Sentiment Classification and Feature based Summarization of Movie Reviews in Mobile Environment

Savita Harer
Department of Computer Engineering
Dr. D. Y. Patil College of engineering, Talegaon, Ambi
Pune University,

Sandeep Kadam
Professor and head of Computer Engineering Department
Dr. D. Y. Patil College of Engg, Talegaon, Ambi, Pune,

ABSTRACT
A new framework is designed for sentiment classification and feature based summarization system in a mobile environment. Posting online reviews has become an increasingly popular way for people to share their opinions about specific product or service with other users. It has become a common practice for web technologies to provide the venues and facilities for people to publish their reviews. Sentiment classification and feature based summarization are essential steps for the classification and summarization of movie reviews. System proposed Random forest method for sentiment classification of movie reviews. Identification of movie features and opinion words are both important for feature based summarization. System identified movie features using a novel approach called Latent Semantic Analysis (LSA) and frequency based approach. Then system identified opinion words using part-of-speech (POS) tagging method. The result of LSA is extended to LSA based filtering mechanism to reduce the size of review summary. System design focused on the sentiment classification accuracy and system response time.

General Terms
Algorithms, Opinion Mining

Keywords
Movie reviews, Sentiment Classification, Summarization, POS tagging.

1. INTRODUCTION
Sentiment classification is performed to determine the semantic orientation of the movie reviews and movie rating score is based on the sentiment classification result. Random forest method and Support vector machines (SVMs) classifiers are used to perform sentiment-classification task on the movie review data. In addition to the accuracy of the classification, system response time is also taken into account in this system design. The problem statement of system is, in cellular-phone environment, it is inappropriate to display detailed movie review due to the small size of the screen. Thus, system employed summarization technique to reduce the size of review summary. Proposed system summarized the movie reviews into positive reviews and negative reviews classes and display the movie rating based on sentiment classification result [1]. System proposes feature-based summarization method for movie reviews. Product feature and opinion-word identification are essential to feature-based summarization. System proposes a Latent-semantic analysis (LSA) approach and frequency based method to identification of movie features. Then size of review summary reduced by LSA-based filtering mechanism. The main modules of this system are the following:

- Collecting Movie Reviews.
- Sentiment Classification.
- Movie feature identification.
- Opinion Based Identification.
- LSA-based filtering for feature Based Summarization.

2. LITERATURE SURVEY
Now a day's rapid development of ecommerce websites motivate people to express their reviews about product or service as per their interest. Online reviews are very helpful for purchasing any product. But, many reviews are long, which describes their opinion regarding product with few sentences. This makes it hard for other people to judge the quality of the product on sale and decide whether it should be bought or not. Another problem is that if there are large numbers of online reviews then it becomes difficult for manufacturers to maintain a record of customer opinions regarding their products. Therefore system proposed a feature based summarization method for summarization of movie reviews into positive and negative review classes. Most of the existing work is focused on product reviews. But, here system focused on specific domain that is movie review. The movie review mining is different from product review mining [2], reason behind that when a person writes a movie review, he/she comments not only on movie elements such as music, dialogue but also on the related people who contributed to its creation such as the director of the movies, the actors and the actresses. On the other hand, there are specific commented features related to product reviews because people may like some features and dislike others. Due to this, it becomes difficult to classify opinion orientation of reviews as positive or negative. Also, there are many comparative sentences in product reviews. So, movie review mining is comparatively, a more challenging and interesting domain than product review mining. Machine learning techniques for sentiment classification of reviews including Naive Bayes, maximum entropy and support vector machine(SVM), to name a few. Pang and Lee [3, 4] compared the performance of these three machine learning techniques in terms of features. Pang and Lee found that support vector machine (SVM) classifier performs better with feature presence than other machine learning techniques. Authors found that feature presence is more important than feature frequency. System considered not only sentiment classification accuracy but also system response time to design an application in mobile environment. If system will use SVM with feature presence, then it will take more time to load SVM model on a system. Liu, Lu and Jou

[1] found that performance of SVM classifier with feature frequency criteria performs better than feature presence.

3. SYSTEM DESIGN

3.1 Use Case View Diagram

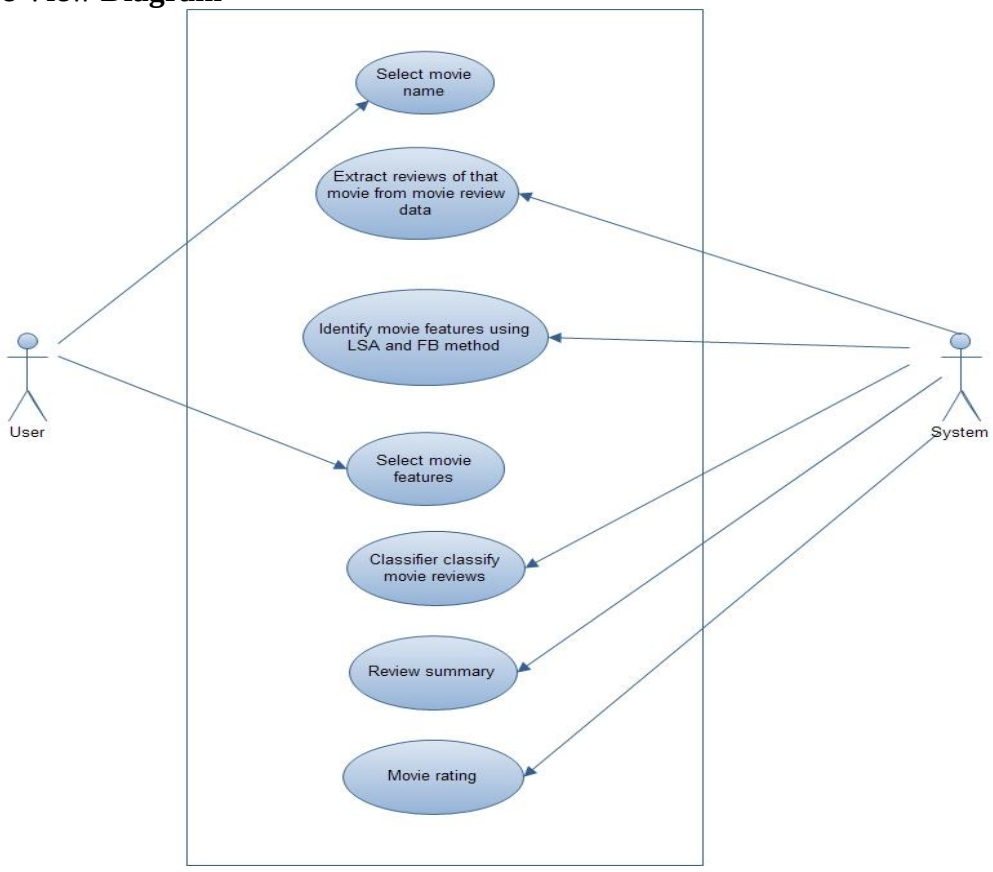


Fig 1: Use case view of movie rating and review summarization

Above Figure 1 shows use case view of movie rating and review summarization system. Here, Actors are users and a system.

3.2 Mathematical Model

Let S be the system,

S = {I, O, Fn, C, S, F}

where,

I = Set of input {M, S, Fe}

M=term-document matrix n*m

n= number of reviews

m=number of terms

S=product feature seed set {S_1, S_2, S_3, S_4, S_5}

S_1 = Scene

S_2 = Plot

S_3 =Director

S_4 =Actor

S_5 =Story

Fe=set of extracted features {f1, f2…, f_K}

k=reduced dimension

O= Output is an association array {AF}

AF= set of {AF_S, AF_P, AF_D, AF_A, AF_{ST}}

where each key represents a product feature seed f and its corresponding value is f's related product feature

Fn= Functions are {f (e1), f (e2), f (e3), f (e4), f (e5), f (svd),

f (avg)}

{e1=i|i is to check whether movie review is available or not} {e2=j|j extract all the reviews of that movie from movie review dataset}

{e3=k|k classify the reviews into positive and negative review classes}

{e4=l|l identify the movie features using frequency based algorithm and LSA based algorithm}

{e5=m|m display the summarized reviews and movie rating score to the end user}

$f(svd) = U\Sigma V^T$

U and V are matrices whose columns and rows are orthonormal vector.

Σ is a diagonal matrix whose diagonal elements are the singular values of M.

$f(avg) = S_f / n$

where,

$$S_f = \sum_{i=1}^{n} \{\text{Frequency } (\text{term}_i)\}^2 \quad (1)$$

C = Constraints are

- Reviews should be available.
- Used for IMDB dataset.
- Collected reviews from rediff website for developing an online application.

S = Success are

- Positive and negative review summary.
- Correct Movie rating.

F = Failure is

- Incorrect movie rating.

4. IMPLEMENTATION

4.1 Dataset

In this paper, system used IMDB dataset consisting of Hollywood movie reviews which are divided into training and testing part. It provides standard training corpus data with 500 positive and 500 negative movie reviews for developing an offline application. In addition to the movie review dataset, system collected Hollywood movie reviews as an online from rediff website without any movie rating information. Then system trains a classification model using IMDB dataset and executes that model on rediff for testing purpose.

4.2 Software Requirements

System used eclipse IDE with JDK 1.7 version on Windows 2007 platform with apache tomcat server and Android plug-ins (ADT) installed with API level 17 for providing further support to the mobile environment.

4.3 Sentiment Classification

Already everyone knows that different classification techniques are used for sentiment classification. System performed sentiment classification using Support Vector Machine technique and Random forest method.

4.3.1 Support Vector Machine

Support vector machine is a supervised machine learning technique. The basic goal of SVM is to find a hyperplane. Hyperplane divide data points into the two classes. The best hyperplane is the one that represents the larger margin between the two classes. System used LIBSVM library [5] to supporting SVM. By default system used SVC-C type of LIBSVM which handles the balancing between training and testing data set. System used radial basis kernel function (RBF) and 5-fold cross validation method for implementation of SVM. System consisting of training dataset as 1000 movie reviews, then divides the dataset into five equal parts. System used cross validation method to found best parameters to train the whole training and testing set. System used LSA and frequency based feature vector as an input to train SVM to remove system response time issue.

4.3.2 Random forest Classifier

Random forest classifier gives better accuracy than other machine learning techniques. As the name suggest, it works as large collection of decorrelated decision trees. It is based on bagging aggregation method. System selected piece of information from specific module called as bootstrap sample. System did not miss the information to the outside world because each and every data point is repeated in decision trees. Then converted the movie review data into ARFF format which makes the list of column name and detailed information about the column. After that system pass that data to the WEKA tool [6]. System integrated LIBSVM library with WEKA tool for implementation of Random Forest classifier. To build classification model system needed two parameters, first is I consisting number of trees in Random forest and second, S that is string value for random number. In this way, random forest classification model developed by system.

4.4 Feature Based Summarization

Feature based summarization is based on movie features and opinion words. System identified movie features using Latent semantic Analysis (LSA) algorithm [1].System constructed term document matrix n*m from dataset. Here n are the number of reviews represented as a row vector and m are number of different words identified in all reviews represented as a column vector. Then, LSA applied Singular value decomposition (SVD) method on term document matrix. SVD is used for dimensionality reduction that separates the term document matrix into three parts that are U, Σ and V. The U matrix is left eigenvectors, Σ is the diagonal matrix of singular values and V is right eigenvectors. Also system identified movie features using frequency based approach. For frequency based algorithm system used term document matrix as it is without a SVD operation. For feature based summarization, identification of opinion words is also very important. System analyzes the polarity of sentence using opinion words. System identified opinion words using part-of-speech tagging method and also frequency information of those words taken into account by using equation (1).

5. EXPERIMENTS

Several experiments are performed to evaluate a system. The movie review dataset includes 500 positive and 500 negative movie reviews.

5.1 Sentiment classification

Table 1 shows the experimental result. System trained a support vector machine classification model using movie review dataset and tested that model on rediff reviews which was collected as an online. For this experiment system considered 3 feature selection criteria's as shown in table1.

Table 1. SVM classification result using movie review dataset

Feature selection criterion	Number of features	Accuracy (%)	Time taken to build SVM model (milliseconds)
Unigrams	5678	97.7	1454
Unigrams with occurrences more than 3	4300	98.4	982
Unigrams using the frequency criterion	335	96.2	383

Table 2 shows the experimental result of Random forest classifier. Here, accuracy of Random forest classifier is better than support vector machine technique. It requires more time

to build classification model than SVM but still it is acceptable because it correctly classified positive movie reviews as positive and negative movie reviews as negative so its accuracy is 100%.

Table 2.Random forest classification result using movie review dataset

Feature selection criterion	Number of features	Accuracy (%)	Time taken to build Random forest model (milliseconds)
Unigrams	5678	100	1412
Unigrams with occurrences more than 3	4300	100	1272
Unigrams using the frequency criterion	335	100	517

5.2 Movie feature identification

In this experiment, system used nouns are the candidates of movie features. System took the seeds that are Scene, Plot, Director, Actor and Story [1]. Table 3 shows top ten movie features for each seed [7].

Table 3. Top ten terms for each seed generated using LSA

Scene	Director	Plot	Actor	Story
scene	director	plot	actor	story
style	mercy	project	scenery	technical
fellow	loses	drawn	interested	opera
pig	challenge	tune	twists	hits
portrayal	perfection	projected	escapes	dragging
breathing	kelly	retains	klondike	characterizations
bogdonavich	broken	post	al	began
actress	summer	understated	pig	novelist
brand	psycho	surrounding	breathing	worthwhile
blue	suffering	lens	goldie	glowing

Table 4. Top ten terms generated using frequency based Approach

Ranking	Terms
1	movie
2	film
3	good
4	story
5	time
6	see
7	people
8	great
9	movies
10	get

Table 4 shows the top ten terms are selected as movie features using frequency based algorithm. Here, all the nouns are ranked according to their frequencies [7].

5.3 Feature Based Summarization

System proposes an LSA-based filtering approach to reduce the size of review summary. System allows the user to choose the feature in which he/she has interest. If user clicks on IMDB checkbox then IMDB movie names are displayed in the "Movie" section spinner menu. Figure 2 shows rediff movie name as Electric City and selected any one features from frequency based, LSA based with reduced dimension 50, LSA based with reduced dimension 500, LSA based with reduced dimension 1000 among these features. When user selects movie name and feature, the system generated positive and negative review summary which is related to movie feature and displays movie rating score as shown in figure 2. Then system calculate movie rating, for example, if there are 20 movie reviews of specific movie and 16 reviews are positive then rating of that movie will be four stars. In addition to movie review data experiment, system employed the movie-review glossary [1], as the basis of comparison. Movie review glossary is created for movie reviewers, films, critics etc. System compared the terms extracted from LSA based and frequency based approaches with the terms filtered in movie review glossary dataset. Then, system took the intersection of movie review data terms and movie review glossary terms as a filtered movie review glossary.

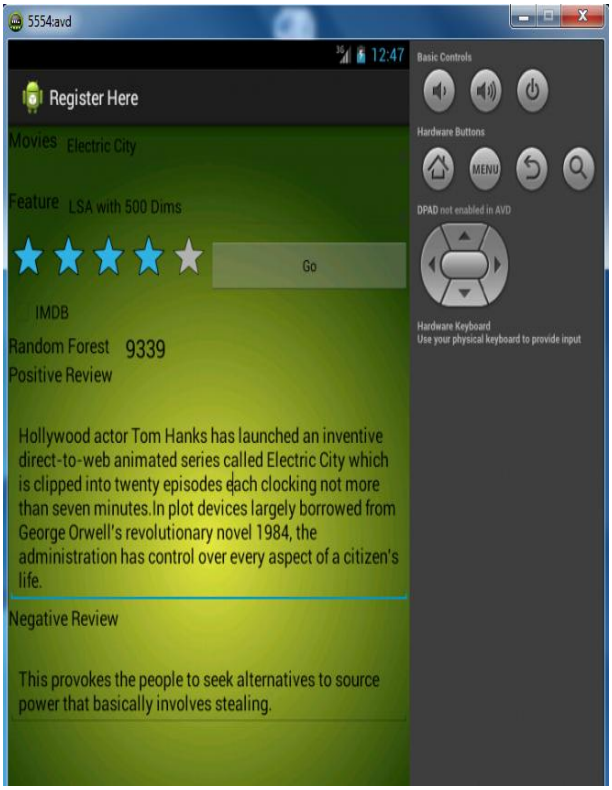
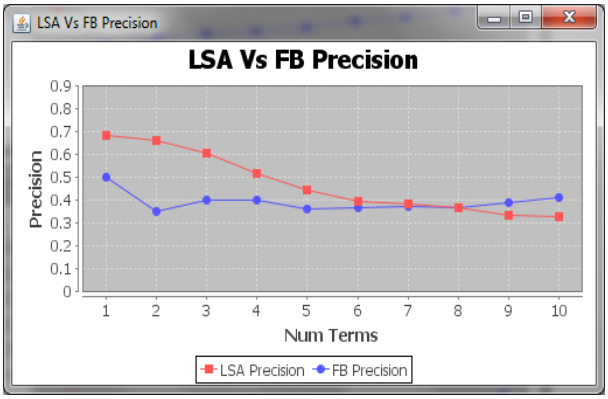
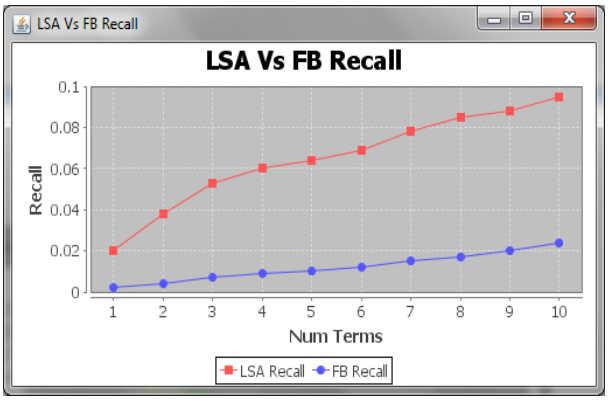


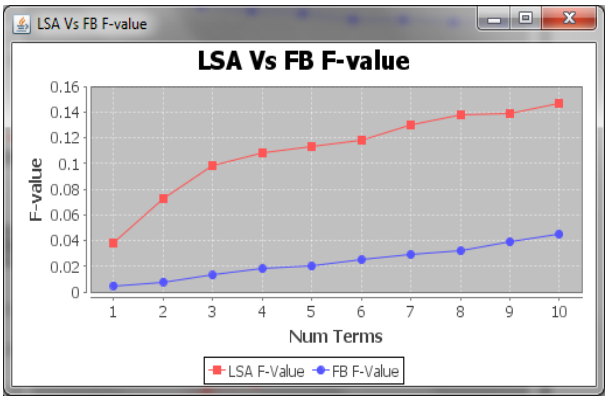
Fig 2: Movie rating and review summarization screenshot



(a)



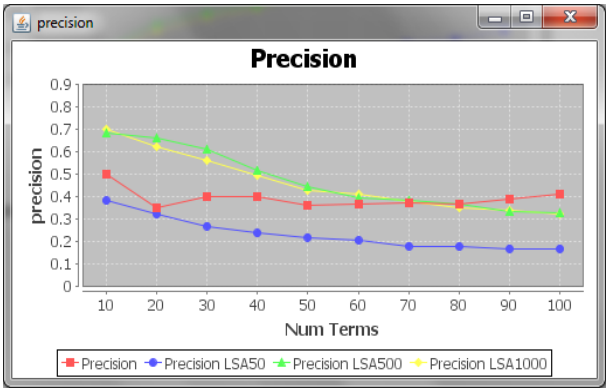
(b)



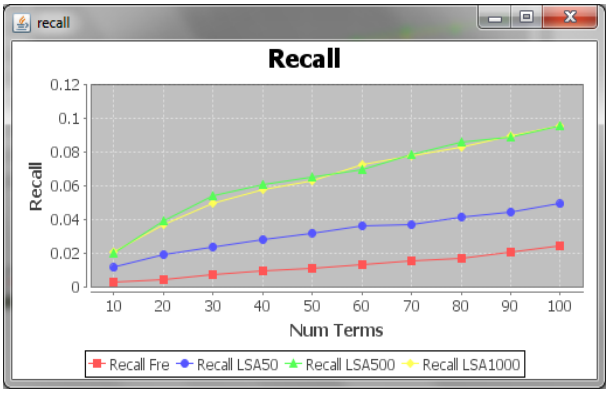
(c)

Fig 3: (a) Precision curve (b) Recall curve (c) F-value curve for movie review glossary dataset

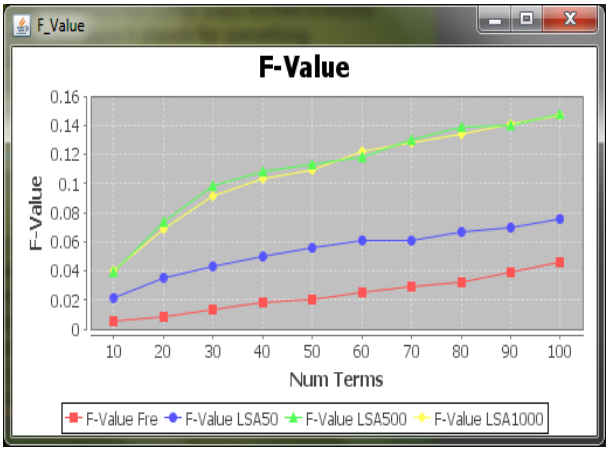
Figure 3 shows the experimental results, here LSA based approach outperforms than frequency based approach by using precision, recall and f-value evaluations evaluation parameters.



(a)



(b)



(c)

Fig 4: (a) Precision curve (b) Recall curve (c) F-value curve for movie review glossary dataset using LSA under different truncated dimensions

In addition to the experiments mentioned above, system conducted experiments on the effect of truncated dimension of LSA. System compared the results of LSA with the frequency based approach. Figure 4 shows the LSA outperforms than frequency based approach when the number of dimensions is more than 500.If the number of dimensions of LSA is 50 then its performance become worse than frequency based approach as shown in precision curve.

Table 5. Time for truncated dimensions of LSA using movie review glossary dataset

Feature	Time taken for fetching results of review summary on mobile system from server (milliseconds)
Frequency based	1036
LSA with reduced dimension 50	5748
LSA with reduced dimension 500	9339
LSA with reduced dimension 1000	5748

System trained a random forest classification model using IMDB dataset and tested that model on rediff reviews. Table 5 shows total time required for fetching the review summary of specific movie feature as per user interest from apache tomcat server to in mobile environment.

6. CONCLUSIONS

Sentiment classification and feature based summarization system in mobile environment is designed and implemented. The experimental result shows that random forest classification model performs better than support vector machine model because it gives better accuracy than other machine learning techniques. System proposed a novel approach called Latent semantic analysis for identification of movie features which outperforms than frequency based approach. The advantage of LSA based approach is that it could be applied to all the languages; it does not need any external dictionary, since LSA is language-independent and it is based on SVD operation. System extended the result of LSA to LSA-based filtering mechanism to reduce the size of movie review summary. The movie rating score is based on the sentiment analysis result. In this way, system implemented as an online and offline in a mobile environment. To develop an online application system used IMDB review dataset for training and testing a model. System used IMDB review dataset for training a model and testing that model on rediff reviews to develop an offline application. System combined movie rating information with review summary and displayed results to the end users. In future, same system design can also be extended to other product-review domains easily.

7. ACKNOWLEDGMENT

I would like to convey a word of gratitude to my guide, Prof. Sandeep Kadam for guiding me throughout the project work and providing me excellent support by valuable guidance.

8. REFERENCES

[1] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou, "Movie Rating and Review Summarization in Mobile Environment", IEEE VOL. 42, NO. 3, MAY 2012.

[2] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. 2006, pp. 43–50.

[3] Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar, Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of EMNLP 2002.

[4] Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar, Thumbs up?: Sentiment Classification using machine learning techniques, In Proceedings of the ACL-02 conference on Empirical Methods in Natural Language, 2002.

[5] (2001), LIBSVM: A library for support vector machines[online].Available:http://www.csie.ntu.edu.tw/cjlin/libsvm.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H.Witten. The weka data mining software: an update. SIGKDD Explor. Newsl, 11(1): pp.10–18, November 2009.

[7] Savita Harer and Sandeep kadam, "Mining and Summarizing Movie Reviews in Mobile Environment," in International Journal of Computer Science and Information Technologies, ISSN: 0975-9646, Vol. 5 (3), 2014.

[8] Andrew L. Maas and Raymond E. Daly and Peter T. Pham and Dan Huang and Andrew Y. and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: HumanLanguage Technologies., pages 142–150, Portland, Oregon, USA, June 2011. ACL.

[9] T. Hofmann, "Probabilistic Latent Semantic Indexing," Proc. 22ndAnn. Int'l ACM SIGIR Conf. Research and Development in InformationRetrieval (SIGIR), pp. 50-57, 1999.

[10] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis,"Mach. Learn., vol. 42, no. 1/2, pp. 177–196, 2001.

[11] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Norwell, MA: Kluwer, 2002.

[12] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in Proc. EMNLP, 2004, pp. 412–418.

[13] T. K. Landauer, P.W. Foltz, and D. Laham, "Introduction to latent semantic analysis," Discourse Processes, vol. 25, pp. 259–284, 1998.

[14] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc.10th ACMSIGKDD Int. Conf.Knowl. Discov.DataMining, 2004, pp. 168–177.

[15] Savita Harer and Yogesh Sayaji, "A Survey On Sentiment Analysis for Movie Domain in Mobile Environment," in International journal of Computer Networking, Wireless and Mobile Communications, ISSN: 2278-9448 , ISSN: 2250-1568,2014.

[16] A. Esuli and F. Sebastiani, "SENTIWORDNET: A publicly available lexical resource for opinion mining," in Proc. 5th Conf. Lang. Res. Eval., 2006, pp. 417–422.