Corresponding Author:***  Prakash P. Rokade,  Nalanda,Garimanagari,  Near Mahila Mahavidyalaya,  Kopargaon ,Dist.-Ahmednagar,  Maharashtra, 423601.  Email: [prakashrokade2005@gmail.com](mailto:prakashrokade2005@gmail.com) | | |

1. **INTRODUCTION**

Sentiment analysis of twitter moods is playing a vital role for business decisions to plan a good business strategy. A customer can view the sentiments online to decide for purchasing any product. A business can also explore the sentiments posted by customers to know the status of the service the industry is o providing. These sentiments are posted by a group, a person, and industry also. Hopefully we may expect that all the users are legal and loyal. But there may be some fake posts which are posted by fake users, competitors. As the input data for sentiments analysis is big data, this should be preprocessed before using for design any model. Duplicates sentiments, fake sentiments, blind sentiments must be removed by preprocessing input big data. After preprocessing this data are transfers into well known structure form. Latter on features like noun, adjective adverb, capital letters, numbers are used for sentiments selection. Dictionary based approach can be used to find positive, negative, neutral sentiments base upon the score of the whole document decision can be made by business intelligence. There are several methods for classification of sentiments. Machine learning approach, corpus base approach and dictionary based approach are the tools for sentiments evaluation. Support vector machine (SVM) is proved better machine learning strategy for sentiments classification.

1. **LITERATURE SURVEY**

A remarkable work is carried out in the research area of sentiment classification. The main focus of this work is on classifying larger pieces of text, like reviews of product or event [1]. Tweets are different from reviews as they have different purpose. Reviews are summary of author’s thoughts. Tweets are limited to 140 characters of text. Sentiment analysis is used to monitor and analyze social aspect. Tweets represent general mood of people through various reactions based on experience or as an impression for news articles [2]. Hu and Liu have given a technique for Feature Based Summarization system (FBS) of customer reviews of products. It also generated sentiment based summary as either positive or negative opinion using adjective words in reviews [3]. Chaovalit and Zhou compared supervised and unsupervised algorithm for classification and got 83.54% of accuracy for supervised method and 77% of accuracy for unsupervised method [4]. Pang O Keefe and Koprinska have given technique to select features using attribute weights and applied Navie Bayes and SVM classifiers for classification of moods [5, 6]. Lequistic features are used to detect the twitter sentiment. Author used hash tagged data set (HASH) and emoticon data set. By tokenization, normalization and parts of speech tagging.Classification is done. Results are evaluated by using unigrams and bigrams [7, 8].The study by Hassan shows that parts of speech features are not playing good role in sentiment analysis for micro-blogging domain. Author introduces classification method for query term sentiment analysis. Here classifier and feature extractor are considering two different components [9]. Each token is assigning a sentiments score called total sentiment index. Base upon the classification algorithm the sentiments are classified as positive or negative polarity sentiments [10]. Political future can be analyzed real time monitoring and analyzing public conversation on social sites [11]. Feature vectors and tagged content of corpus can be used to make model by using machine learning approach. This model is used to classify or categories untagged corpus of text document [13]. For language consistency twitter is more informal. Emoticons are used express the opinion. Many tweets are ambiguous and these are maximizing the opinion for readers; but deflect the opinion to a machine learning algorithm [14]. Sentiment classification algorithm (SCA) and SVM are used to evaluate the performance of the approach used accuracy, recall, precision are some parameters on which sentiment analysis performance is evaluated [15, 16].

1. **PROPOSED APPROACH**

## Mathematical Model

Let M be the model which describes the extraction,preprocessing,lebling and evaluating the sentiments .

S= {Tw, Pt, Sl, Se}

Where

* Tw = Twitter sentiments.
* Pt=PreprocessingofTweets
* Sl=Labling the sentiments as positive,negative or neutral

Sl= {Pv, Nv, Ne}

Pv={P1,P2,…,Pn}=PositiveClass

Nv={N1,N2,…,Nn}=NegativeClass

Ne={Ne1,Ne2,…,Nen}=NeutralClass

* Se=Sentiment evaluation

**Rsearch Design**

A proposed research design for sentiments analysis using feature base support vector machine techniques is given below.

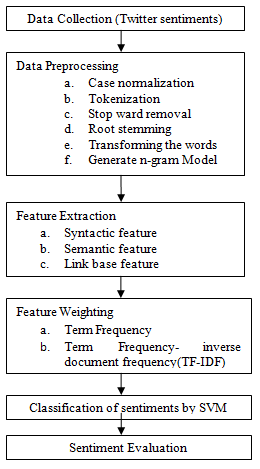


Fig. 1 Flow of Proposed Sentiment Analysis Approach

1. **Data Collection**

A correct input may leads us to get a correct output. Sentiments data is available on twitter website.

1. **Data Preprocessing**
   1. **Case normalization**

The tweets are available in combined case that is it may contain upper and lower case characters. In case normalization the entire document or sentence is converted in to lower case pattern generally.

* 1. **Tokenization**

A document is split in to sentences. Sentences may be divide in to words. By removing certain characters like punctuation marks these words are now tokens.

* 1. **Stop ward removal**

A set of stop words list is provided to remove them from sentiments. The frequently used stop words are “A, AN, THE, IS, THAT, SHALL, WILL”.

* 1. **Root stemming**

In this process derived words are reduced to their stem. The disadvantage of stemming is that all inflected forms must be listed in table. For example “Careful”, “careless”, “carefully” are reduced to “care”.

* 1. **Transforming the words**

A set of define rules are used to transform the word to a specific form. For example a word clarifies can be replaced by clarify.

Table 1 Words With Their Equivalent Stem

|  |  |
| --- | --- |
| **Words** | **Stem** |
| Equality, Equally | Equal |
| Engineering,Engineer,Engineered | Engineer |
| Manually,Manual,Man | Man |

* 1. **Generate n-gram Model**

N-gram models sequences, natural language using its statistical properties. The probability of world to be consider if it proves the condition on some number of previous words. For example one word in bigram model, two words in trigram model. Words are model such that every n gram will compose n words. Consider three gram sequence of characters generated from “good evening “ are “goo”, “ood”, “od”, ”ev”, “eve” and so on. It can also be used for good approximate matching.

However the value of n can be extended to higher level grams. The n-gram model can be better explained with the following examples:

Text: “Do exercise daily and eat healthy food..”

Unigrams: “Do”, “exercise”, “daily”, “and”, “eat”,” healthy”,” food”.

Bigrams: “Do exercise”, “exercise daily”, “daily and”, “eat healthy”.

Trigrams: “Do exercise daily”, “exercise daily and”, “daily and eat”.

Unigrams presents the simplest model for the n-gram approach. It consists of all the individual words present in the text. The bigram model defines a pair of adjacent words. Each pair of words forms a single bigram. The higher order grams can be formed in the similar way by taking together the n adjacent words. Higher order n-grams are more efficient in capturing the context as they provide better understanding of the word position.

* 1. **Removal of handles like # etc.**

Users 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.@alecmgo). An equivalence class token (USERNAME) replaces all words that start with the @ symbol.

**C. Feature Extraction**

In this relevant features are extracted the noun, adjective, adverb and their combinations are used as sentiment features. The feature words are extracted and scoring for sentiment is calculated. There are four categories of feature.

**Syntactic feature**

Related to traditional parts of speech like noun, verb, and prepositions.

**Semantic feature**

It is the study of meaning. It focuses on meaning base properties of noun. A semantic properties is specified in square brackets and plus or minus sign indicating the presence or absence of that property

Boy is [+human], [+male], [-adult]

A girl is [+human], [+female], [-adult]

**Link base feature**

Stylistic elements are give an idea or feeling to the literature. The text will be made distinctive in some way using this technique. It improves performance of sentiment classification.

**D. Feature Weighting**

**Term Frequency**

The number of times any important terms occurring in a sentiment tweet is called term frequency for that term.

**Term Frequency- inverse document frequency (TF-IDF)**

There are two rating of TF-IDF phase regularity and inverse papers regularity.

**Support Vector Machine**

SVM is supervised machine learning approach which analyzes input data gives classification. Given set of input sentiments, it analyzes them and group in to positive, negative or neutral sentiments. Like linear classification SVM can efficiently perform nonlinear classification also this is called as kernels trick that map inputs into high features space.

SVM constructs one or more hyper planes which can be used for classification, regression. If one hyper plane is not sufficient to classify the data additional hyper planes are considered for input sentiments classification. The width of the hyper plane shows its strength for classification. This width is called as margin. Less margin leads to poor classification where as big margin gives better classification. All input sentiments data points are called as vectors and the vector points which are very close to the margin from both side are called as support vectors.

**SVM benefits**

* SVM has regularization parameter which is useful to avoiding over fitting.
* It can model nonlinear decision boundaries.
* It is robust for high dimensional space.
* It can model real world problems like text classification bio-informatics analysis.
* The generalization quality and training to SVM is better than many traditional methods.

The performance of SVM is decided by the classification rate or error rate. We may consider the time performance to indicate how fast it provides result for given set of data.

New sentiments are they mapped in to the same space and category in which they may fall is predicted.

**Working of SVM**

**Identify correct hyper plane)**

**Case 1:**

We have three hyper planes P Q R now sentiments can be classified using star and circle as follows

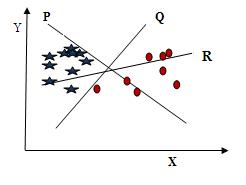


Fig. 2(a) SVM Based Classification Using Three Hyper Planes

Here we will select the hyper plane which will devide to classes of sentiment better. Hyper plane Q is excellent for this moto.

**Case 2:**

We have three hyper planes P, Q and R. all these are dividing the classes for input sentiments

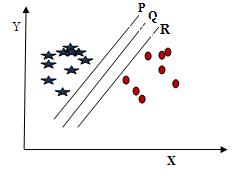


Fig. 2(b) SVM Based Classification Using Three Hyper Planes

We can maximize the distances between nearest sentiments data points to decide correct hyper plane for sentiment classification. These distance is called as margin as shown in following figure

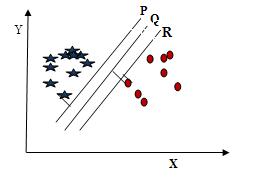


Fig. 2(c) SVM Based Classification Using Three Hyper Planes

In above figure hyper plane Q has high margin as compared to hyper plane P and R. so hyper plane Q is correct for better classification of sentiments. If we select low margin hyper plane, there is more changes of miss classification of sentiments.

**Case 3:**

Consider there are two hyper planes P and Q used for classification of sentiments here hyper plane P has better margin than hyper plane Q. but SVM gives preference to accurate classification of sentiments than maximizing margin between data points. Hyper plane P has classification error where as hyper plane Q classifies sentiment correctly.

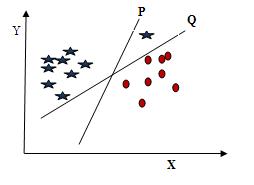


Fig. 2(d) SVM Based Classification Using Three Hyper Planes

**Case 4:**

Sometimes we may not have linear hyper plane between classes for classifying sentiments.

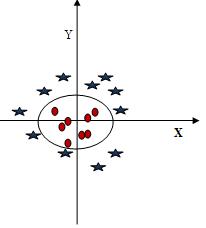
****

Fig. 2(e) SVM Based Classification Using Three Hyper Planes

SVM will introduce additional feature z = x2 + y2 .The new plot for sentiments data points will be on axis x and z.Or values for z will be positive as z = x2 + y2.So red color sentiments data point will be close to the origin of x and y axis and star shaped sentiment data points will be away from the origin with higher value of z.

**Sentiment Evaluation**

The performance of different text clasiifer is evaluated by using some performance measures. Precision, recall gives us the relevant and retrieved sentiments. True (TP), true negative (TN), false positive (FP) and false negative (FN) are important performance measures**.**

* + Precision=TP/(TP+FP)
  + Recall=TP/(TP+FN)

The most important averages are micro average, which count search document equally important, and macro average, which count search category equally important.

**DISCUSSION AND CONCLUSION**

We can compare priority based features for the given input sentiment data. Based on the manual priority we have selected, the sentiment analysis result will vary. Also we can compare the output of SVM for n-grams. Here we will consider the cases up to 3 gram. For each case, the time required for getting the polarity for a given set will vary. Finally n gram approach and priority based feature selection can be merged to get more results for further sentiment analysis. We hope the proposed architecture will give a better analysis of the sentiments which will be proved fruitful for business decisions.

**REFERENCES:**

|  |  |
| --- | --- |
| [1] | B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?Sentiment classi¯cation using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79{86, 2002}.) |
| [2] | Turney, Peter (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". Proceedings of the Association for Computational Linguistics. |
| [3] | M. Hu and B. Liu,”Mining and Summarizing Customer Reviews”, Proceedings of the 10th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, (2004). |
| [4] | P. Chaovalit and L. Zhou, “Movie Review Mining: A Comparison between Supervised and  Unsupervised Classification Approaches”, In System Sciences, HICSS'05, Proceedings of the 38th Annual Hawaii International Conference on IEEE, **(2005)**, pp. 112c- 112c. |
| [5] | T. O‟Keefe and I. Koprinska, “Feature Selection and Weighting in Sentiment Analysis”, Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia. (2009). |
| [6] | A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, 2010, pp. 1320-1326 |
| [7] | Efthymios Koulompis, Theresa Wilson, Johanna Moore: Twitter Sentiment Analysis: The Good the Bad and the OMG!.In: Proceeding of the Fifth International AAAI Conference on Weblogs and Social Media, 2011 |
| [8] | F. M. F. Wong, S. Sen, and M. Chiang, “Why Watching Movie Tweets Won’t Tell the Whole Story?,” *Arxiv preprint arXiv:1203.4642*, no. c, p. 6, Mar. 2012. |
| [9] | Hassan Saif, Yulan He and Harith Alani: Semantic Sentiment Analysis of Twitter. In: Proceedings of the 11th International Semantic Web Conference, 2012. |
| [10] | Gann W-JK, Day J, Zhou S. (2014). Twitter analytics for insider trading fraud detection system. Presented at second ASE international conference on Big Data. |
| [11] | Jensen MJ, Jorba L, Anduiza E. Introduction. In E.Anduiza, M. Jensen, & L. Jorba (Eds.), Digital media and political engagement worldwide: A comparative study. New York, NY: Cambridge University Press,2012, 1-15 |
| [12] | Kothari, A. A., & Patel, W. D. (2015). A Novel Approach Towards Context Based Recommendations Using Support Vector Machine Methodology.Procedia Computer Science, 57, 1171-1178. |
| [13] | Abinash Tripathy, Ankit Agrawal, Santanu Kumar, “Classification of Sentimental Reviews Using Machine Learning Techniques,” *3rd International Conference on Recent Trends in Computing*, *2015* |
| [14] | Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE) , Issue 1, Volume 2 , January 2015 |
| [15] | Github:https://github.com/aababu/sentiment-analysis/blob/master/wordstrength.txt |
| [16] | Sourceforge:https://sourceforge.net/projects/java-ml/files/java-ml/ |

**BIBLIOGRAPHY OF AUTHORS**

|  |  |
| --- | --- |
|  | Prakash P.Rokade has received his B.E.degree in Computer Pune University, Maharashtra; Indiain 2005.He has received his M.Tech. degree Computer Engineering from Bharti idyapeerth, Pune, Maharashtra,India in 2011 and presently pursuing his Ph.D. in Computer Science and Engineering Koneru Lakshmaiah Education Foundation, formerly K L University, Vaddeswaram , Andhra Pradesh, India.Hisr research interest includes Sentiment Analysis, Opinion Mining,and Machine Learning. |
|  |  |
|  | Dr. Aruna Kumari D has received her Ph.D. degree in Computer Science and Engineering from the K L University, Vaddeswaram, Andhra Pradesh,India. Currently, She is Professor Koneru Lakshmaiah Education Foundation, formerly K L University. Her teaching and research areas include in data Mining, Machine Learning and has published more than 50 papers in many National, International journals.She is honoured by DST Young Scientist Award (Govt. of India). |