DOI: 10.21917/ijsc.2012.0044

# SENTIMENT ANALYSIS FOR ONLINE FORUMS HOTSPOT DETECTION

**K. Nirmala Devi[1] and V. Murali Bhaskaran[2]**
[1]*Department of Computer Science and Engineering, Kongu Engineering College, Tamil Nadu, India*
E-mail: k_nirmal@kongu.ac.in
[2]*Paavai College of Engineering, Tamil Nadu, India*
E-mail: murali66@gmail.com

***Abstract***
***The user generated content on the web grows rapidly in this emergent information age. The evolutionary changes in technology make use of such information to capture only the user's essence and finally the useful information are exposed to information seekers. Most of the existing research on text information processing, focuses in the factual domain rather than the opinion domain. Text mining plays a vital role in online forum opinion mining. But opinion mining from online forum is much more difficult than pure text process due to their semi structured characteristics. In this paper we detect online hotspot forums by computing sentiment analysis for text data available in each forum. This approach analyzes the forum text data and computes value for each piece of text. The proposed approach combines K-means clustering and Support Vector Machine (SVM) classification algorithm that can be used to group the forums into two clusters forming hotspot forums and non-hotspot forums within the current time span. The experiment helps to identify that K-means and SVM together achieve highly consistent results. The prediction result of SVM is also compared with other classifiers such as Naïve Bayes, Decision tree and among them SVM performs the best.***

***Keywords:***
***Sentiment Analysis, Hotspot, Text Mining, K-means, SVM***

## 1. INTRODUCTION

Data mining is the process of nontrivial extraction of implicit, previously unknown, and potentially useful information from data that can help the businesses to make proactive and knowledge driven decisions. It uses machine learning, statistical and visualization techniques to discover and present knowledge that previously went unnoticed. Opinion mining is an important sub discipline within data mining and natural language processing (NLP), which automatically extracts, classifies, and understands the opinion generated by various users. These techniques also help to enhance the value of existing information resources that can be integrated with new products and systems as they are brought on-line.

The growth of tremendous amount of online information from various forums has made very difficult for the customers to acquire information that are useful to them. This has motivated on the detection of hotspot forums [4] where useful information are quickly made available for those customers which might make them benefit in decision making process. In topic-based classification, topic related words are important.

Efficient statistical and machine learning techniques can be applied to process the enormous amount of online data. An emergent technique called Emotional polarity computation also known as sentiment analysis [5] can also be performed during online text mining. However, in opinion classification, topic-related words are not very important. But, opinion words that indicate positive or negative opinions are important, e.g., great, excellent, amazing, horrible, bad, worst, etc. Most of the methodologies for opinion mining apply some forms of machine learning techniques for classification. Customized-algorithms specifically for opinion classification have also been developed, which exploit opinion words and phrases together with some scoring functions. In this paper we detect the hotspot forums by computing text sentiment analysis.

This method quantifies the user attention on any forum with which hotspot forums can be identified. The proposed work is then integrated with K-means clustering and Support Vector Machine (SVM) algorithm. It optimally groups the forums into two clusters, forming hotspot forums and non-hotspot forums within each time window.

The rest of the paper is structured as follows: Section 2 discusses related works that describes various existing semantic orientation based sentiment classification approaches. The proposed Support Vector Machine along (SVM) algorithm is discussed in Section 3. The experimental results were discussed in Section 4. Finally Section 5 concludes the paper.

## 2. RELATED WORKS

This section focuses various streams of related work such as analysis of review mining, sentiment classification, machine learning techniques for predicting hotspots.

### 2.1 ANALYSIS OF REVIEW MINING

Mining of online reviews has become a flourishing frontier in today's environment as it can provide a solid basis for predicting future events. For example Zhou et al in 2005 [1] has stated that online reviews became more useful and influence the sales as it provides important information about the product to potential consumers.

A multi-knowledge based approach is proposed where WordNet, statistical analysis and movie knowledge are integrated. The experimental results have shown the effectiveness of the approach in movie review mining and summarizing.

Hu et al [3], in his work has proposed a method in which a generated and semantic orientation labeled list containing only adjectives are used for analyzing. Finally it is observed that machine learning is used to depict the interacting structure of reviews.

### 2.2 SENTIMENT CLASSIFICATION

The documents available on the web can be classified based on various metrics including topics, authors, structures, and so forth. Classification based on sentiments has become a new frontier to text mining community. The task of sentiment

classification is to determine the semantic orientations of words, sentences or documents. Most of the early work on this topic used words as the processing unit. Turney et al [3] has introduced the cosine distance in Latent Semantic Analysis (LSA) space as the distance measure, which has lead to better accuracy. An automatic sentiment classification at document level has been done by Pang and Vaithyanathan [6] in which several machine learning approaches are used with common text features to classify movie reviews from IMDB. It has been pointed out that direct marketing is a promotion process which has motivated customers to place orders through various channels [8].

Li, M. Huang, and X. Zhu [5] stated that in order to work for this, one is needed to have an accurate customer segmentation based on a good understanding of the customers, so that relevant product information can be delivered to different customer segments. Thelwall et al. [10] has stated that analyzing Twitter has given insights into why certain events resonate with the people.

It is found that the customers, who are used to having only a limited range of product choices due to physical and/or time constraints, are now facing the problem of information overload. An effective way of increasing customer satisfaction and consequently customer loyalty has been done that has helped the customers identify products according to their interests. This again has called for the provision of personalized product recommendations [7] [9]. Hofmann and Puzicha in their work have used the Latent Class Model (LCM) to circumvent the aforementioned problems.

Paltoglou and Thelwall [2] have explored in their work that incorporating sentiment information into Vector Space Model (VSM) values using supervised methods was helpful for sentiment analysis.

### 2.3 MACHINE LEARNING TECHNIQUES FOR PREDICTING HOTSPOTS

For predicting online hotspot forums two machine learning techniques [4] have been proposed by Nan Li and Dash. It includes K-means and SVM. K-means clustering is applied to achieve a clustering view for all the forums within each time window. The centers of clusters form the hotspot forums. SVM based approach forecasts the hotspot forums for the current time window by using the data from the past time window.

Unlike other learning methods, SVM's performance is related not to the number of features in the system, but to the margin with which it separates the data.SVM achieves a clustering result by exactly classifying each forum as either hotspot forum or non-hotspot forum.

## 3. PROPOSED WORK

The proposed work helps in detecting hotspot forums and achieves highly consistent results by applying an efficient optimization algorithm with SVM. Fig.1 depicts the conceptual diagram of proposed approach.

### 3.1 PRE-PROCESSING

The data set used in our experimental research is acquired from forums.digitalpoint.com and after data cleaning they are formatted to 37 different forums and 1616 threads. The data collection is initiated by crawling the forum names of first 50 forums. The parsed forum names are then stored in a table.

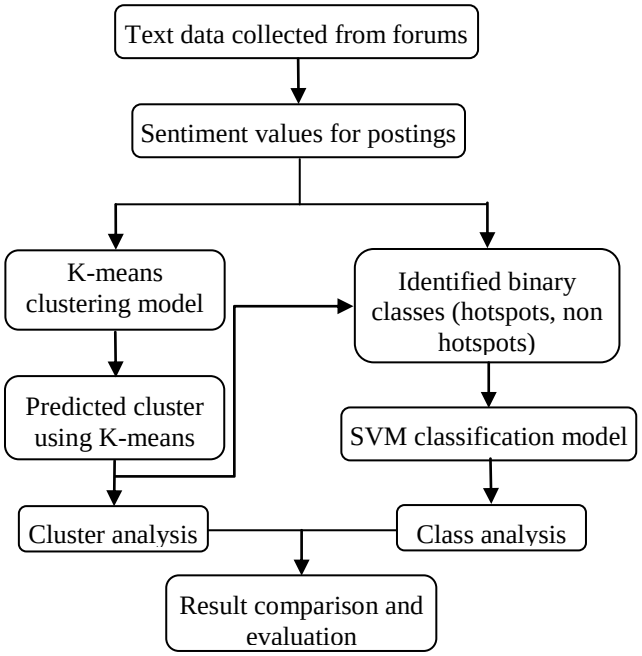


Fig.1. Conceptual diagram of the proposed approach

Then all the thread posts and the reply posts contained in the corresponding web pages are parsed and they are stored separately in a table. After crawling process is achieved data cleaning is done where noise data and irrelevant data are manually removed. Noise data include forums with picture postings that are not clearly shown online. Irrelevant data are from forums where the posting contents are not related to the forum threads at all. The threads that have no replies and the forums that have no threads across the time window are also removed. The data before cleaning and after cleaning are listed in Table.1. Finally after cleaning, 37 forums are narrowed down within the time span from January to October and each time window is of a month length over the year 2011.

Table.1. Data view before cleaning and after cleaning

| | **Before cleaning** | **After cleaning** |
|---|---|---|
| **Time period** | 2007 Jan to 2011 Oct | 2011 Jan to 2011 Oct |
| **Number of forums** | 50 | 37 |
| **Number of threads** | 2430 | 1616 |
| **Number of replies** | 39239 | 19370 |

### 3.2 FEATURE EXTRACTION

The pre-processing work is followed by feature extraction process. For each forum five features are extracted across each time window such as the number of threads, the average number of replies of threads, the average sentiment value of threads, the fraction of positive threads among all the threads and the fraction of negative threads among all the threads. Sentiment value for each thread can be calculated by computing text sentiment.

### 3.3 SENTIMENT COMPUTATION ON FORUM TEXT

Feature extraction includes text sentiment analysis which aims at calculating an integer value for each piece of text. It is a semantic orientation based approach where the sentiment values for all keywords are added to achieve the sentiment value for the whole article. The replies of thread are decomposed into a set of keywords. For each keyword a sentiment value is assigned. The sum of the sentiment values for all the keywords will give the sentiment value for the thread. Suppose for a thread t, its replies are decomposed into a set of key words. For each key word $w_i$ (i=1, 2,...,n) let the sentiment value be $s_i$. Then the sentiment value $S_t$ of the thread t can be calculated as using Eq.(1)

$$st = \sum_{i=1}^{n} si \, . \tag{1}$$

Calculation of sentiment value is based on SentiStrength. SentiStrength is an algorithm for text sentiment analysis that helps in estimating the sentiment values for texts.

### 3.4 FORUM CLUSTERING USING K-MEANS

K-means is the simplest unsupervised learning algorithm that solves the well-known clustering problem. The procedure follows a simple and easy way to group a given data set through a certain number of clusters. The main idea is to define K centroids and it should be placed in a cunning way because different initial centroids cause different result. So, the better choice is to place them as far away as possible.

After the features are extracted clustering can be carried out using K-means algorithm in Rapid miner tool. Each forum may be represented as a datapoint in a vector space. During the feature extraction process a vector is used to represent the emotional polarity of any forum and it is composed of five elements: the number of threads, the average number of replies of threads, the average sentiment value of threads, the fraction of positive threads among all the threads and the fraction of negative threads among all the threads. These datasets are given as the input to the k-means clustering where a clustered view of all the forums is obtained. The hotspot and non-hotspot forums being obtained, within each time window are those closest to the theoretical centers of clusters.

### 3.5 FORUM CLASSIFICATION USING SVM

Support Vector Machine is a state of art classification algorithm and it is known to be successful in a variety of applications. It outperformed most of the other classification algorithms in text categorization tasks. The hot spot forecasting can be carried out using Support Vector Machine (SVM) algorithm. It forecasts the clustering view of the forums in a sliding window manner and those results will be compared with the K-means. The standard SVM takes a set of input data and it optimally predicts, for each given input, which of two possible classes comprises the input. It is employed to realize hotspot forecasting. In order to forecast the hotspot forums within the current time window the clustering result obtained by K-means approach from the previous time window is used. SVM performs forum classification iteratively and tries to find the optimized solution. A well trained SVM is utilized to carry out the prediction for the next time window by inputting the data from the current time window.

Suppose the current time window is 'ti', if a forecast for 'ti+1' is expected, we first train the SVM by inputting vectors of 'ti' and setting the output as the clustering result for 'ti' by K-means. Then the trained SVM generates classification outputs for data of 'ti'. Finally, SVM result is compared with the result of K-means of 'ti+1'.

For each SVM, the input is a forum's representation vector and the optimized output is achieved by classifying each forum as either hotspot forum or non-hotspot forum. The accuracy in predicting hotspot forums is improved with the proposed model and the consistency of the model is validated for its performance.

## 4. EXPERIMENTAL RESULTS

The data that we have collected for our empirical studies are from forums.digital point.com. A list of posts in the form of threads and replies has been crawled from January 2007 to October 2011. The data view before and after cleaning is depicted in Table.1. After cleaning the data are narrowed to 37 forums from January 2011 to October 2011 and then the features are extracted that includes computing sentiment values for threads.

The feature extraction is then followed by K-means clustering and classification using Support Vector Machine (SVM) among the 37 leaf forums for each time window in 2011. Clustering and classification is done using Rapid miner tool. The classification model for forums from forums.digitalpoint.com is shown in Fig.2.

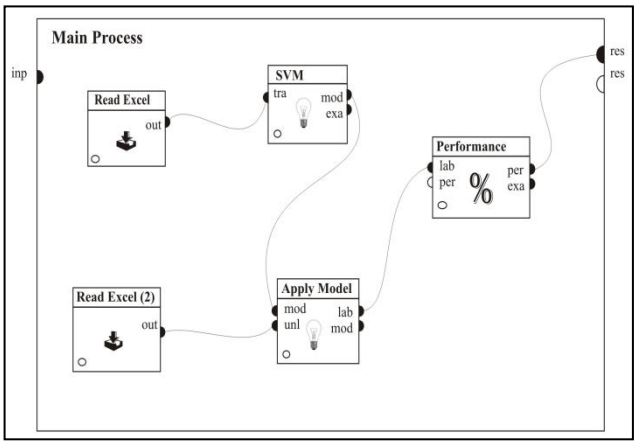


Fig.2. Classification model using SVM

The results that have been obtained using Support Vector Machine (SVM) present a noticeable consistency with the results achieved by K-means clustering. The forums that are most popular among the users based on average number of threads include 'Search Marketing, Publisher Network, adcenter, General Marketing', etc. The forums that are popular based on average number of replies include 'Affiliate Programs-Google, Affiliate Network, Payments, Google-Google+', etc. Table.2 shows the initial data view for user attention that consists of average number of threads and average number of replies for the 37 forums across the 10 time windows.

Table.2. Data view for forums from digitalpoint.com over the time window

| Forum Id | Forum name | Avg. number of threads | Avg. number of replies |
|---|---|---|---|
| 4 | Guidelines / Compliance | 4.0 | 9.475 |
| 5 | Placement / Reviews / Examples | 4.9 | 4.71428 |
| 6 | Reporting & Stats | 4.8 | 14.89583 |
| 7 | Payments | 4.2 | 17.95238 |
| 8 | AdWords | 5.0 | 8.48 |
| 9 | Analytics | 4.2 | 8.61904 |
| 10 | Google-Google+ | 4.5 | 16.28888 |
| 11 | Affiliate Network | 4.8 | 22.22916 |
| 12 | Sitemaps | 4.4 | 9.40909 |
| 13 | Google API | 4.6 | 9.19565 |
| 14 | Product Search | 4.5 | 14.48888 |
| 16 | Publisher Network | 5.0 | 14.48 |
| 17 | Search Marketing | 5.1 | 9.74509 |
| 18 | Yahoo API | 4.9 | 9.26530 |
| 20 | AdCenter | 5.0 | 11.82 |
| 21 | All Other Search Engines | 4.6 | 16.82608 |
| 23 | Solicitations & Announcements | 3.8 | 9.07894 |
| 24 | ODP / DMOZ | 4.8 | 12.20833 |
| 26 | General Marketing | 5.0 | 10.74 |
| 28 | Keywords | 4.6 | 7.97826 |
| 29 | Sandbox | 4.3 | 8.46511 |
| 32 | Facebook API | 3.0 | 13.4 |
| 33 | Twitter | 3.0 | 14.26666 |
| 34 | Social Network-Google+ | 4.4 | 10.36363 |
| 35 | Link Development | 4.7 | 14.89361 |
| 37 | Digital Point Ads | 4.0 | 9.5 |
| 38 | Google AdWords | 3.3 | 4.18181 |
| 39 | Yahoo Search Marketing | 3.7 | 9.59459 |
| 40 | Microsoft adCenter | 3.6 | 7.75 |
| 43 | Commission Junction | 4.7 | 10.06382 |
| 44 | Affiliate Programs-Google | 4.3 | 23.79069 |
| 45 | Pepperjam | 4.4 | 10.29545 |
| 46 | Azoogle | 3.8 | 13.02631 |
| 47 | Amazon | 4.3 | 14.06976 |
| 48 | EBay | 4.2 | 9.30952 |
| 49 | ClickBank | 4.9 | 14.22448 |
| 50 | Chitika | 4.5 | 9.77777 |

The forums that are mostly identified as hotspots by both K-means clustering and Support Vector Machine (SVM) over the time window from January 2011 to October 2011 are shown in Table.3. The result shown in this section further presents a noticeable consistency between K-means and SVM. Therefore, a strong connection between hot spot distribution and text sentiment for online forum is confirmed by both the techniques. This has verified the hot spot forum detection and forecast with the aid of text sentiment analysis.

Table.3. Forums identified mostly as hotspots by k-means and SVM

| Forum ID | Forum name |
|---|---|
| 11 | Affiliate Network |
| 14 | Google |
| 10 | Google+ |
| 6 | Reporting & Stats |
| 49 | ClickBank |

## 4.1 PERFORMANCE EVALUATION

The consistency between K-means and Support Vector Machine (SVM) algorithms is validated using five metrics. They are accuracy, sensitivity, specificity, positive predictive value and negative predictive value. A set of these five metrics are applied for each time window which are defined as follows.

$$\text{Accuracy} = TP + TN / (TP + TN + FP + FN) \quad (2)$$

$$\text{Sensitivity} = TP / (TP + FN) \quad (3)$$

$$\text{Specificity} = TN / (TN + FP) \quad (4)$$

$$PPV = TP / (TP + FP) \quad (5)$$

$$NPV = TN / (TN + FN) \quad (6)$$

where, TP denotes the number of forums that are estimated as hotspots by both K-means and SVM.

TN denotes the number of forums that are estimated as non -hotspots by both K-means and SVM.

FP denotes the number of forums that are estimated as hotspots by SVM whereas non-hotspots by K-means.

FN denotes the number of forums that are estimated as non-hotspots by SVM whereas hotspots by K-means.

Using Eq.(2)-(6), the performance is evaluated for each time window. The time windows are those that are used in SVM classification process. Table.4 suggests that the proposed classification algorithm gives an optimized accuracy result than that of the other classification algorithms.

Similarly the performance can be evaluated using other four metrics and the results can be compared. Fig.3 shows a graphical view of the accuracy result.

Table.4. Comparison of accuracy when using different algorithms and proposed SVM

| Time window | Accuracy (%) | | |
|---|---|---|---|
| | Using Naïve Bayes | Using Decision Tree | Using SVM |
| 2 | 64 | 80 | 84 |
| 3 | 60 | 54.1 | 60 |
| 4 | 61.54 | 60 | 61.54 |
| 5 | 96.54 | 99 | 99 |
| 6 | 60 | 62.22 | 60 |
| 7 | 84.38 | 80.99 | 81.25 |
| 8 | 68.57 | 64 | 65.71 |
| 9 | 90 | 93.1 | 94.59 |
| 10 | 48.65 | 58.2 | 60 |

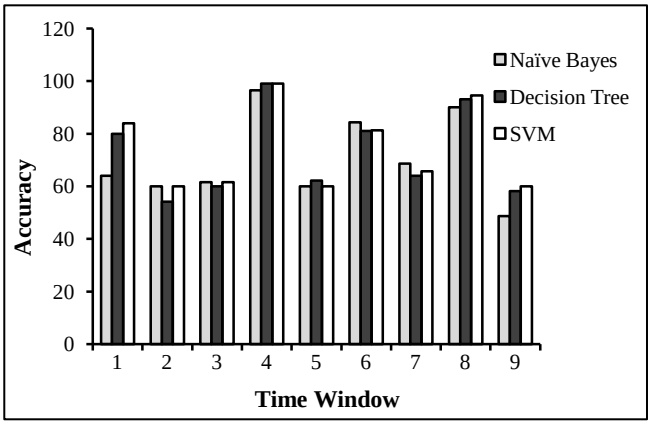


Fig.3. Accuracy comparison while using SVM and other algorithms

## 5. CONCLUSION

This paper proposes a new approach for predicting hotspot forums. In this approach emotional polarity of the text is obtained by computing a value for each piece of text. After calculating the sentiment values the method is then integrated with K-means clustering and SVM classification algorithms for forums cluster analysis. Computation indicates both K-means and SVM produce consistent grouping results. The hot spot based semantic engine can aggregate the content in the forums to determine whether stories on a particular company are positive, negative or neutral.

Using this hotspot predicting approaches the marketing department understands what their specific customers' timely concerns regarding the goods and services. Thus the efficient detection of hotspot forums based on sentiment analysis might make internet social network members benefit in the decision making process.

The results generated from this proposed approach can also combined with market basket analysis to yield comprehensive decision support information. Further work also is done regarding supervised learning algorithms other than Naïve Bayes, decision tree and SVM.

## REFERENCES

[1] Chaovalit P and Zhou L "Movie review mining: a comparison between supervised and unsupervised classification approaches", *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

[2] Paltoglou G and Thelwall M, "A study of information retrieval weighting schemes for sentiment analysis", *Proceedings of the 48th Annual meeting of Association for Computational Linguistics*, pp. 1386–1395, 2010.

[3] Hu M and Liu B "Mining and summarizing customer reviews", *Proceedings of ACM Transactions on Knowledge and Data Mining*, pp.168-177, 2004.

[4] Nan Li and Desheng D Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast" *Decision Support Systems*, Vol. 48, No. 2, pp. 354–368, 2010.

[5] Fangtao Li, Minlie Huang and Xiaoyan Zhu, "Sentiment analysis with global topics and local dependency", *Proceedings of 24th AAAI Conference on Artificial Intelligence*, pp. 1371–1376, 2010.

[6] Pang B, Lee L, and Vaithyanathan S, "Thumbs up? Sentiment classification using machine learning techniques", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 79-86, 2002.

[7] Popescu A and Etzioni O, "Extracting product features and opinions from reviews", *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 339-346, 2005.

[8] Sindhwani S and Mellville A, "Document-Word Co-regularization for Semi-supervised Sentiment Analysis", *Eighth IEEE International Conference on Data Mining*, pp. 1025 – 1030, 2008.

[9] Thelwall M, Kevan B, Paltoglou G, Cai D and Kappas A, "Sentiment strength detection in short informal text", *Journal of the American Society for Information Science and Technology*, Vol. 61, No. 12, pp. 2544–2558, 2010.

[10] Thelwall M, Buckley K, and Paltoglo G, "Sentiment in Twitter Events", *Journal of the American Society for Information Science and Technology*, Vol. 62, No. 2, pp. 406–418, 2011.