Mining Movie Reviews using Machine Learning Techniques

N. Sudha
Research Scholar
Department of Computer
Science and Engineering
Annamalai University

M. Govindarajan, PhD
Assistant Professor
Department of Computer
Science and Engineering
Annamalai University

ABSTRACT

Sentiment analysis has been observed as an important subject in data mining because of the wide range of direct applications such as analysis of products, customer profiles, and political trends and so on. It is the process of identifying people's attitude and emotional state from language to language. In Natural Language Processing, sentiment analysis is an automated task where machine learning is used to rapidly determine the sentiment of large amounts of text or speech.In this research work a comparative study of effectiveness in which some of the Machine learning techniques like naïve bayes and support vector machine. The results observed and noted that naïve bayes performs better in terms of accuracy, precision, recall and F-Measure for movie review.

General Terms

Data mining, Classification, Data sets, Social network and Algorithms

Keywords

Sentiment analysis, opinion extraction, reviews, Support vector machine, Naïve bayes

1. INTRODUCTION

With the immense development of the Internet, a large number of people would begin to express their views and ideas about all kinds of things on the internet; this is because of its rapid growth due to the increasing event of social network contacts, online discussion forums, blogs, movie reviews, digitals libraries and quick streaming news stories. The computational study of peoples' opinions, sentiments, attitudes and emotions expressed in text is called Sentiment analysis (SA) or Opinion Mining (OM). The two terms SA or OM are interchangeable.

They express a mutual meaning. However, some researchers declared that OM and SA has slightly been different in terms of notions. Opinion Mining extracts and analyzes people's opinion aboutan entity while Sentiment Analysis identifies the sentimentexpressed in a text then analyzes it[13]. The main aim of Opinion mining is to find the polarity of comments (positive, negative or diplomatic) by taking out features and components of the object that have been commented on each document [2,14]. Sentiment analysis is a form of natural language processing to address the mood of the public about a particular product or movie reviews. For instance, Sentiment mining is based on relating multiple sentiments with a document, which gives more exact division into multiple classes, (e.g., happy, sad, angry, disgusting, etc.)

Sentiment analysis is the process of finding the opinion of an individual. Sentiment classification has been categorized into two different types. The first one is supervised and the second one is un-supervised classification techniques. Supervised classification techniques are used in machine learning pattern to classify the review. It is mainly used in movie reviews . Corpus is made to describe data in the document, and then the corpus is trained using the classifiers such as Naïve Bayes , Maximum Entropy and Support Vector Machines. These are the machine learning approaches trained on different feature sets[8]. Sentiment analysis classification can be performed at three levels. They are document-level, sentence-level, and aspect-level [1].

Document-level consists of an opinion document as proving a positive or negative opinion or sentiment. It considers the entire document a basic information unit. In document level the sentiment classification has been focused using machine learning techniques [7].

Sentence-level intends to classify emotions expressed in every sentence. SA refers to classify the sentiment in respect to a particular feature of entities, determining whether an opinion expressed on each feature is either positive, negative or diplomatic. Sentiment analysis provides People with information on new to choose a

product and also helps the organization to improve the quality of the product. The further research is based on the topics as below. Section 2 deals with literature review. Section 3 deals with the classification methods for sentiment mining. Section 4 deals with the results and discussion. Section 6 conclude the work done and discuss the scope for future research.

2. LITERATURE REVIEW

Lina Zhou et al., examined movie review mining using machine learning and semantic orientation. Supervised classification technique and text classification techniques are used in the proposed machine learning approach to classify the movie review. A corpus is formed to represent the data in the documents and all the classifiers are prepared using this corpus. Thus, the proposed technique is more effective. However, the machine learning approach employs supervised learning, the proposed semantic orientation approach employs "unsupervised learning" since it does not require prior training in order to mine the data.[4]. Poirier et al investigated two varying opinion extraction methods. The first relied on machine learning technique based Naive Bayesian classifier where the second applied NLP techniques to process opinions and build dictionaries which determine a comment's polarity based on its words. Both approaches were estimated with contents from flixster.com. The results showed that using a low-level NLP approach with a small corpus tend to good training: a lexicon building cost and negation detection designing remained sensible. When the corpus was great, ML approach could be spread easily [9].

Turney's applied on classification of reviews is perhaps the most approach to ours. He uesd a specific unsupervised learning technique based on the mutual information between document phrases and the words "excellent" and "poor", where as the mutual information is calculated using statistics gathered by a search engine. In contrast, we used several completely prior-knowledge-free supervised machine learning techniques with the goal of empathizing the inherent trouble of the task [14].

Liu et al formulated a movie-rating and review-summarization system in a mobile environment was created and Movie-rating information is based on sentiment-classification results. Condensed movie review descriptions are produced from a feature-based summarization. The authors suggested a new approach based on latent semantic analysis (LSA) for product features identification. Also, summary size was decreased, based on product features from LSA. Both sentiment-classification accuracy and system response time were taken into discussion in system designing. Rating and review-summarization system is flexible to be extended to other product-review domains well [5].

Jakob et al estimated whether an anaphora resolution algorithm could improve a baseline opinion mining system's performance. Based on two different anaphora resolution systems, an analysis was performed. Experiments on a movie review corpus showed that unsupervised anaphora resolution algorithm greatly improved target extraction in opinions [3].

Zhang et al. applied word dependency structure to classify the sentiment using rule based semantic analysis[15]. Maas et al. applied both supervised and un-supervised techniques by getting semantic term document information to learn word vector [6].

3. METHODOLOGY

Different approaches to classify the machine learning methods which includes Naïve Bayes and Support Vector Machine.

3.1 Naïve Bayes

A Naive Bayes classifier is a simple probabilistic model based on the Bayes rule along with a strong independence assumption. The Nave Bayes model involves a simplifying conditional independence assumption. That is a given class (where people don't express opinions in the same way; they use opinion words as positive or negative comments), the words are conditionally independent of each other. This assumption does not affect much the accuracy in text classification but makes really fast classification algorithms applicable for the problems.

3.2 Support vector machine

Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization, they are large-margin, rather than probabilistic classifiers, 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 hat 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; letting cj$\in\{1, -1\}$(corresponding to positive and negative) be the correct class of document dj, the solution can be written as

$$\vec{w} = \sum_i \alpha_j c_j \vec{d}, \quad \alpha_j > 0,$$

where the αj, are obtained by solving a dual optimization problem. Those such that is greater than zero are called support vectors, since they are the only document vectors contributing to. Classification of test instances consists simply of determining which side of hyperplane they fall on.

4. EXPERIMENTAL RESULTS

4.1 Dataset Description

We used documents from the movie-review corpus. The basic data set consist of 2000 movie reviews, 1000 labelled positive and 1000 labelled negative (so they have a uniform class distribution). These were downloaded from Bo Pang's web page:

http://www.cs.cornell.edu/people/pabo/movie-eview-data

4.2 Results and discussion

Table 1: classification accuracy for NB and SVM classifiers

Dataset	Algorithm	Accuracy
Movie review data	Naïve Bayes(NB)	87.9%
	Support vector Mechine (SVM)	86.5%

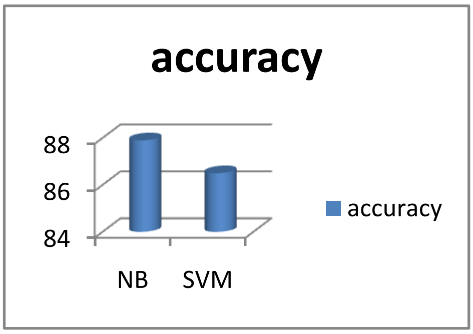

Figure 1: Classification accuracy for movie review data

Table 2: Classification Precision Recall f-measure

Techniques used	Precision	Recall	F-Measure
Naïve Bayes(NB)	0.862	0.879	0.844
Support Vector Machine (SVM)	0.749	0.866	0.803

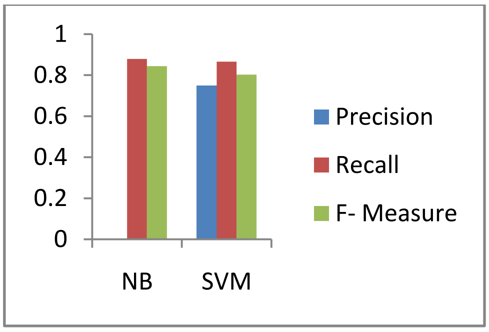

Figure 2: Precision Recall and F-measure

The data set described in section 4 is being used to test the performance of classifiers. Experimental results are conducted for sentiment classification using online movie review data. 2000 instances (1000 positive and 1000 negative) were used for evaluation. Following Tables 1 and 2 and Figures 1 and 2 gives the classification accuracy, precision and recall for the various classifiers used for classifying the opinion into positive or negative. It is seen from Fig.1, that the classification accuracy achieved by Naïve Bayes is much better than that of Support vector machine. Naïve Bayes achieves 87.9 % better classification accuracy than the other classifiers.

5. CONCLUSION

In this research work on mining opinions from unstructured documents. The focus was on extracting relations between movie reviews and opinion expressions. Opinion in movie reviews is analyzed/classified as positive/negative. Features are extracted from reviews using Inverse document frequency and reviews are classified through use of the Naïve Bayes, Support vector machine classifier. Experimental results show that Naïve Bayes achieve the best classification. In this research, two different algorithms have been implemented. In future, other similar classification algorithms under supervised learning methodology like decision tree learning, meta algorithm, artificial neural network and others can be considered to implement.

6. REFERENCES

[1] Bing Liu,"Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers", pp. 1- 167,May 2012.

[2] Dave.D, Lawrence.A, Pennock.D," Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of InternationalWorld Wide Web Conference, pp. 20-24, May 2003.

[3] Jakob, N., &Gurevych. I,"Using anaphora resolution to improve opinion target identification in movie reviews", In Proceedings of the ACL Conference Short Papers, pp. 263-268,July 2010.

[4] Lina Zhou, PimwadeeChaovalit, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii International Conference on system sciences, pp. 112c-112c, January 2005.

[5] Liu, C. L., Hsaio, W. H., Lee, C. H., Lu, G. C., &Jou," E. Movie rating and review summarization in mobile environment", IEEE Transactions on, Systems, Man & Cybernetics: Part C - Applications & Reviews, Vol 42, pp. 397-407,2012.

[6] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts "Learning Word Vectors for Sentiment Analysis", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol.1, pp.142-150, June 2011.

[7] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis" ,Foundations and Trends in Information Retrieval, vol. 2, nos. 1/2, pp. 1-135, January 2008.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up? Sentiment Classification Using Machine Learning Techniques", proceedings of the ACL-02 conference on Empirical methods in natural language processing,Vol. 10, pp. 79-86, January 2008.

[9] Poirier, D., Bothorel, C., De Neef, E. G., &Boullé, M. Automating opinion analysis in film reviews: the case of statistic versus linguistic approach. In Affective Computing and Sentiment Analysis, Springer Netherlands,Vol. 45,pp. 125-140, July 2011.

[10] RanjaniGandhi, V, Priya. N," Literature Survey on Data Mining and Statistical Report for Drugs Reviews", IJIRCCE, Vol. 3 Issue 3, pp. 1734-1739, March 2015.

[11] Richa Sharma, Shweta Nigam, Rekha Jain, "Opinion Mining of Movie Reviews at Document level",International Journal on Information Theory (IJIT), Vol.3, No.3, pp. 13-21, July 2014 .

[12] Ronen Feldman," Techniques and Applications for Sentiment Analysis", Communications of the ACM, Vol. 56 No. 4, pp. 82-89, April 2013.

[13] TsytsarauMikalai, PalpanasThemis. Survey on mining subjective data on the web. Data Mining and KnowledgeDiscovery, Vol. 24, pp. 478–514, May 2012.

[14] Turney, P, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Vol.21. No.4, pp. 417-424, July 2002.

[15] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, and W. Zuo, "Sentiment Analysis of Chinese Documents: From Sentence to Document Level", Journal of the American Society for Information Science and Technology, vol. 60, No. 12, pp. 2474-2487, December 2009.