ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Exploiting Movie reviews using Unigram feature Propagation in micro blogging.

Christy A and Meera Gandhi G*.

School of Computing, Sathyabama University, Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 119, India

ABSTRACT

In the era of digital world, Micro logging plays a vital role as a communicator. Micro blogging messages encounter severe challenges in extracting sentiments due to some of its inherent characteristics like size, simplicity and informal writing. Unigram Feature Propagation algorithm is proposed in our article in order to extract opinion from micro blogging website. Features, the main components required for opinion mining are extracted using Chi-Square Keyword Extractor and Key graph Keyword Extractor methods. In turn the above methods are further trained using Machine learning techniques and the accuracy of the results obtained are investigated. From the movie reviews data test, positive and negative reviews are classified. Our experimental results illustrate the Key graph Keyword outperforming Chi-Square Keyword Extractor method. Accuracy of the method is tested by identifying a topic-dependent opinion target feature set. If a word or phrase appears in a sentence, it is extracted as opinion target. The features extracted using Feature Propagation algorithm is manually built for extracting the positive and negative movie reviews using Precision, Recall and Accuracy Metrics. Association Strength between the Successive terms is then calculated using the Scoring method to achieve the effectiveness of every sentence in the document.

Keywords: Unsupervised Feature Propagation, Chi Square, Key graph, Machine Learning, Precision, Recall

***Corresponding author**

INTRODUCTION

In recent times, internet users are making their opinion in buying a product after reading about the opinions and experiences of other internet users. According to a survey, by Bo Pang and Lillian Lee (2008) 81% of Internet users have done online research on a product before buying at least once and 27% of users feel that the reason for these online activities was to get perspectives within their community. In websites such as eopinions.com and amazon.com review information is presented in a semi structured and stereotyped fashion. Blogs in fact can contain subjective content but the desired material within the blogs can vary in content, style, presentation and even in the level of grammaticality. Extracting the key features for opinion mining remains a challenging task.

MOTIVATION

The growth of machine learning methods in Natural Language Processing and Information retrieval has led to the discovery of Opinion Mining. Opinion Mining can otherwise be called as Appraisal Extraction, Affective Computing, etc. According to literature survey, Opinion mining with the usage of list of keywords has shown up to 60% accuracy. Given the training data, its correlation with the positive class can be discovered via a data-driven approach. Applying machine learning techniques based on Unigram models can achieve over 80% accuracy.
Xinjie Zhou et al (2016) have considered #Hashtags# in Chinese microblogs for identifying opinion targets, as they often indicate fine-grained topics. Opinion targets can be classified into two types as Explicit Target and Implicit Target. In this paper, we have tried to extract the opinion targets from Movie reviews. The extraction of opinion targets falls under the categories Explicit and Implicit targets.

1. Explicit Target: In certain reviews, the opinion will be explicitly specified. Consider the following statements
Review 1: It's a damn cute story
Review 2: Solid, Interesting Story
Review 3: Boring to watch
From the reviews stated above Review 1 & 2, specifies Positive feedback directly, whereas Review 3 brings in Negative feedback.

2. Implicit Target: The opinion cannot be interpreted directly from the statements and due to this the possibility of occurrence of False Positives and False negatives can be more. For example, Reviews 4 & 5 indicates positive feedback whereas, Review 6 is negative feedback.
Review 4: The movie doesn't let us go bored
Review 5: His life is ruined. It is a comedy film.
Review 6: It is surprising to see one who actually listens throughout any given scene

Identification of opinion target does not depend only on the Bag of words, but with the important features of the document.
In the proposed work, our process of Opinion Mining involves two stages:
i) Annotation of relevant features by training the features using Machine learning algorithms
ii) Extraction of opinions associated with these features

LITERATURE SURVEY

Pierre P. Senellart and Vincent D. Blondel has reviewed and discussed methods for automatic extraction of synonyms from different kinds of sources such as large corpora of documents, the web, newspapers, etc. They have proposed three methods to discover similar words, such as the straight forward method involving a document vector space model and the Cosine similarity measure [1].

Xiaohui Yan et al (2013) have identified topics within short texts as instant messages and tweets by capturing the document-level word co-occurrence patterns l2]. Chris Clifton and Robert Cooley (2000) has treated document as a collection of entities. This is done by using natural language technology to extract entities from the document [3]. By identifying frequent item sets, clustering is done based on documents inter relations. Kathy Lee et al (2011) has identified topics from twitter messages using bag of words and supervised learning techniques [4]. Xinjie Zhou et al (2016) has presented an opinion mining system for opinion target extraction and opinion summarization using Chinese micro blogs called CMiner. Topics were extracted using hash tags by extracting the noun phrases in each sentence [5].

Jiaqi Zhu et al (2016) have developed an algorithm to detect personalized and abnormal behaviors of Internet users. It is based on the concept that there exist sequential relations in successive documents published by a specific user [6]. The authors have proposed Sequential Topic Patterns (STPs) and formulated the problem of mining User-aware Rare Sequential Topic Patterns (URSTPs) in document streams on the Internet. Naresh Kumar Nagwani (2015) has proposed three cases for similarity measurements between the pairs of documents using Similarity Measurements for Text Processing (SMTP). Case 1 is based on absence and presence of features in the pair of text documents, i.e., the features appearing in both of the documents, second case covers the features appearing in only one document and the third case covers the features appearing in none of the documents[7].

Kang Liu et al (2015) have extracted targets and opinion words from online reviews using graph co-ranking. The graph depicts two types of relations: semantic and opinion relations. A co-ranking algorithm proposed is used to estimate the confidence of each candidate, and the candidates with higher confidence will be extracted as opinion targets/words [8]. Zhen Hai et al (2014), has proposed a method to identify opinion features from online reviews from two different corpora, one a domain-specific corpus and one domain-independent corpus using a measure called domain relevance (DR), obtained by checking the relevance of a term to a text collection. Those terms which are less generic and more domain specific (DR score greater than a threshold) are then confirmed as opinion features [9].

Information Extraction (IE) problems like extracting the key components of reviews as well as documents can be carried out by sub classification and ranking and then ordering the text on the basis of its positivity. Converting the contents of a text into a feature vector or similar representation makes it most salient and important features available for text mining. Documents are represented as a feature vector wherein the entries correspond to individual terms [10][11][12].

Term frequencies using tf.idf is considered as the best metric for topic based test classification whereas presence, a binary valued feature vector in which the entries indicate whether a term is present (value 1) or not (value 0) performs better in polarity classification. Positional of a term in a document can affect the subjectivity status of the enclosing textual unit. Part of Speech tagging considering the noun, verb, and adjectives can play the role of word sense disambiguation. Feature selection using machine learning techniques plays a vital role in Text Mining.

From the literature survey, it is evident that identification of key features plays a vital role in Document summarization, classification and Opinion mining related tasks. This identification of key features can be performed by training with annotated set of features by training them through machine learning algorithms The proposed work follows the Context Aware Hash key segmentation proposed by Xinjie Zhou et al (2016) through the algorithm Unsupervised Label Supervision[2].

METHODOLOGY

Annotating the relevant features plays a vital role in Opinion Mining. Fig 1 shows the architecture adopted for our approach. Since Opinion Mining systems are domain-oriented, a collection of domain dependent documents are considered for the approach. Identification of relevant features is done by separating the words using lexical analysis, whereby the documents can be converted to term-document representation.. Each word is looked up in a lexicon and is assigned a Part Of Speech. Each token is an instance of the type, so the number of tokens is much higher than the number of types. Part-of –Speech (POS) tagging is a module which assigns to each term of a document a Part Of speech (POS) tag. The output is fed into Bag of Words Creator which annotates the words with the corresponding grammar.

For example, in the Bag of Words listed below, all the features used for representing positive reviews are either Noun (NN), Adjective (JJ) , Verb (VB) or Adjectives. It cannot be predicted which form can feature a positive review. A document may signify a positive feedback, which in turn may have negative bag of words. The words such as entertain[VB(POS)], masterpiec[NN(POS)],)],glori[NN(POS)], fascin[JJ(POS)] are positive sounds representing positive reviews whereas unfortun[JJ(POS)], dull[JJ(POS)], depress[VBP(POS)], violenc[NN(POS)], dull[JJ(POS)], reluct[NN(POS, horribli[NNS(POS)] are some of the negative sounds representing positive feedback

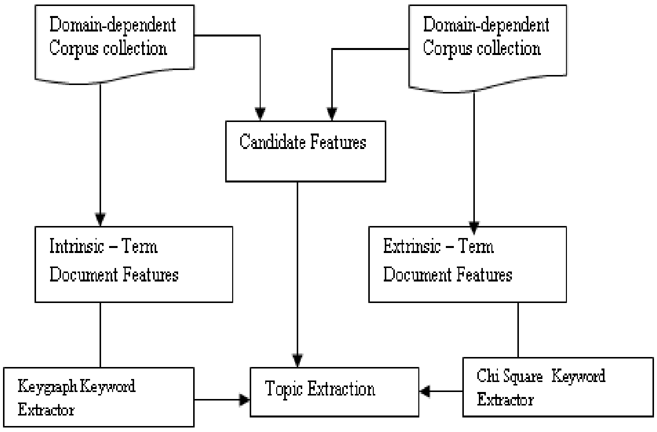
Domain-dependent Corpus collection
Domain-dependent Corpus collection
Candidate Features
Intrinsic – Term Document Features
Extrinsic – Term Document Features
Keygraph Keyword Extractor
Topic Extraction
Chi Square Keyword Extractor

Fig 1: SYSTEM ARCHITECTURE

Wrong choice of features, especially an opinion target word will directly influence the results of Part-of-Speech tagging and Candidate Extraction. By correctly identifying the pattern and adding them to the classifier, we can significantly improve the overall pattern performance. The Chi-square Keyword Extractor and Key graph Keyword Extractor are used to extract the most frequently occurring terms. The Chi-square Keyword Extractor analyzes documents and extracts relevant keywords using co-occurrence statistics. Key graph Keyword Extractor extracts relevant keywords using the graph-based approach. First, the most frequently occurring terms are selected and added as the initial nodes of the graph. The association strength between successive terms is then calculated using the scoring method: association (term1, term2) = min (frequency of term1, frequency of term2) summed for every sentence in the document. All the terms in the graph are rated based on the equation (1).

$$\textbf{Score (t) =} = \sum_{,i=1}^{t} \mathbf{min(freq\,(t), freq\,(w))} \qquad (1)$$

The output table contains (Keyword term, Score, Associated document) tuples. The features thus extracted are passed through the various classification algorithms to study its variance of outliers. The algorithm, the updated version of Unsupervised Label Propagation called as Unsupervised Feature Propagation is depicted in Table 1.

Table 1: Unsupervised Feature propagation algorithm

```
Algorithm: Unsupervised Feature Propagation
Input :
A collection of D documents
Candidate Similarity: $S € R_+^{M X M}$
Prior Document Tagging: $Y_v € R_+^{1 x M}$ for v € V
Filtering Matrix: $F_v € R_+^{M X M for}$ v € V
Threshold: $T^{min}$ and $T^{max}$

Output:
Feature Vector   : $O_v € R_+^{1 x M}$
Tag Cloud
For all v € V do
$O_v \leftarrow Y_v$
```

```
Repeat for all $v \in V$ do
    $D_v \leftarrow \sum_{u \in V, u \neq v} W_{vu} (O_u \times S) \times F_v$
    $O_v \leftarrow P Y_v + P^{const} D_v$
end for
Until convergence
```

Given a collection D documents, our purpose is to get the features required for opinion mining. Features trained and classified are stored in S matrix having M rows and M columns. From each document, sentences are extracted and then they are stored in Term-Document vector, named Y. Features which are less than the threshold T^{min} and T^{max} are filtered, namely F. Features once trained are stored in S, term vector, Documents are preprocessed and each cleaned entry has to be transformed into a numerical vector. From this transformation, we obtain a set of vectors V, with one vector per document consisting of terms v. Since the number of different words to be that appear in a large collection of documents can be quite large, the ***feature selection*** technique is used to select a set of the most relevant words to be used as the dimensions of the vectors.

For all terms v of V, check the terms relevant for opinion mining by checking the similarity with the features tagged (S) and by training with the classification of features obtained through machine learning techniques (F). The set of different terms is much larger and the technique that we use for selecting features consists of first computing the distributions of terms in the documents and then selecting the terms in each document that have the highest probabilities. The entries of the document-term matrix need to be preprocessed, which might involve removing some entries from the sentence, normalizing features, removing special symbols, eliminating stop words and stemming. From this preprocessing we get a collection D of the features extracted. Since the numbers of different features that appear in a large collection of documents are large, a feature selection technique is used to select the most relevant words to be used as the dimension of the vector as Oy and they can be represented in the form of Tag cloud.

RESULTS AND DISCUSSION

In our approach, the identification of feature sets plays a major role in opinion mining. It will directly influence errors especially on opinion target words. Dataset of movie reviews, tagged as positive reviews and negative reviews across various domains are used. The dataset has a total of 1000 positive reviews and 1000 negative reviews. Given the set of documents related to movie reviews as the input to KNIME, the **OPEN NE NLP Tagger** is used which converts a piece of text into a feature vector. Feature vectors are taken as input to the Document Data Extractor, which extracts information from a document into data columns. The Stop words are filtered and the frequent occurrence of unigram features is extracted. Unsupervised Feature propagation algorithm has extracted salient features using Chi-Square keyword extractor and Keyword Key graph extractor methods. It has been ranked according to the order of highest priority and the frequency of occurrences which are represented in the form of Tag Cloud as shown in Fig 2.0 and 3.0.

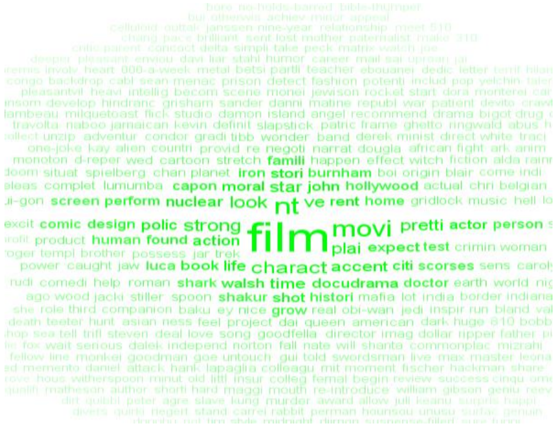

Fig 2: Tag cloud of Chi-Square keyword extractor

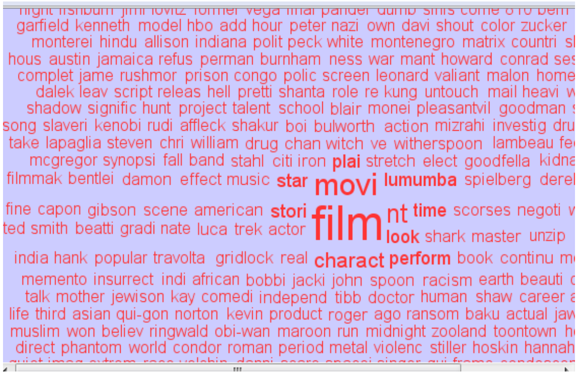
garfield kenneth model hbo add hour peter nazi own davi shout color zucker
monterei hindu allison indiana polit peck white montenegro matrix countri
hous austin jamaica refus perman burnham ness war mant howard conrad
complet jame rushmor prison congo polic screen leonard valiant malon home
dalek leav script releas hell pretti shanta role re kung untouch mail heavi
shadow signific hunt project talent school blair monei pleasantvil goodman
song slaveri kenobi rudi affleck shakur boi bulworth action mizrahi investig
take lapaglia steven chri william drug chan witch ve witherspoon lambeau
mcgregor synopsi fall band stahl citi iron plai stretch elect goodfella
filmmak bentlei damon effect music star movi lumumba spielberg
fine capon gibson scene american stori film nt time scorses negoti
ted smith beatti gradi nate luca trek actor look shark master unzip
india hank popular travolta gridlock real charact perform book continu
memento insurrect indi african bobbi jacki john spoon racism earth beauti
talk mother jewison kay comedi independ tibb doctor human shaw career
life third asian qui-gon norton kevin product roger ago ransom baku actual
muslim won believ ringwald obi-wan maroon run midnight zooland toontown
direct phantom world condor roman period metal violenc stiller hoskin hannah

Fig 3: Tag Cloud of Keyword Keygraph extractor

The features thus extracted are trained using Machine Learning techniques and experiments results have shown KEYGRAPH being the efficient classifier than the Chi-square method and the error rates obtained by Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are depicted in Table 2.0.

Table 2: Error rate using Classification method

	Chi-Square Keyword Extractor		Keygraph Keyword Extractor	
Method	**MAE (%)**	**RMSE (%)**	**MAE (%)**	**RMSE (%)**
Zero R	16.9432	23.691	5.1628	6.9577
Linear Regression Model	17.9413	25.0108	5.4722	7.0743
CVParameterSelection	16.9432	23.691	5.1628	6.9577
RegressionByDiscretization	16.9432	23.691	5.1628	6.9577
Decision Table	16.9432	23.691	5.1628	6.9577

Accuracy of the method has been tested by identifying a topic-dependent opinion target feature set. If a word or phrase appears in a sentence, it is extracted as opinion target. The features extracted using Feature Propagation algorithm is manually built for extracting the positive and negative movie reviews using Precision, Recall and Accuracy Metrics and the results before applying Machine learning techniques as well as after applying machine learning techniques are obtained and they are depicted in Table 3.

$$Precision = \frac{tp}{tp+fp} \qquad (2)$$

$$Recall = \frac{tp}{tp+fn} \qquad (3)$$

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \qquad (4)$$

Movie ratings come at a greater level of precision (like 4.3) than the individual users. Every individual user who really thinks that a film is worth 4.3 has to pick 4 or 5, but its average movie review ratings score could well be 4.3 or 4.5. If the rating categories available to the user are indeed too common, this would show up in the relationship with the movie score: movies with an average score of 4.5 would be less predictable that movies with an average score of either 4 or 5. To test this conjecture, the linear regression models tested repeatedly on two subsets of the data: one comprising the movies with an average positive review of 4.2 and negative review of 3.9.The fit of the regression for the first group was better than for the second (RMSE of 23.691 vs. 6.957) when compared with Chi-square keyword extractor and Key graph keyword extractor.

Table 3: Performance Metrics of Positive and Negative Reviews

	Before Machine Learning Technique			After Machine Learning Technique		
	Precision (%)	**Recall (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**Accuracy (%)**
Positive Review	0.42	0.34	0.38	0.48	0.38	0.42
Negative Review	0.39	0.32	0.35	0.43	0.30	0.32

The features which forms the basis of opinion mining are selected based on the manually assigned threshold and it frequency using Chi-square and Key graph method are depicted in Fig. 4 and the outliers are estimated using linear regression model and it shows 80% of features being skewed in one direction as in Fig. 5. Error rate calculation for different classifiers

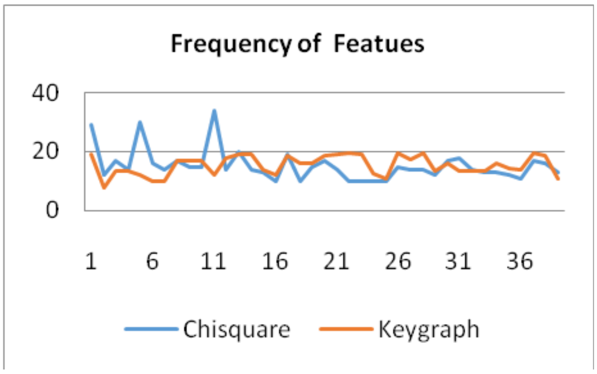

Fig 4: Frequency of Features

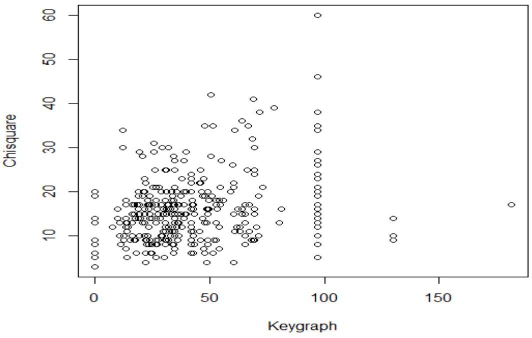

Fig: 5 Features using linear regression

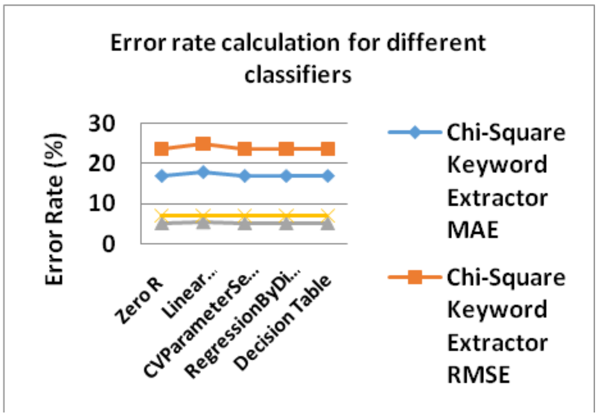

Fig 6: Error rate calculation for different classifiers

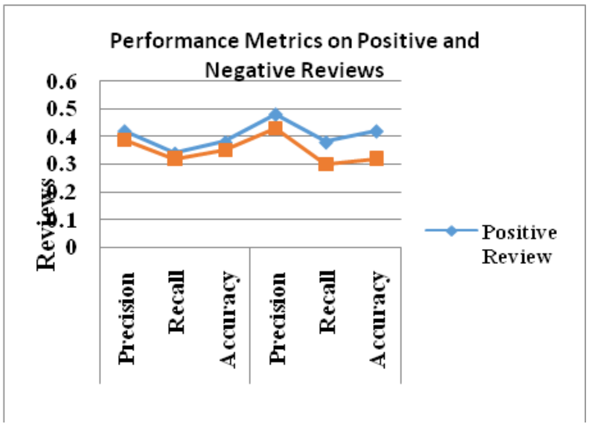

Fig 7: Performance metrics on positive and negative reviews

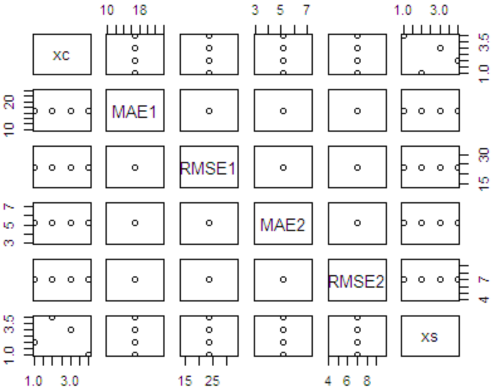

Fig 8: Scattor Plot matrix for error rate

In addition, it was very surprising to see the backoff model perform worse than the unigram multinomial model, it also improves accuracy towards the unigram model, a weight between 0.7 to 0.9 for the unigram model results in the best performance. Scatter plot matrices determine the linear correlation between multiple variables. The variables are written in a diagonal line from top left to bottom right. Then each variable44 is plotted against each other. For example, the middle square in the first column is an individual scatter plot of MAE1 and RMSE1, with MAE1 as the X-axis and RMSE1 as the Y-axis. The same plot is replicated in the middle of the top row. In essence, the boxes on the upper right hand side of the whole scatter plot are mirror images of the plots on the lower left hand. In the Scatter Plot, there is a correlation between Xc and Xs, because the plot looks like a line. Presume probably less of a correlation between MAE1 and RMSE1 in addition to MAE 2 and RMSE2.

CONCLUSION

Feature selection is an important process in our opinion mining. If the exact feature is not extracted the entire process will be in vain. Our algorithm Unsupervised Feature propagation will identify the features and select the relevant features using Chi-square and Key graph Keyword extractor. The features lying within a threshold are accepted as the features used for opinion mining. The keywords are further trained using machine learning technique and selected as the optimum feature sets. The features are trained for the domain under study and the experimental results show the effectiveness of our method. We would extend our algorithm of opinion mining with micro blogs and news forums in future.

REFERENCES

[1] Pierre P. Senellart and Vincent D. Blondel, “Automatic Discovery of Similar Words”, DOI: 10.1007/978-1-
4757-4305-0_2, 2002

[2] Xueqi Cheng, Xiaohui Yan and Jiafeng Guo , “BTM: Topic Modelling Over Short Texts”, IEEE Transactions on Knowledge and Data Engineering, Vol. 26, Issue 12, Pp. 2928-2941, 2014

[3] Chris Clifton and Robert Cooley, “TopCat: Data Mining for Topic Identification in a Text Corpus”, 2000

[4] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary , “Twitter Trending Topic Classification”, 11th IEEE International Conference on Data Mining Workshops, Pp. 251-257, 2011

[5] Xinjie Zhou, Xiaojun Wan and Jianguo Xiao , “CMiner: Opinion Extraction and Summarization for Chinese Microblogs”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol 28, Issue : 7, Pp. 1650-1663, 2016.

[6] Jiaqi Zhu , Kaijun Wang , Yunkun Wu, Zhongyi Hu and Hongan Wang, Mining User-Aware Rare Sequential Topic Patterns in Document Streams”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol 28, Issue 7, Pp. 1790-1804, 2016.

[7] Naresh Kumar Nagwani, “A Comment on “A Similarity Measure for Text Classification and Clustering”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 27, No. 9, Pp. 2589 – 2590, 2015.

[8] Kang Liu , Liheng Xu and Jun Zhao , “ Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING , Vol. 27, Issue 3 ,PP. 636-650 , 2015.

[9] Zhen Hai , Kuiyu Chang , Jung-Jae Kim and Christopher C. Yang, “Identifying Features in Opinion - Mining via Intrinsic and Extrinsic Domain Relevance”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol 26, Issue 3, Pp . 623-634, 2014.

[10] Chien Chin Chen, Meng Chang Chen, “TSCAN : A Content Anatomy Approach to Temporal Topic Summarization”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol. 24, Issue : 1, 170-183, 2012.

[11] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, A. Jaimes, “ Sensing trending topics in Twitter”, IEEE Transactions on Multimedia (pre-print), 2013

[12] Malú Castellanos , “HotMiner: Discovering Hot Topics from Dirty Text”, Survey of Text Mining, Pp 123-157, 2004.

[13] Xia Hu , Lei Tang, “Exploiting Social Relations for Sentiment Analysis in Microblogging,”,ACM 978-1-4503-1869-3/13/02.

[14] Zhao Yanyan, Qin Bing, Che Wanxiang, et al. Appraisal Expression Recognition Based on Syntactic Path [J]. Journal of Software. 2011, 22(5): 887-898

[15] Zhao Yanyan, Qin Bing, Liu Ting. Sentiment Analysis [J]. Journal of Software. 2010, 21(8): 1834-1848