International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.4., 2015
(July-Aug)

RESEARCH ARTICLE

ISSN: 2321-7758

MINING AND SUMMARIZING FROM CUSTOMER REVIEWS FOR COMPETITIVE INTELLIGENCE

M.KEERTHANA[1], M.LOVELINPONNFELCIAH[2]

[1]M.phil. Scholars, [2]Assistant Professor

Department of Computer Science, Bishop Heber College (Autonomous), Thiruchirapalli, India.

ISSN:2321-7758
International Journal of Engineering Research-Online
A peer-reviewed, open access International scientific journal
www.ijoer.in

ABSTRACT

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.. In this paper, we design and develop various strategies required for sentiment analysis of movie domain. 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 thought of sentiment classification accuracy & system response time.

INTRODUCTION

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. So, movie review mining is comparatively, a more challenging and interesting domain than product review mining. Sentiment analysis is a category 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.

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. Additional problem is that if there are huge 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.

This paper discovers and designs a mobile system for movie rating and review summarization in which semantic orientation of comments, the restriction of small display capability of devices, and system response time are considered. 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.

Sentiment Classification

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

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, is as large as possible. This hunt corresponds to a constrained optimization problem. We employed SVM to perform the classification and libsvm package is used in the system. The kernel role 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.

Feature Based Summarization

Feature based summarization is based on movie features and opinion words. System recognized movie features using Latent semantic Analysis (LSA) algorithm. 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 and identification of opinion words is also very important. System examines 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.

Pre-Processing

The following are the major steps adopted to produce a matrix of meaningful sentiment descriptors whose weights were determined based on their presence in a twitter comment and their overall presence in the corpus of tweets collected on a particular movie.

1) Consolidate all tweets into two separate corpora as shown in table 1 and perform the following steps for each corpus.

2) Remove stop words.

3) Reduce verbs to lemmas using a simple nonaggressive stemming algorithm.

4) Discard rare words by giving a lower limit to the frequency of accepted words equal 3. The weighted movie descriptor frequency matrix is calculated using vectorization as

wi=tfi*log(D/dfi) ----------------(1)

5) Take the transpose of the matrix obtained in step 4 in order to cluster the words instead of the comments.

6) Take the transpose of the matrix found in step 5 was then trimmed by including only those attributes (rows) which were elements of the sentiment lexicon described above.

Latent semantic analysis (LSA)

In movie-feature identification, we compared LSA based approach with frequency based approach. We performed experiments using movie review glossary dataset. Latent semantic analysis algorithm is used to identify movie features and the seeds include scene, director, plot, actor, and story. 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.

Performance

The runtime time and space requirements of the naıve Bayes and SVM classifiers are both roughly linear in the number of features. However, the training time for the SVM classifiers was usually much faster than for the naıve Bayes classifier. Even though the naıve Bayes classifier generally converged within 20 or so EM iterations, in some cases it took more than 100 iterations to reach convergence, which could take up to several hours on data sets with large numbers of features. The SVM classifiers generally converged within a few minutes at poorest and often within seconds.

CONCLUSIONS

Sentiment classification and feature based summarization system 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 because LSA islanguage-independent and LSA 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.

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]. LIBSVM: A library for support vectormachines[online].Available:http://www.csie.ntu.edu.tw/c jlin/libsvm. (2001),

[7]. 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.

[8]. SavitaHarer 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.

[9]. 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.

[10]. T. Hofmann, “Probabilistic Latent Semantic Indexing,” Proc. 22ndAnn. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 50- 57, 1999.

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

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