International Journal of Engineering, Science and Mathematics
Vol. 7Issue 5, May 2018,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

PERFORMANCE EVALUATION OF NAÏVE BAYES AND MAXIMUM ENTROPY MODEL THROUGH OPINION MINING

Ashwani Kumar[*]
Deepika Goyal[**]

Abstract

Opinions, sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining. The inception and rapid growth of the field coincide with those of the social media on the Web, e.g., reviews, forum discussions, blogs, micro blogs, Twitter, and social networks, because, we have a huge volume of opinionated data recorded in digital forms. Since early 2000, opinion mining has grown to be one of the most active research areas in natural language processing. It is also widely studied in data mining, Web mining, and text mining. In fact, it has spread from computer science to management sciences and social sciences due to its importance to business and society as a whole. In recent years, industrial activities surrounding opining mining have also thrived.

Opining mining is a type of natural language processing for tracking the mood of the public about a particular product or topic. Opining mining, which is also called Sentiment analysis, involves in building a system to collect and examine opinions about the product made in blog posts, comments, reviews or tweets. Opining mining can be useful in several ways. For example, in marketing it helps in judging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features. In our thesis we proposed a opining mining system of online mobile phone reviews using naïve bayes classification and maximum entropy model. The technique will perform Opining mining and give acquiescent result and will also compare the performance of naïve bayes classification and maximum entropy model.

Keywords:

Naïve Bayes;
Maximum Entropy Model;
Opinion Mining.

Author correspondence:
Ashwani Kumar,
M.Tech Program, Advanced Institute of Technology and Management
Maharishi Dayanand University,Rohtak, Haryana

[*] M.Tech Program, Advanced Institute of Technology and Management, Maharishi Dayanand University, Rohtak, Haryana.

[**] Head, Dept. of Computer Science and Engineering , Advanced Institute of Technology and Management, Maharishi Dayanand University, Rohtak, Haryana.

1. Introduction

Opinion Mining is a type of natural language processing for tracking the inclination of the public about a specific product or topic. Opinion Mining, which is also called sentiment analysis, involves in building a system to gather and analyse opinions about the product described in blog posts, comments, reviews or tweets. Opinion mining can be useful in several ways. For example, in marketing it helps in judging the success of an advertisement campaign or the launch of a new product, determine which versions of a product or service are popular and even identify which demographics like or dislike particular features.

There are a number of challenges faced in Opinion Mining. The first is an opinion word that is considered to be positive in one situation may be considered negative in another situation. A second challenge is that people don't always express opinions in an identical way. Most traditional text processing relies on the fact that small variation between two pieces of text doesn’t make a huge impact on the meaning of the context very much. In opinion mining, however, "the movie was great" is very different from "the movie was not great". People can be ambiguous in their statements. Most reviews will have both positive and negative comments, which is somewhat controllable by analyzing the sentences one at a time. However, in the more informal medium like twitter or blogs, the more likely people are to combine various sentiments in the same sentence which is easy for a human to understand, but quite difficult for a computer to parse. Many a times even other people have difficulty in understanding someone thoughts based on a small piece of text information because it lacks context. For example, "This version of phone was as good as the last one” is entirely dependent on what the person expressing the opinion thought of the previous model.

Opinion mining concentrates on attitudes, whereas traditional text mining focuses on the analysis of facts. There are few fields of research outweigh in opinion mining such as sentiment classification, feature based opinion classification and opinion summarization. Sentiment classification deals with classifying the complete document according to the opinions towards certain product or objects. Feature-based opinion classification considers the opinions on features of certain product or object. Opinion summarization task unlike traditional text summarization only the features of the product are mined on which the customers have expressed their views. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the meaning in the classic text summarization. Languages that have been generally focused are English and Chinese. Presently, there are very few researches who conduct opinion based classification languages such as like Spanish, Italian and Thai.

2. Literature Review

In this section detailed literature review of the different techniques in opinion mining is presented. This literature review helps us to know the methods in analysis of opinions or sentiments. There are various methods that can perform opinion mining. Some of them are based on supervised approach and some are based on unsupervised approach. These all reviews are summarized below:

Jalaj S. Modha et. al. [6] in this paper sentiment analysis for unstructured data on web was discussed; they presented a study of existing methods, approaches to perform opinion mining. Currently, only subjective statements are taken into account while performing opinion mining and objective statements carrying opinions or sentiments are ignored. So, they proposed a new approach to classify and handle both subjective and objective statements for sentimental. First they classify the data into opinionated and non- opinionated categories after this opinionated data is divided into subjective and objective statements and in the last step both subjective and objective statements are categorized into positive, negative and neutral. K. Bun et al. [7] they proposed an information system that will extract important topics in a news archive per week. A user can know what the important news events happened in the last week by obtaining a weekly report. In general, related research on subject identification is divided into two types. They presented a study over term weighting method that extracts useful terms which are relevant to collected documents and modelled also. Second is TF-IDF mostly used for term weighting in Natural language processing and information extraction process.

Dengya Zhu et. al. [8] produced more enhanced formula „R-TF-IDF "which is better than TF-IDF". Here TF-IDF formula is multiplied by an adjusting factor. Importance of term frequency in a document is increased by this factor whereas the terms having relatively higher term frequency weighting and appearing less frequent are punished. Mukhrjee A. et. al. [9] performed work on techniques F-measure and EFS algorithm. F-measure suggested the concept of implicitness of text and is a unitary measure of text relative contextuality and formality. Certain part of speech are used to calculate F score which defines the contextuality and formality. A low F score defines contextuality (implicitness) suggesting greater use of verbs, pronouns, interjections and adverbs. A high F score defines formality (explicitness) suggesting use of nouns, adjectives and prepositions in the text. EFS Algorithm uses both the advantages in its working. For ranking the features following the filter model, it uses a number of feature selection criteria. After ranking, using the wrapper model a feature based on classification accuracy is found by the algorithm using candidate feature subsets, to find the final feature set.

Ying Chen et. al. [10] Traditional Blog systems could not perform well information retrieval and organization process performance due to lack of semantic support for query processing. In this paper they designed Blog ontology and domain subject classification ontology with analyses of the existing technologies in Blog systems and concentrating on semantic retrieval of Blog information resource. SPARQL query is used for semantic retrieval with predefined rules. Jacques Savoy et. al. [11] in this paper they proposed Z score values based on n-gram of characters, lemmas. The classification scheme is very simple and new in approach. The calculated Z score value tells the differences between the expected occurrence frequencies and observed frequencies. If the term has a large positive Z score, it belongs to the specific vocabulary, and if it has a large negative Z score means the term is underused. The classification rule is to sum the Z score values over all terms appearing in a text. From the experiments it is shown that the Z score scheme is more efficient as compare to F-measure, SVM model and the Naïve Bayes approach.

3. Research Method (10pt)

Figure 4.1 shows the proposed framework of our approach which has three modules: Data extraction, Pre-processing and Classification. In this proposed approach, number of steps are used conceptualize, design and perform an effective opinion mining of online airline reviews that is to be achieved by using Naïve Bayes Classification and Maximum entropy classifier and to evaluate which technique is more accurate for opinion mining.

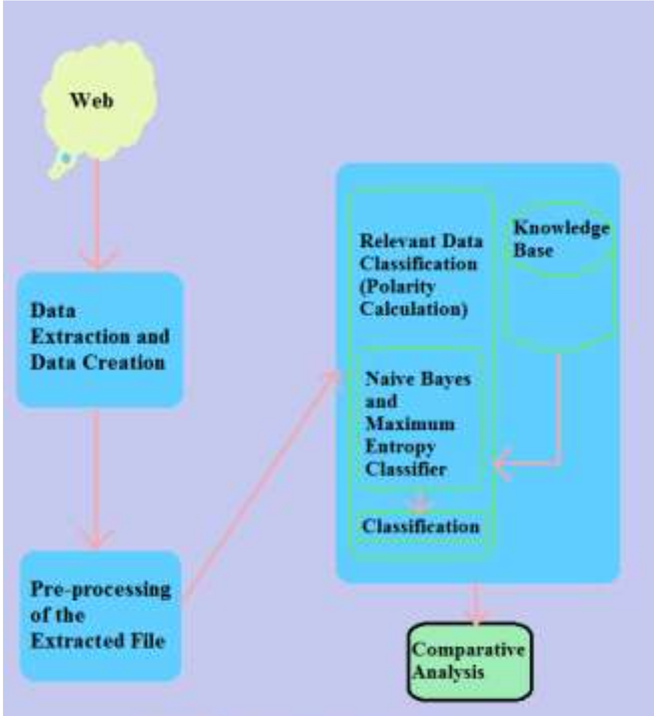
Web
Data Extraction and Data Creation
Pre-processing of the Extracted File
Relevant Data Classification (Polarity Calculation)
Knowledge Base
Naive Bayes and Maximum Entropy Classifier
Classification
Comparative Analysis

Figure 1 : Proposed Framework

The proposed approach of opinion mining has following steps:
1. First step: Extract the data to be analyzed from the web (Twitter, Blogs etc.). In our work we have extracted data from twitter for airline review database.
2. Second step: For preprocessing and polarity calculation of the extracted data we have created a training dataset for positive, negative, average sentiment words and stop words in SQLite.
3. Third step: Preprocessing- In the pre-processing of the data the words that does not carrying any sentiments or opinion are removed from the database. Another task performed in pre-processing is stemming. It is the process of reducing derived words into their root forms e.g. word sadness is reduced into root form sad. So, after pre-processing we get only the meaningful data on which we can apply the techniques.
4. Fourth Step: Here we calculated the polarity of the processed data. Polarity provides us the count of positive, negative and average sentiment words in the entered dataset which is used by the techniques as an input for further processing.
5. Fifth Step: Classification is done using Naïve Bayes Classification and Maximum Entropy Model.

Naïve Bayes Classification is based on supervised learning. It is a statistical method for classification. It computes the probabilities of the outcomes to determine whether a sample belongs to a particular class or not. It is used for both diagnostic and predictive problems.

The Max Entropy classifier is a probabilistic classifier which belongs to the class of exponential models. Max Entropy model does not assume that the features are conditionally independent of each other. The MaxEnt is based on the Principle of Maximum Entropy and from all the models that fit our training data, selects the one which has the largest entropy. The Max Entropy classifier can be used to solve a large variety of text classification problems such as language detection, topic classification, sentiment analysis and more.

Due to the minimum assumptions that the Maximum Entropy classifier makes, we regularly use it when we don't know anything about the prior distributions and when it is unsafe to make any such assumptions. Moreover Maximum Entropy classifier is used when we can't assume the conditional independence of the features. This is particularly true in Text Classification problems where our features are usually words which obviously are not independent. The Max Entropy requires more time to train comparing to Naive Bayes, primarily due to the optimization problem that needs to be solved in order to estimate the parameters of the model.

The first step of constructing this model is to collect a large number of training data which consists of samples represented on the following format: (xi,yi) where the xi includes the contextual information of the document (the sparse array) and yi its class. The second step is to summarize the training sample in terms of its empirical probability distribution.

The framework has two phases for Opinion Mining. First the creation of training data set of various opinion words in SQLite. Second is an interface in Spyder which is used for performing testing data and classification. This approach uses Python for implementation. First a review is entered and then pre-processing of the data is done so as to remove all the meaningless data. It helps in real time opinion mining by reducing the noise of the data and improves the classifier performance as well as increases the speed of the classification process. In the next step classification techniques Naïve Bayes Classification and Maximum Entropy Model have been applied. Both the techniques give result in the form of a probability function.

4. Results and Analysis

This study examined the performance of two advance machine learning techniques Naïve Bayes and Maximum Entropy Model in case of opinion mining of online airline reviews. Comparison of performance of both the techniquesis was done on four airline review datasets. The results shows that probability function value of Maximum Entropy Model is greater than probability function value of Naïve Bayes in all the four datasets.

With the help of confusion matrix parameters True Positive rate, accuracy, False Posotive rate and error rate is evaluated and it was determined that the True Positive rate and accuracy of Maximum Entropy Model is greater to that of Naïve Bayes across all the four datasets. The FP rate and error rate of Maximum Entropy Model is less than Naïve Bayes in all the four datasets. So with this work it is concluded that Maximum Entropy Model performs better than Naïve Bayes in case of opinion mining of online airline reviews.

5. Conclusion

This study examined the performance of two advance machine learning techniques namely Naïve Bayes and Maximum Entropy Model in case of opinion mining of online airline reviews. We compared the performance of both the techniques on four datasets. Our results with the help of confusion matrix parameters TP rate, accuracy, FP rate and error rate shows that TP rate and accuracy of Maximum Entropy Model is greater to that of Naïve Bayes across all the four datasets. The FP rate and the error rate of Maximum Entropy Model is less than Naïve Bayes in all the four datasets. So with our work we conclude that Maximum Entropy Model performs better than Naïve Bayes in case of opinion mining of online airline reviews.

References

[1] Abhimanyu Chopra, Abhinav Prashar and Chandresh Sain, "Natural Language Processing" , International Journal of Technology Enhancements and Emerging Engineering Research, Volume 1, Issue 4 131 ISSN 2347-4289.

[2] Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa, "Natural Language Processing (Almost) from Scratch", Journal of Machine Learning Research 12 (2011) 2493-2537, 2011.

[3] Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[4] Bing Liu and Lei Zhang, "A Survey of Opinion Mining and Sentiment Analysis", University of Illinois at Chicago Chicago, IL, November 2010.

[5] Bing Liu. "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, May 2012.

[6] Jalaj S. Modha, Prof & Head Gayatri S. Pandi Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 12, December 2013

[7] K. Bun and M. Ishizuka, "Topic extraction from news archive using TF*PDF algorithm" In Proceedings of Third International Conference on Web Information System Engineering.

[8] Dengya Zhu, and Jitian XIAO,"R-tfidf, a Variety of tf-idf Term Weighting Strategy in Document Categorization" published in 2011 IEEE Seventh International Conference on Semantics, Knowledge and Grid (SKG).

[9] Mukhrjee, A. and B. Liu, 2010, "Improving gender classification of weblog authors" EMNLP" 10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing Pages 207-2017 Association for Computational Linguistics Stroudsburg, PA, USA.

[10] Ying Chen, Wenping Guo, Xiaoming Zhao, "A Semantic Based Information Retrieval Model for Blog" Third International Symposium on Electronic Commerce and Security, 2010, IEEE.

[11] Jacques Savoy, Olena Zubaryeva, "Classification Based on Specific Vocabulary" published in 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 978-0-7695-4513-4/11 2011 .

[12] Fred Popowich, "Using Text Mining and Natural Language Processing for Health Care Claims Processing", SIGKDD Explorations. Volume 7, Issue 1 - Page 59.

[13] Hemlatha, Saradhi Varma and A.Govardhan, "Sentiment Analysis Tool using Machine

Learning Algorithms", International Journal of Emerging Trends and Technology in Computer Science, Volume 2, Issue 2, March-April 2013 ISSN: 2278-6856.

[14] Zhongwu Zhai, Bing Liu and Hua Xu, "Clustering Product Features for Opinion Mining", WSDM"11, February 9–12, 2011, Hong Kong, China. Copyright 2011 ACM 978-1-4503-0493-1/11/02.

[15] G.Vinodhini and RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012.