Movie Review Classification and Feature based Summarization of Movie Reviews

Sabeeha Mohammed Basheer[#1], Syed Farook[*2]

[#1] MTech student(CSE),MES College of Engineering,Kuttippuram,India

[*2] Assistant Professor (CSE),MES College of Engineering,Kuttippuram,India

***Abstract*-- Sentiment classification and feature based summarization are essential steps involved with the classification and summarization of movie reviews. The movie review classification is based on sentiment classification and condensed descriptions of movie reviews are generated from the feature based summarization. Experiments are conducted to identify the best machine learning based sentiment classification approach. Latent Semantic Analysis and Latent Dirichlet Allocation were compared to identify features which in turn affects the summary size. The focus of the system design is on classification accuracy and system response time.**

***Keywords*-- LSA, PLSA, LDA, Naive Bayes, Maximum Entropy, SVM**

I. INTRODUCTION

A. *Sentiment Classification*

The task of determining whether a movie review is positive or negative is similar to the traditional binary classification problem. Given a review, the classifier tries to classify the review into positive category or negative category. The classification result will be the basis of the rating. With the proportion of positive and negative reviews, the system could provide the rating information to end users.

B. *Feature based Summarization*

Summarization technique is employed to reduce the size of information. The system will summarize the reviews (including positive reviews and negative reviews) and provide the user an overview about the reviews. A Latent Semantic Analysis (LSA) based feature-identification approach works best to identify features. Features and opinion word identification are essential in feature-based summarization.

LITERATURE SURVEY

A. *Sentiment Classification Methods*

1) Naive Bayes:

One approach to text classffication is to assign to a given document 'd' the class 'c*'.

$$c = \operatorname{argmax}_c p(c \mid d)$$

The Naive Bayes(NB) [5] classifier isderived from Bayes rule

$$p(c \mid d) = \frac{p(c)\,p(d \mid c)}{p(d)}$$

where P(d) plays no role in selecting c. To estimate the term P(d | c), Naive Bayes decomposes it by assuming fi are conditionally independent given d's class:

$$p_{NB}(c \mid d) = p(c)\left(\prod_{i=1}^{m} p(f_i \mid c)^{n_i(d)}\right)$$

Naive Bayes is optimal for certain problem classes with highly dependent features.

2) Maximum Entropy:

Maximum entropy classification [5] (MaxEnt, or ME) is an alternative technique which has proven effective in a number of natural language processing applications. Its estimate of P(c | d) takes the following exponential form:

$$p_{ME}(c \mid d) = \frac{1}{z(d)\exp\left(\sum \lambda_{i,c} F_{i,c} C_{d,c}\right)}$$

where Z(d) is a normalization function. $F_{i,c}$ is a feature/class function for feature $F_{i,c}$ and class c, definedas follows:

$$F_{i,c}(d,c') = ¿\{1\}\{n_i(d)>0\}$$

$$F_{i,c}(d,c) = \{0\}\{otherwise\}$$

Unlike Naive Bayes, MaxEnt makes no assumptions about the relationships between features, and so might potentially perform better when conditional independence assumptions are not met. The underlying philosophy is that we should choose the model making the fewest assumptions about the data while still remaining consistent with it.

3) Support Vector Machine:

Support vector machines [4] (SVMs) have been shown to be highly effective at traditional text categorization, generally outperforming Naive Bayes. They are large-margin, rather than probabilistic, classiffiers, in contrast to Naive Bayes and MaxEnt. 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.

B. Feature Identification Methods

1) Latent Semantic Analysis:

Vector Space Model (VSM) cannot deal with synonymy and polesemy. To address these issues, latent semantic analysis (LSA)[6] has been developed. LSA projects an original vector space or term-document matrix into a small factor space. The dimensional reduction of a matrix is accomplished using singular value decomposition which decomposes an original matrix into three matrixes, a document eigenvector matrix, an eigenvalue matrix, and a term eigenvector matrix. In turn, an original matrix can be approximated by multiplying these three matrixes with only high eigenvalues. Because of orthogonal characteristic of factors, words in a factor have little relations with words in other factors, but words in a factor have high relations with words in that factor.

2) Probabilistic Latent Semantic Analysis

Under the assumption of exchangeability, the occurrence of words can be modeled using probabilistic theory. The probabilistic latent semantic analysis [7] (PLSA) assumes that documents are generated throughout the following three steps. First, a document d is generated or selected with probability P(d). Second, topic z is picked with probability P(z|d). Third, each word w in a topic is generated with probability P(w|z). Then,the probabilities of word-document occurrences, P(d,w), can be represented with $P(d,w)= \sum P(z)P(w|z)P(d|z)$. Using EM algorithm which is a general solution in estimating unknown parameters, PLSA estimates topic probabilities P(z), document probabilities given topics P(d|z), and word probabilities given topics P(w|z).

3) Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [1] incorporates the generative process of documents with Dirichlet distribution. According to LDA process, each document is generated in the following three steps. First, the number of words used in a document is determined by sampling with the Poisson distribution. Second, a distribution over topics for a document is elicited from the Dirichlet distribution. Third, based on the document-specific distribution, topics are generated, and then words for each topic are generated. LDA also provides topics in which words have probability values.

III. IMPLEMENTATION DETAILS

A. Movie Review Classification

Machine learning algorithm would take movie review as input and predict whether the review is negative / positive about the movie based on what was said. This task includes the following steps. First the movie review is converted into ARFF format and then preprocessing is carried out after converting the text field into word vector. Next step is classification using machine learning algorithm with the help of weka[9] . The dataset used to perform sentiment classification consist of 1000 positive and 1000 negative movie reviews available at dataset [3] .

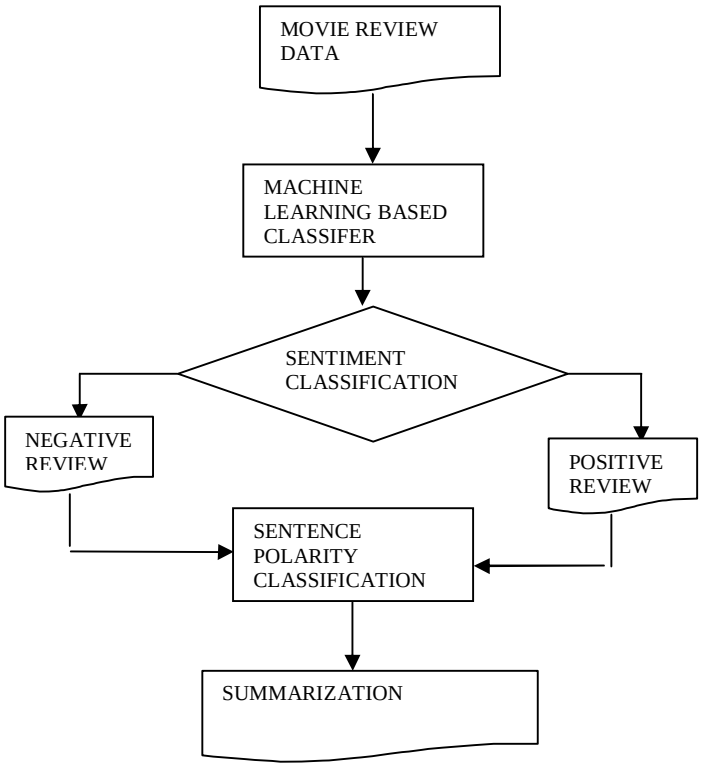

Fig. 1 Movie Review classification and summarization flow

B. Feature Based Summarization

1) Sentence Polarity Detection

After movie review classification separate extracts of the positive and negative aspects of a movie review must be generated. A positive movie review may include negative comments about specific aspects and vice versa. This is done with the help of a knowledge base. An opinion lexicon[8] is used as the knowledge base. Each adjective and adverb in a sentence is compared with the opinion lexicon. If the number of positive opinion words are

greater than the number of negative opinion words, then the sentence can be claimed as positive or vice versa. In this way each sentence can be marked as belonging to either positive or negative extracts of a movie review. Extracts thus generated can be further summarized with the help of these feature identification techniques..

Classification type	Accuracy	Time taken to build the model
Support Vector Machine	81.6%	1.65 seconds
Naive Bayes	81.35%	0.01 seconds
Maximum Entropy	81.4%	0.25 seconds

2) Feature Identification

LSA is employed to find out related feature terms of specified seed features,and these related terms could be regarded as being semantically related to the specified features. These related terms can be employed to select summary sentences.

IV. EXPERIMENTAL RESULTS

A. Movie Review Classification

Several experiments are performed to evaluate our system. In sentiment classification experiment three different machine learning algorithms namely, SVM classifier, Naive Bayes classifier and Maximum Entropy Classifier are compared to find out the most suitable classifier for this task. Different feature combinations are used to evaluate the system performance.

Feature type	Description
Frequent Unigrams	Presence of most frequent unigram as a feature
Discriminative features	Words with predictive ability as a feature

TABLE I
FEATURE DESCRIPTION

Classification type	Accuracy	Time taken to build the model
Support Vector Machine	78.8%	29.66 seconds
Naive Bayes	80.55%	0.54 seconds
Maximum Entropy	61%	34.44 seconds

TABLE2
FREQUENT UNIGRAMS: NO. OF FEATURES:1157- DATASET[3]

TABLE3
DISCRIMINATIVE FEATURES: NO. OF FEATURES:50- DATASET[3]

The above experimental results show that NB based sentiment classification performs better than the other two machine learning based classification approaches taking into account both speed and accuracy. In addition to that when the number features is reduced based on its predictive ability there is a drastic improvement in both accuracy and time taken to build the model.

Inorder to verify the accuracy of the above results same experiments were carried out on another benchmark movie review dataset [10]. The experimental results are tabulated below.

TABLE4
FREQUENT UNIGRAMS: NO. OF FEATURES:1351-DATASET[10]

Classification type	Accuracy	Time taken to build the model
Support Vector Machine	81.75%	13.16 seconds
Naive Bayes	81.15%	6.07 seconds
Maximum Entropy	74.3%	48.82 seconds

TABLE5
DISCRIMINATIVE FEATURES:NO. OF FEATURES:32-DATASET[10]

B. *Feature Identification*

Two different feature identification tasks, namely LSA [6] and LDA [1] are compared. The results of each of which are then used in generating the summary. The identified features were compared with the standard movie review glossary data [3] to analyse their precision, recall and f-value measures. The results are shown in the following figures . From fig.2 it can be seen that precision values are higher for LSA and as the number of terms increases a significant improvement can be seen for LDA based feature identification. As far as recall is concerned, again LSA shows better results when compared with LSA but its performance begins to degrade when the number of terms increases and LDA

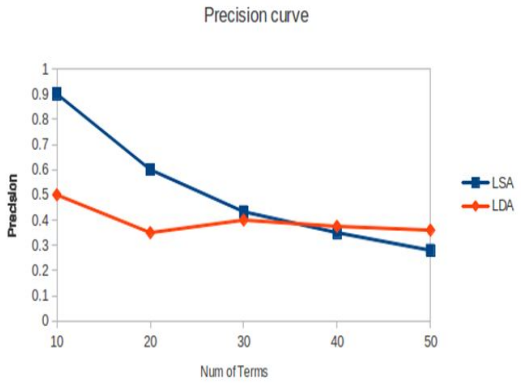

Fig.2 Precision Curve

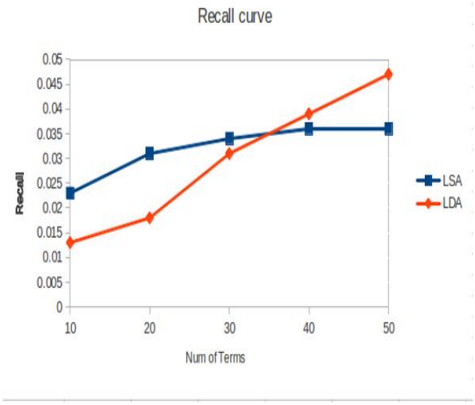

Fig. 3 Recall curve

Classification type	Accuracy	Time taken to build the model
Support Vector Machine	99.1%	0.06 seconds
Naive Bayes	99.3%	0.01 seconds
Maximum Entropy	98.8%	0.15 seconds

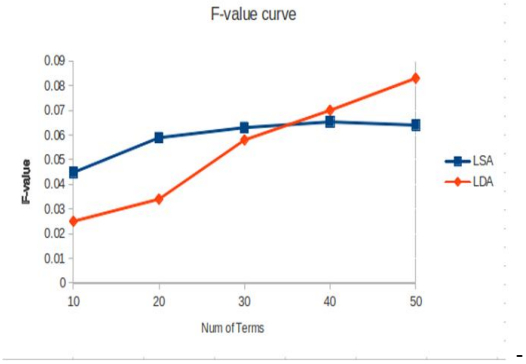

fig. 4 F-value curve

V. CONCLUSIONS

A movie review classification and feature based summarization system is designed and implemented. Sentiment classification using machine learning approach is applied to the movie reviews. The experimental results shows that Naive Bayes classifier is the best suited approach for this task taking into account its accuracy and the time taken to build the model. The reduction in the number of features based on its predictive ability has immense effect in the system performance. Furthermore, LSA based filtering approach to reduce the size of the summary based on users preferred aspect is implemented. Furthermore, LSA based filtering approach to reduce the size of the summary based on users preferred aspect works better than LDA based approach.

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn.Res., 3: pp. 993–1022, March 2003.

[2] Sangno Lee, J. Baker, Jaeki Song, and J.C.Wetherbe. An empirical comparison of four text mining methods. In System Sciences (HICSS),2010 43rd Hawaii International Conference on, pp. 1 –10, jan. 2010.

[3] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and E. Jou. Movie rating and review summarization in mobile environment. Systems, Man, and Cybernetics, Part C:Applications and Reviews, IEEE Transactions on,42(3): pp.397–407, May.

[4] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt,and B. Scholkopf. Support vector machines. Intelligent Systems and their Applications, IEEE,13(4): pp. 18–28, Jul/Aug.

[5] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan.Thumbs up?:Sentiment classification using machine learning techniques.In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, EMNLP '02, pp. 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[6] Susan T. Dumais. Latent semantic analysis. Annual Review of Information Science and Technology, 38(1):1pp. 88–230, 2004.

[7] Thomas Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99,pp. 50–57, New York, NY, USA, 1999. ACM.

[8] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '04, pp. 168–177, New York, NY, USA, 2004. ACM.

[9] 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.

[10] 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: Human Language Technologies., pages 142–150, Portland, Oregon, USA, June 2011. ACL.