ISSN:0975-9646

Mining and Summarizing Movie Reviews in Mobile Environment

Savita Harer[#1], Sandeep Kadam[*2]

[#]*Department of Computer Engineering, University of Pune*
Dr. D. Y. Patil College of Engineering, Ambi, India

[*]*Professor of Computer Engineering Department, University of Pune*
Dr. D. Y. Patil College of Engineering, Ambi, India

***Abstract*— In this paper, we design and develop various strategies required for sentiment analysis of movie domain in mobile environment. The main objective of review mining and summarization is extracting the features on which the reviewers express their opinions and determining whether the opinions are positive or negative. The sentiment classification is done by various classifiers such as maximum entropy, naive bayes, Support vector machine (SVM) model and, Random forest technique, to name a few. Movie rating score based on sentiment classification result. The movie feature extraction is done by various methodologies such as Latent semantic analysis (LSA) algorithm and Frequency based algorithm. The result of LSA is extended to filtering mechanism to reduce the size of review summary. We design our system by consideration of sentiment classification accuracy & system response time. The same design can be extended to other product review domain easily.**

***Keywords*— natural language processing, movie reviews, summarization, sentiment classification.**

I. INTRODUCTION

Sentiment analysis is a type of natural language processing which tracking the mood of public about particular product or service. In earlier days, when we wanted to purchase any product from the merchants we asked those of our relatives for their opinions who had knowledge about that product. But now days, the Internet compose people to explore for other people's opinions from the different websites before purchasing a product or seeing a movie. Sentiment analysis widely used in business application to determine their product quality and maintaining their reputation in the market. Meanwhile, cellular phones become certainly the most important part of our lives. Since the digital contents displayed on cellular phone is limited because the cellular phones are physically small. Hence, a compressed description of documents will aid the delivery of digital content in cellular phones. This paper discovers and designs a mobile system for movie rating and review summarization in which semantic orientation of comments, the limitation of small display capability of cellular devices, and system response time are considered [3]. We get the opinion of people through search engine along with the different websites. Most of the websites provides user ratings in percentage and search engine monitors best matching web pages according to its pattern. But current search engine does not provide semantic orientation of the content in review. Sentiment classification is done by binary classification. The system will provide summary about the movie reviews. The movie rating depends on sentiment classification result. Section II, describes review of related work. In Section III, system design is presented. In section IV, several experiments are introduced. In section V, the conclusion is presented.

II. REVIEW OF RELATED WORK

TABLE I
SUMMARY OF CLASSIFICATION TECHNIQUES

Sr. no	Author	Feature	Dataset	Classifier	Accuracy
1	Kaiquan Xu (2011)	Linguistic feature	Amazon reviews	Multiclass SVM	61%
2	Xue Bai (2011)	Information gain	Movie reviews	Naive Bayes	92%
3	Yulan He (2010)	Self trained	Movie reviews	Lexical approach	74.7%
4	Zhu Jian (2010)	Odds ratio	Movie reviews	Back propogation	86%
5	Rudy (2009)	Document frequency	Movie reviews	SVM,Hybrid	89%
6	Qinglianng Miao (2009)	POS,Apriori	Amazon reviews	Lexical approach	87.6%
7	Kennedy and Inkpen (2006)	frequencies	Movie review	SVM	86.2%
8	Gamon (2005)	Stemmed terms	Car reviews	Naive Bayes	86%
9	Bai (2005)	Dependence among words	Movie reviews	Two-stage Markov Blanket	87.5%
10	Pang and Lee (2004)	Based on minimum cuts	Movie reviews	Naive Bayes	86.4%

Table 1 shows different classification techniques of product reviews [12].

III. SYSTEM DESIGN

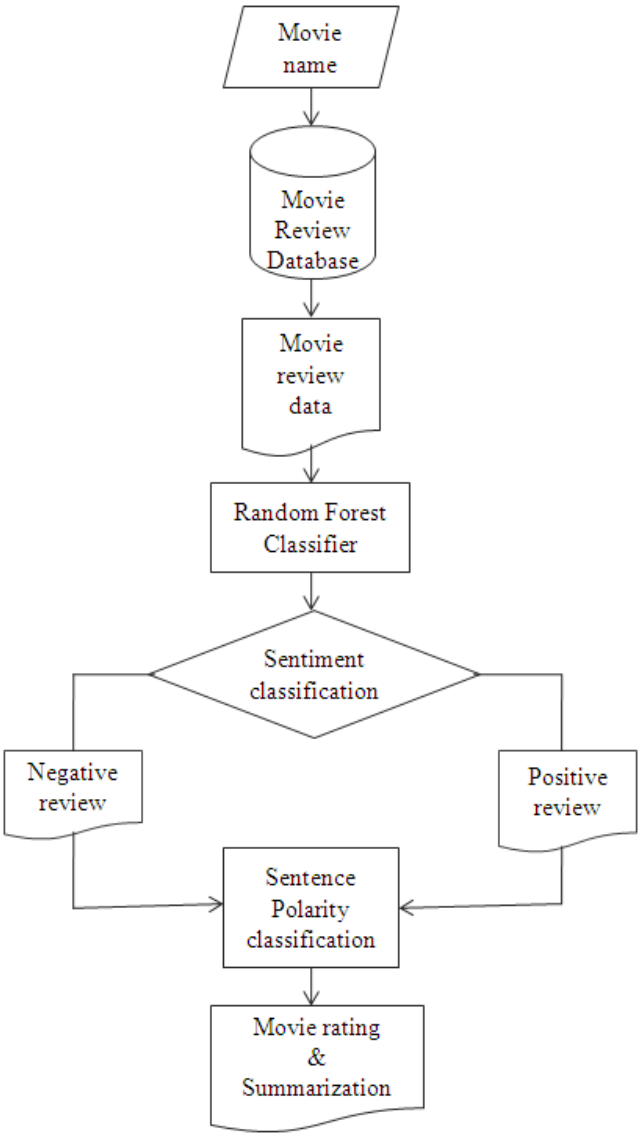


Fig.1 Movie rating and review summarization flow

Fig.1. shows the movie rating and review summarization system flow [3].The input of our system is name of the movie. This input helps in retrieving the movie reviews from IMDB dataset of that movie. Now, extracted movie reviews become the input of Random forest classifier. It classifies the movie reviews into positive review and negative review classes. By using opinion words, we determined the polarity of sentence. The movie rating score based on sentiment classification result. The output of our system is positive or negative movie review summarization depending upon whether the inclination of the opinion is more towards the positive or the negative side.

A. Dataset

In this paper, we used IMDB dataset which consisting collection of Hollywood movie reviews from internet blogs without any rating information. We randomly selected 500 positive reviews and 500 negative reviews as the data for classification-model building.

B. Sentiment Classification

Different machine learning algorithms such as maximum entropy, naive bayes, Support vector machine and, random forest used for sentiment classification of product reviews. We employed following two classifier for classification of movie reviews into positive and negative review classes.

1) Support Vector Machine

Training data contains the positive and negative reviews. Data provided in training classifier uses positive and negative reviews but they do not deal with complicated data which is very hard to classify. The reviews are said to be of low variance if they are giving only positive or only negative opinions otherwise they are having high variance. In the two-category case, the basic idea behind the training procedure is to find a hyperplane, represented by vector w, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible. This search corresponds to a constrained optimization problem. We employed SVM to perform the classification and libsvm package is used in the system. The kernel function used in the system is the radial basis function (RBF) and K-fold cross validation (i.e., K = 5) is conducted in the experiment. The movie rating score is based on sentiment classification result.

2) Random Forest

Random forest algorithm is highly accurate classifier which runs efficiently on large dataset. It works as large collection of decor related decision trees based on bagging technique using weka tool. It handles thousands of input variables without variable deletion. It is an effective classification model for estimating missing data and maintains accuracy when large proportion of the data missing.

C. Feature based Summarization

Feature based summarization focused on the movie features on which the reviewers articulate their opinions. Identification of movie features and opinion words are both essential in feature based summarization. We propose a Latent semantic analysis (LSA) algorithm and frequency based algorithm to identify movie features. We compared these two movie-feature-identification approaches, in the experiment section. In addition to feature identification, opinion words about the movie features are important as well. Hu and Liu‘s [7] work as pioneer on feature extraction algorithm. Liu Bing carried out the successively extraction of frequent features, opinion words and occasional features through the cooccurrence relationship among them [3]. For opinion word identification we considered POS-tagging information and frequency of the opinion words. The result of LSA algorithm is extended to LSA based filtering mechanism is planned to employ the semantically related terms to reduce the size of review summary. LSA based filtering approach allows the users to choose the features in which they are interested. The system will generate summary of positive and negative movie review which is related to movie feature F.

IV. EXPERIMENT

Several experiments are performed to evaluate our system. Our application runs on mobile platform therefore we considered classification accuracy as well as system response time to design a system. Sentiment classification is performed by Support vector machine (SVM) model and Random forest technique. We identified movie features using Latent semantic analysis (LSA) algorithm and Frequency based algorithm and compare LSA based approach with frequency based approach using the movie review glossary dataset.

A. Sentiment classification

TABLE II
SVM CLASSIFICATION RESULT USING IMDB DATASET

Feature selection criterion	**Number of features**	**Accuracy (%)**	**Total time to load SVM model (msec)**
Unigrams	5678	97.7	236
Unigrams with occurrences more than 3	4300	98.4	140
Unigrams using the frequency criterion	335	96.2	449

Table II shows the experimental result, system using unigrams with presence feature will have 5678 features, and it takes about 236msec to load SVM classification model. It is infeasible on mobile platform because we considered both sentiment classification accuracy and system response time. Hence, numbers of features are very important for the system response time. The second one that is unigram with occurrences more than 3 reduces the number of features. In the third experiment number of features are 335 and its accuracy is about 96.2 as well as it takes about 449msec to load SVM classification model on our system. Therefore, frequency as filtering criterion reduces the number of features and it is feasible on mobile platform.

As shown in Table III, sentiment classification of movie reviews performed by Random forest classifier. We compared the results of SVM classification technique with Random forest model. The system using the frequency criterion will have 335 features, and it gives better accuracy that is 100% than support vector machine technique, and it takes about 186 msec to load Random forest classification model on system.

TABLE III
RANDOM FOREST CLASSIFICATION RESULT USING IMDB DATASET

Feature selection criterion	**Number of features**	**Accuracy (%)**	**Total time to load Random forest model (msec)**
Unigrams	5678	100	688
Unigrams with occurrences more than 3	4300	100	586
Unigrams using the frequency criterion	335	100	186

B. Movie feature identification

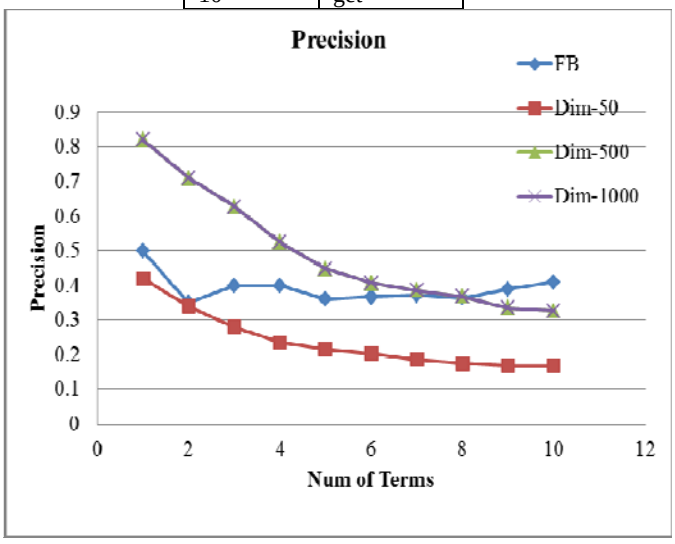
TABLE IV
FIVE ASPECTS GENERATED USING LSA

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

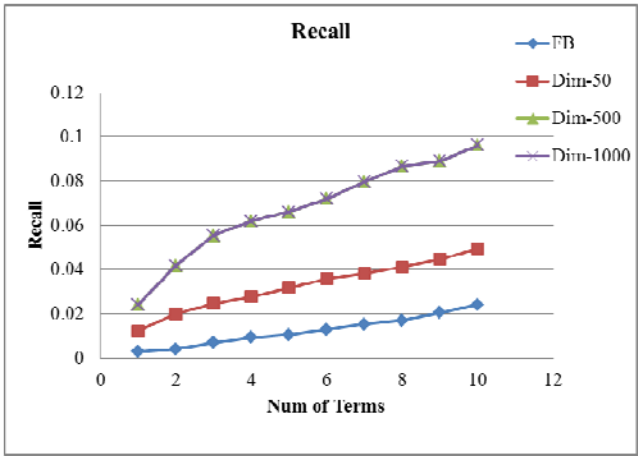
In movie-feature identification, we compared LSA based approach with frequency based approach. We performed experiments using movie review glossary dataset [3]. Latent semantic analysis algorithm is used to identify movie features and the seeds include scene, director, plot, actor, and story. Table IV shows top ten features for each seed can be regarded as being semantically related to the seed.

TABLE V
TOP TEN TERMS USING FREQUENCY BASED APPROACH

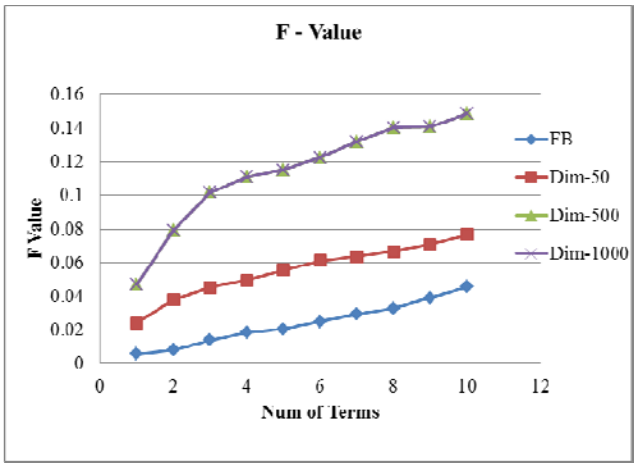
Ranking	**Terms**
1	movie
2	film
3	good
4	story
5	time
6	see
7	people
8	great
9	movies
10	get



(a)



(b)



(c)

Fig.2 Precision, recall, and F-value curves for movie-review-glossary using LSA under different dimensions. (a) Precision curve. (b) Recall curve. (c) F-value curve.

Frequency based approach identify the terms which are used in movie reviews. Above Table V shows the terms using frequency based algorithm. In this approach, all the nouns are ranked according to their frequencies, and then, the top ones are selected as movie features.

In addition to the experiments mentioned above, we further conducted experiments on the effect of truncated dimension of LSA in movie-feature identification using movie review glossary dataset [3]. Above Fig. 2 shows the result, where precision, recall, and F-value curves are presented. Latent semantic analysis (LSA) outperforms than frequency-based approach when the number of dimensions is more than 500.For LSA, differences are minor when the number of dimensions is more than 500.On the other hand, if the number of dimensions of LSA is 50 then performance becomes worse than frequency-based approach.

V. Conclusions

In this paper, we design and implemented sentiment analysis of movie domain in a mobile environment. We implemented support vector machine and random forest classifier for sentiment classification of movie reviews into positive and negative review classes. In a system-performance-analysis experiment, the number of features plays an important role for building classification model on a system. So, we used frequency criteria to reduce the number of movie features, and it can take less time to load the classification models on our system. We determined the polarity of a sentence using opinion words. As a result, Random forest classification technique is the best mean for mining and summarization of movie reviews. It gives better accuracy than other machine learning techniques. In feature-based summarization, identification of movie-features and opinion words plays an essential role. So, we implemented a LSA algorithm and frequency based algorithm which identified features about movie reviews. Furthermore, we propose an LSA-based filtering mechanism to reduce the size of review summary. The movie rating score is based on sentiment-classification results. We combined movie rating information with review summary, and display result to end users with the rating and summarization information about the movie. In future work, our project design can also be extended to other product-review domains easily.

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.

References

[1] 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.

[2] B. Snyder and R. Barzilay, “Multiple aspect ranking using the good grief algorithm,” in Proc. HLT-NAACL, 2007, pp. 300–307.

[3] 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.

[4] Kaiquan Xu , Stephen Shaoyi Liao , Jiexun Li, Yuxia Song, “Mining comparative opinions from customer reviews for Competitive Intelligence”, Decision Support Systems 50 (2011) 743–754.

[5] Kennedy and D. Inkpen, “Sentiment classification of movie reviews using contextual valence shifters,” Computational Intelligence, vol. 22, pp. 110–125,2006.

[6] 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.

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

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

[9] 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.

[10] P.Turney 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceeding of Association for Computational Linguistics, pp. 417--424.

[11] Qingliang Miao, Qiudan Li, Ruwei Dai, “AMAZING: A sentiment mining and retrieval system”, Expert Systems with Applications 36 (2009) 7192–7198.

[12] 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.

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

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

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

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

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