Rough Set and Entropy based Feature Selection for Online Forums Hotspot Detection

K. Nirmala Devi
Kongu Engineering College
Perundurai
Tamil Nadu, India

V. Murali Bhaskaran, PhD
Principal
Dhirajlal Gandhi College of Technology
Salem, Tamil Nadu, India

ABSTRACT
The exponential growth of web arouses much attention on public opinion. The rapid progress of online forums, micro blogs and new reports are having large volume of public opinion information. These are proving to be extremely valuable resources in helping to anticipate, detect and forecast societal events. But most of the online data is unstructured or semi structured and that is difficult to decipher automatically. Therefore, it is very much essential to analyze in time and understands the trends of their opinion correctly. The hotspot detection is one of the promising research areas in web mining and it helps to make appropriate decision in timely manner. Feature selection is an essential component in text categorization to identify the relevant features and reduces the dimensionality of data to gain the improved higher accuracy. The proposed system integrates rough set approach with entropy for detecting the online forums hotspot. The experimental results demonstrate that the proposed hybrid feature selection method outperforms with Naïve Bayes and Support Vector Machine based hotspot detection models.

General Terms
Feature Selection, Classification, Accuracy, Web mining

Keywords
Hotspot, Opinion, Sentiment Analysis, Rough Set, Entropy, Naïve Bayes, Support Vector Machine

1. INTRODUCTION
Sentiment analysis or 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 users. These techniques are easily integrated with existing information resources to enhance the values as well as to promote the new products. Rapid growth of web and information age arouses much attention on public opinion and most of these are in unstructured and semi structured format. It is difficult to decipher automatically as well as it is very much difficult for the customers to acquire information that are useful to them. This has motivated on the online forums hotspot detection, where the useful information are quickly made available to the customers which might make them benefit in decision making process.

Feature selection [1] [2] is a process of obtaining the significant minimal subset of attributes based on certain constraints in order to eliminate the unwanted and irrelevant features. It is an essential component in text categorization to identify the relevant features and reduces the dimensionality of data to gain the improved higher accuracy. It also guides to choose the optimal features with correct numbers which may not be too many parameters or too less parameters. Since, too many parameter selections shows duplicates as well as noise information and too less parameter selection may losses the relevant attributes. The major advantage of feature reduction is to decrease the runtime and increases the accuracy.

Recently, rough set theory gaining popularity to generate decision rules [2] [3] for various decision making activities. The relevant, essential and significant minimized attributes are obtained to make intelligent decision in financial and economic forecasting [4] [5]. Furthermore, the rough set is being used for extracting the rules and those rules have a high impact to expose new patterns in the available data.

Entropy [6] [7] based feature selection is used to select the significant features with respect to class attribute and it does not require additional information regarding the distribution of data. The aim of this research paper is to examine the detecting power of various features for online forums hotspot detection based on the hybrid feature selection.

The proposed system integrates rough set approach with entropy for detecting the forum as hotspot or not in the specified time slot. The experimental results demonstrate the proposed hybrid feature selection method outperforms with improved accuracy.

The rest of the paper is organized as follows. Section 2 presents a brief overview of hotspot detection techniques. The outline of the proposed system is presented in the Section 3. The experimental results and analysis is represented in the Section 4. Section 5 concludes the proposed paper.

2. RELATED WORKS
Sentiment Analysis also called opinion mining [8] is the field of study analyses people's opinions, sentiments, evaluations, appraisals, attitudes and emotions towards entities. The inception and rapid growth of social media such as micro blogs, blogs, forums [9], and twitter have large volume of opinionated data recorded in digital form. These social media are proving to be valuable resources to analyze in decision making process. Mining of online reviews has become a flourishing frontier in today's environment as it can provide a solid basis for predicting future events [10]. Therefore, internet public opinion monitoring and analyzing [11] have become a hot issue in recent years.

The author Liu [11] has proposed a method that incorporates sentiment information into Vector Space Model (VSM) was helpful for sentiment analysis. Most VSM uses the TFIDF for representing vectors and TFPDF [12] [13] also used for representing the vectors with improved accuracy. But the conventional approaches used in the information retrieval such as TFIDF and TFPDF alone might not give good results

for forum data. Since the document features in forums are sparse nature. In this paper, a novel method to detect the hotspot forums with integrated sentiment features.

Community Question Answering [9] [14] has recently become a popular social media where users can post questions on any topic of interest and get answers from enthusiasts. It can help users focus on the most popular products or events and track their changes by exploiting hot topics and analyzing trend of a specific topic over a time. The opinion mining is also applied in the micro blogs and twitter for forecasting the stock price movements [15] [16] as well as movie success of the box office [17].

The rough set theory is a dominant mathematical tool introduced by Pawlak [18] [19] to handle imperfect, incomplete or noisy data. It is used to identify the data dependencies for knowing the significant attributes to reduce the dimensionality. The irrelevant and insignificant attributes are eliminated with minimal information loss. It describes a method to represents the data to perform decision making and it is widely used in many areas such as attribute relevant analysis, attribute reduction, attribute dependencies checking etc., due to the following merits.

• The rough set only requires the actual data as its universe of discourse and no extra data is needed as like fuzzy sets, where membership function is required for representation.

• It includes both quantitative and qualitative data

• In order to make decisions the decision rules are generated

• Noise free significant and reliable minimum rules are generated without redundancy and these rules are supported by realistic information

A rough set is associated with a pair of precise sets, called lower and the upper approximation. The objects surely belongs to the set is in lower approximation and the possibly belonging elements are in the upper approximation. The boundary region for the rough set is the difference between lower and the upper approximation of the rough set. In rough set, the data is represented in information table. The row in the table represents an object and column shows the features and class label is called as decision attributes. In the proposed case, the row contains the forum records; column shows the features gathered from the historical forum data decision attribute is whether forum is hotspot / not hotspot.

3. PROPOSED SYSTEM

The proposed system aims to examine the detection power of features for online forums hotspot based on the integrated rough set approach with entropy feature selection. There are six major processes in the proposed system as noted below.

1. Data Preprocessing and Transformation

2. Granulate conditional and decisional attributes by entropy

3. Apply rough set theory to generate reducts and core

4. Extract rules and perform training

5. Forecast based on extracted rules

6. Evaluate the performance

3.1 Data Preprocessing and Transformation

The data set used in our experimental research is acquired from forums.digitalpoint.com [20] [21] [22] and after data cleaning they are formatted to 111 different forums and 10789 threads. The data collection is initiated by crawling all the URL links of 142 forums and its links are stored in the data base. Then all the topic posts and the comment posts contained in the corresponding web pages and their links are parsed and they are stored in the data base. After crawling process is achieved data cleaning is done where noise data and irrelevant data are 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. Finally after cleaning, 39 forums are narrowed down within the time span from January to December and each time window is a half month length (i.e., Fifteen days duration) over the year 2011. The data before cleaning and after cleaning are listed in Table 1.

Table 1. Data view before cleaning and after cleaning

	Before Cleaning	After Cleaning
Time period	2007 Jan to 2011 Dec	2011 Jan to 2011 Dec
Number of forums	142	111
Number of threads	46749	10789
Number of replies	420281	316173
Number of views	769319	548766

In this process, the forum data base, this contains date, forum id, threads, posts / replies, views used as initial indicators. Using above five indicators to find the useful features such as TFIDF, TFPDF, AVG Sentiments etc are highly related to hotspot forums detection. The extracted features of forums are shown in the Table 2.

Table 2. Extracted features of forum

Feature Name	Feature Name
F1 - TFIDF	F9 – Num. of Negative Replies
F2 - TFPDF	F10 – AVG Threads
F3 – TSFISF	F11 – AVG Replies
F4 – Num. of Threads	F12 – AVG Views
F5 - Num. of Replies	F13 – AVG Sentiments Replies
F6 – Num. of Views	F14 – Proportion of Positive replies
F7 – Num. of Sentiment Replies	F15 – Proportion of Negative Replies
F8 – Num. of Positive Replies	

3.2 Granulate Conditional and Decisional Attributes by Entropy

The proposed system uses fifteen extracted features as conditional attributes (C) and hotspot / not hotspot is employed as decisions attribute (D). This process granulates the conditional and decisional into a linguistic value for

further process. Therefore, the decision attribute is granulated in to two linguistic values (Hotspot (L1), Not hotspot (L2) and the conditional attribute is granulated with five linguistic values (Lowest Low (L1), Low (L2), Normal (L3), High (L4), Highest high (L5)). Table 3 shows the numerical intervals for the linguistic values and based on this, membership function for decision attribute is generated. Table 4 contains the five linguistic values and their corresponding numeric ranges for the conditional attribute of Num. of Threads. The partial records for conditional and decision attributes are shown in the Table 5.

Table 3. Parameters of Decision Attribute

Linguistic Value	Universe of Discourse		
	LB	Midpoint	UB
Not Hotspot (L1)	5	10	25
Hotspot (L2)	20	35	90

Table 4. Parameters of Conditional Attribute – Num. of Threads

Linguistic Value	Universe of Discourse		
	LB	Midpoint	UB
Lowest Low (L1)	1	10	20
Low (L2)	15	30	40
Normal (L3)	35	50	60
High (L4)	55	70	80
Highest High (L5)	75	90	100

Table 5. Partial records of Conditional and decision attribute

Time Slot	F1	F2	F3	…	F14	F15	Decision Attribute
T1	L3	L4	L4	…	L3	L2	L2
T2	L3	L3	L4	…	L5	L3	L2
. .	. .	. .	. .	. .	. .	. .	. .
T23	L2	L2	L1	…	L2	L1	L1
T24	L4	L3	L2	…	L3	L3	L2

The lowest entropy indicates the highest discriminative information of the class and at the same time the high entropy having a low discriminative power. The entropy values associated for features of forum in Fig. 1. The features with lowest entropy are selected. The top 10 indictors are used by the rough set to generate the reducts.

3.3 Apply Rough Set to Generate Reducts and Core

This process applies rough set theory to construct decision rules from linguistic values. The top 10 features selected by entropy are used by the rough set to generate the reducts. The Table 6 shows the generated reducts. The core is generated form reducts and it contains {F2, F3, F4, F6, F11, F13, F14, and F15}

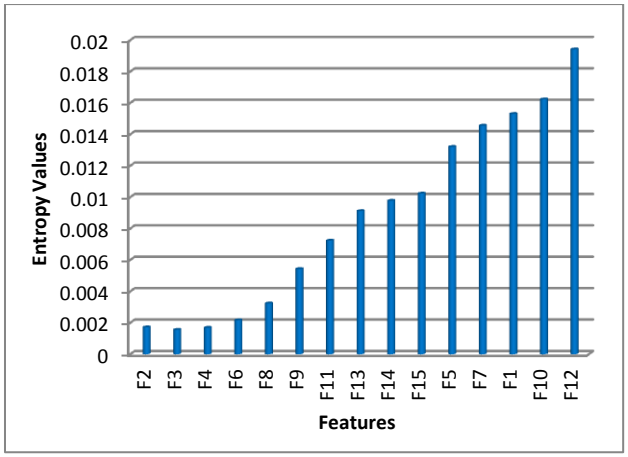

Fig 1: Entropy values for features of forums

Table 6. Generated reducts

Reduct No	Reduct
1	{F2, F3 ,F4,F5,F6, F8, F9, F11, F12, F13, F14, F15}
2	{F1, F2, F3 ,F4,F6, F8, F9, F10, F11, F13, F14, F15}
. .	. .
10	{F2, F3 ,F4, F5, F6, F8, F9, F10, F11, F13, F14, F15}

3.4 Extract Rules and Perform Training

The generated rules are supported with the objects of universe and are represented in the **Table 7**. The rule support specifies the number of objects meet the generated decision rules in the stock data set. For example, Rule 1: If F2 = L3 and F3 = L4 and F4 = L3 and F6 = L4 and F8 = L3and F9=L3 and F11=L2 and F13 = L4 and F14 = L3 and F15 = L2 then Decision = Hotspot forum and it indicates that there are twenty one objects that meet the criteria of Rule 1.

Table 7. Partial rules generated by rough set

Rule No	Rules	Support
1	F2 = L3 and F3 = L4 and F4 = L3 and F6 = L4 and F8 = L3and F9=L3 and F11=L2 and F13 = L4 and F14 = L3 and F15 = L2	21
162	…	2

3.5 Forecast Based on Extracted Rules

This process uses the rules from Table 7 to perform matching between forum objects and the rules extracted from the training set for predicting the future status of forum as hotspot or not hotspot. The 'if' part contains conditional attributes and 'then' part indicates the decision attribute.

3.6 Evaluate the Performance

The performance the proposed system is evaluated with accuracy, sensitivity and specificity. The sensitivity measure shows the correctly classified hotspot instances and specificity shows the correctly classified not hotspot instances. In order to calculate the performance measures equations (1), (2), (3) and (4) are used.

$$Accuracy = \frac{No.of\ Correctly\ classified\ instances}{Total\ instances} \quad (1)$$

$$Sensitivity = \frac{No.of\ Correctly\ classified\ Positive\ instances}{Total\ instances} \quad (2)$$

$$Specificity = \frac{No.of\ Correctly\ classified\ Negative\ instances}{Total\ instances} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\ Actual_t\ -\ Predicted_t\ )^2} \quad (4)$$

4. EXPERIMENTAL RESULTS

The proposed system uses fifteen features as conditional attributes (C) and hotspot / not hotspot is employed as decisions attribute (D). This paper proposes a hybrid technique based on entropy and rough set approach to select the suitable features for detecting the forum as hotspot or not. First, the features are derived from the historical forum data. Second, the entropy method is used to select the relevant features from the extracted features. Next, the selected features are then applied with the rough set based approach to obtain the minimum reducts for finding the hotspot of the forum. Finally, the performance of the proposed approach result is evaluated and compared with the others. The experimental results of proposed system and others are shown in the Table 8.

Table 8. Result of performance measures

Parameters	Without Feature Selection (Rough Set and Entropy)		With Feature Selection (Rough Set and Entropy)	
	NB	SVM	NB	SVM
Accuracy	70.41	79.16	72.81	82.26
Sensitivity	67.85	74.61	71.21	77.12
Specificity	61.22	66.42	66.34	70.14
RMSE	4.74	3.34	4.35	2.84

The Fig. 2 shows the performance measures of proposed feature selection with NB and SVM. The experimental results demonstrate that the proposed hybrid feature selection method integrated with Naïve Bayes (NB) and Support Vector Machine (SVM) outperforms with improved accuracy in hotspot detection than the methods without feature selection.

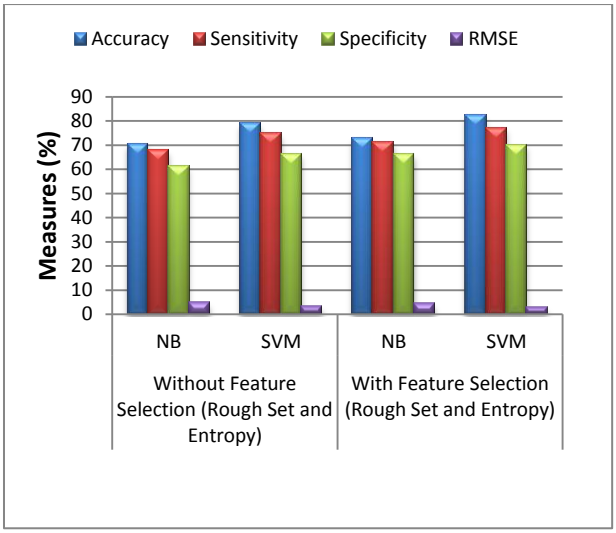

Fig 2: Performance measures of proposed with others

The experimental results in Fig. 2 indicate that the SVM with RST and Entropy method can achieve higher accuracy rate than regular SVM algorithm. The result shows the improvement of accuracy level of SVM with feature selection from 79.16% to 82.26%. When comparing the accuracy of this with SVM there is a 3.1% accuracy improvement is obtained. The same way when comparing the accuracy of proposed with Naïve Bayes there is a 9.45% accuracy improvement. This shows the proposed SVM with RST and Entropy method better than Naïve Bayes.

5. CONCLUSION

The proposed system proposes a novel approach to examine the detection power of features for online forums hotspot based on the integrated rough set approach with entropy based feature selection. It uses fifteen features as conditional attributes (C) and hotspot / not hotspot is employed as decisions attribute (D). The experimental results demonstrate that the proposed hybrid feature selection method outperforms with Naïve Bayes and Support Vector Machine based hotspot detection models. Thus the efficient detection of hotspot forums based on sentiment analysis with hybrid feature selection might make social network members benefit in decision making process. Further this can be extended to identify the hot topic based events from the detected hotspot forums.

6. REFERENCES

[1] Bo-Tsuen Chen, Mu-Yen Chen, Hsiu-Sen Chiang and Chia-Chen Chen. 2011. Forecasting Stock Price Based on Fuzzy Time-Series with Entropy-Based Discretization Partitioning, Springer-Verlag Berlin, pp.382–39.

[2] Richard Jensen and Qiang Shen. 2009. New Approaches to Fuzzy-Rough Feature Selection, IEEE Transactions on Fuzzy Systems, vol. 17, no.4, pp.824-838.

[3] Hameed. A,Qaheri, Hassanien A.E, and Abraham. A. 2008. A Generic Scheme for Generating Prediction Rules Using Rough Set, Neural Network World, vol.18, no.3, pp.181-198.

[4] Wang Ruizhong. 2012. Analyses the Financial Data of Stocks Based Rough Set Theory, In Proceedings of Eighth International Conference on Computational Intelligence and Security, pp. 387-390.

[5] Francis E.H. Tay, Lixiang Shen. 2002. Economic and financial prediction using rough sets model, European Journal of Operational Research, vol.141, pp.641–659.

[6] Chung-Ho Su,Tai-Liang Chen, Ching-Hsue Cheng and Ya-Ching Chen. 2010. Forecasting the Stock Market with Linguistic Rules Generated from the Minimize Entropy Principle and the Cumulative Probability Distribution Approaches, Entropy, vol. 12,pp. 2397-2417.

[7] Salim Lahmiri. 2014. Entropy-Based Technical Analysis Indicators Selection for International Stock Markets Fluctuations Prediction Using Support Vector Machines, Fluctuation and Noise Letters, vol.13, no.2, pp.1450013-1 to 1450013-16.

[8] Liu, B. 2012. Sentiment Analysis and Opinion Mining, Morgan & Claypool publishers, San Rafael, USA.

[9] Nirmala Devi. K and Murali Bhaskaran. V. Text Sentiment Computation for Online Forums Hotspot Detection, International Journal of Information and Communication Technology – Inderscience, 2015, in press .

[10] Peng. W. 2012. Predicting Collective Sentiment Dynamica from Time Series Social Media, in Proceedings of the Confrenece WISDOM'12.

[11] Liu, H. 2010. Internet public opinion hotspot detection and analysis based on Kmeans and SVM algorithm, in Proceedings of the Conference of Information Science and Management Engineering – ISME – 2010, pp.257-261.

[12] Bun, K.K and Ishizuka, M. 2002. Topic extraction from news using TF * PDF algorithm, in Proceedings of the 3rd International conference on Web Information Systems Engineering, pp.73-82.

[13] Chen,K., Luesukprasert,L. and Chou, S. 2007. Hot topic extraction based on timeline analysis and multidimensional senetence modeling, IEEE Transactions on Knowledge and Data Engineering, pp. 1016-1025.

[14] Zhang, D. and Li, F. 2011. QuestionHolic: hot topic discovery and trend analysis in community question answering systems, Expert Systems with Applications, vol.38, no. 6, pp.6949-6855.

[15] Khoza, M. and Marwala, T. 2011. A rough set theory based predictive model for stock prices, in Proceedings of CINTI 12th IEEE International Symposium on Computational Intelligence and Informatics.

[16] Nirmala Devi,K. and Murali Bhaskaran, V. 2015. Forecasting Indian Stock Market Using Particle Swarm Optimization and Support Vector Machine, International Journal of Applied Engineering Research, vol.10, no.1, pp.1891-1900.

[17] Pang,B. Lee,L. and Vaithyanathan, S. 2002. Thumbs Up? Sentiment classification using machine learning techniques, in Proceedings of the Conference on mperical methods in Natural Language Processing, pp. 79-86.

[18] Pawlak. Z. 1991. Rough Sets, Theoretical Aspects of Reasoning about Data, Dordrecht: Kluwer Academic.

[19] Z. Pawlak.Z, Grzymala-Busse.J, Slowinski.R and Ziarko.W. 1995. Rough sets, Communications of the ACM, vol.38, no.II , pp.89-95.

[20] Nirmala Devi K and Murali Bhaskaran V. Sentiment Analysis for Online Forums Hotspot Detection, ICTACT Journal on Soft Computing, vol. 2, no. 2, pp.280-284, 2012.

[21] Nirmala Devi K and Murali Bhaskaran V. Online Forums Hotspot Prediction Based on Sentiment Analysis, Journal of Computer Science, vol. 8, no. 8, pp.1219-1224, 2012.

[22] Digital Point Forums http:// forums.digitalpoint.com