ANNQUEST NOVEMBER 2016
ISSN: 2321-3043
www.stannsannquest.com

A review of Machine Learning techniques used in Opinion Mining

Aparna Bulusu, Dept. of Computer Science, mail id:awudali@gmail.com
A. Usha Rani, Dept. of Physics & Electronics, mail id:a_usha_26@yahoo.com
St. Ann's College for Women, Mehdipatnam

Abstract

The proliferation of internet technologies and smart phones has led to an increasing number of people using social networking and micro blogging platforms. Consumers are constantly expressing opinions about various kinds of products and services in multiple formats in huge numbers. This has led to the development of a new discipline called opinion mining that focuses on automatically classifying the sentiments of users without manual intervention. Opinion mining has acquired immense research and commercial interest due to its potential applications for users and businesses alike. While many methods are available for automatically classifying sentiments of web based articles and reviews, machine learning techniques like naive bayes, support vector machines, maximum entropy, decision trees etc are being widely used in opinion mining tasks. This paper attempts to understand the problem of opinion mining and how machine learning algorithms are being implemented to analyze the sentiments of user opinions

Keywords: Opinion mining, Data mining, Machine learning algorithms

Introduction

Huge amounts of data are being generated at an extremely fast pace, thanks to the availability of high speed internet and allied technologies to a vast majority of the population. Increasing numbers of users are now active on various social platforms , blogs etc and keep posting reviews/status updates etc continuously. Consumers are checking out user reviews , social network results, trending topics, product ratings etc for making purchasing decisions. Success of businesses is now dependent on the positive sentiment generated through their consumer feedbacks and reviews. Ecommerce websites are the biggest source of user reviews. However manually sifting through hundreds of user reviews is practically not feasible. The problem for businesses is therefore making sense of the large number of customer opinions of their products which may be sold across multiple merchant sites.[1]This task of understanding and classifying user opinions or sentiments involves two areas of study: Text mining in order to collect, extract and classify the huge amounts of data into clusters followed by automatic sentiment analysis. Sentiment classification is therefore a special case of text categorization, where the criterion of classification is the attitude expressed in the text, rather than the "content" or topic. [2]

The problem of Opinion Mining

Opinion mining or sentiment analysis, involves creating an automated system to collect and classify opinions about a product. Automated opinion mining often uses machine learning, a type of artificial intelligence, to mine text for sentiment.

Sentiment Analysis can be performed at various levels of granularity. The first is document level where a whole document is classified as expressing a positive or a negative opinion.[3] For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities. The next level is the Sentence level, where each sentence is examined to express a positive, negative, or neutral opinion. This level also involves subjectivity classification , which distinguishes objective sentences that express factual information from subjective sentences that express subjective views and opinions. A more fine grained level of classification is Entity and Aspect level (or feature level) which directly looks at the opinion itself. It is based on the idea that an opinion consists of a sentiment (positive or negative) and a target (of opinion). The goal of this level of analysis is to discover sentiments on entities and/or their aspects. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses. In addition to the various levels at which opinions can be mined, the opinions themselves are can also be classified into different types[4].

Opinions can be either regular straight forward opinions where they state a direct opinion about an issue or can be of the comparative type where they discuss similarities or differences between related entities.

Challenges: Opinion mining isn't a trivial task and poses a lot of challenges. A key challenge is that natural language can be used in different ways for expressing various opinions. i.e. it is based usually on the context which may be easy for humans to process and understand but might be difficult for systems based on algorithms to figure out. Another challenge is that it is very difficult for systems to classify subjective opinions into strictly positive or negative groups as most reviews tend to include multiple and contradictory sentiments which makes it difficult to classify them as either positive or negative. According to Bing Liu [5],some of the major issues faced in opinion mining are feature extraction, grouping of synonyms, opinion orientation classification, detecting mix words, spam and sarcasm and identifying the strength of an opinion.

Machine Learning - Techniques

Machine learning is an inter disciplinary field that lies at the intersection of computer science, engineering and statistics. Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning has been defined as a "Field of study that gives computers the ability to learn without being explicitly programmed".[6] It is a tool with applications in many fields like spam filtering, search engines, recommender systems etc. Machine learning tasks are typically classified into two broad categories:

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a trainer, and the goal is to learn a general rule that maps inputs to outputs. Some of the popular supervised learning algorithms include k-nearest Neighbors, Naive Bayes, Support Vector Machines, Decision Trees etc.

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).Algorithms like k-means and DBScan fall under this category.

The aim of Machine Learning is to develop an algorithm so as to optimize the performance of the system using example data or past experience. Machine learning in turn consists of certain steps like: Data pre processing , Feature selection and/or feature reduction , Representation , Classification and Post processing. The classification phase of this process is where the actual mapping between patterns and labels happens. Machine learning techniques are being used favourably to provide good results in classifying sentiments. A few of the more common machine learning techniques are as follows.

Machine learning techniques applied within the context of Opinion Mining:

- **Naive Bayes Classifier**: This method is being used primarily for document level sentiment classification. Bayesian classifiers assign the most likely class to a given example described by its feature vector. [7] Learning such classifiers can be greatly simplified by assuming that features are independent given class. In simpler terms, a small learning set is used to train a classifier to classify documents into positive or negative sentiment after extracting key features from them. The naive bayes classifier then works on the rest of the document set to classify them into their respective classes by using conditional probability as follows:

  $$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

  ,the above equation can be understood as calculating conditional probability based on prior probabilities as follows:

  $$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

  Despite the assumption that features are independent, Naive Bayes is remarkably successful in practice, often competing with much more sophisticated techniques. Naive Bayes has proven effective in many practical applications, including text classification, medical diagnosis, and systems performance management.[8]

- **Support Vector Method**: A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. Given labelled training data , the algorithm outputs an optimal hyper plane which categorizes new examples. Using a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.SVM is employed by manually classifying a small portion of test data into labels or numerical categories. SVM algorithm then uses mathematical concepts to create convex hulls for data belonging to same label thereby determining the hyper plane which separates the labels. Once the training algorithm determines the new hyper plane, it can be run on pre processed data to determine their underlying sentiment. In classic examples of determining sentiment among movie reviews, the best results were found in when analysis was performed on unigrams in a presence-based frequency model run through a Support Vector Machine (SVM), with 82.9 percent accuracy [3].

- **Maximum Entropy**: The principle of maximum entropy states that, subject to precisely stated prior data ,such as a proposition that expresses testable information, the probability distribution which best represents the current state of knowledge is the one with largest entropy. Maximum Entropy models are feature-based models. In a two class scenario, it is equivalent to using logistic regression to find a distribution over the classes. Maximum Entropy makes no independence assumptions for its features, unlike Naive Bayes. Features like bigrams and phrases are therefore used extensively in Maximum Entropy models. The model is represented as:

PME(c|d, λ) = exp[Σiλifi(c, d)] Σc 0 exp[Σiλifi(c, d)]

where c is the class (negative/positive), d is the data/tweet/review being analysed, and λ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the lambdas so as to maximize the conditional probability. Theoretically, Maximum Entropy performs better than Naive Bayes because it handles feature overlap better. However, in practice, Naive Bayes can still perform well on a variety of problems[9]

Conclusion

The explosion of user-generated content on the Web has led to new opportunities and significant challenges for companies, that are increasingly concerned about monitoring the discussion around their products. Sentiment analysis is a Natural Language Processing and Information Extraction task that aims to obtain writer’s feelings expressed in positive or negative comments, questions and requests, by analyzing a large numbers of documents tracking user sentiment and provides useful insight on how to improve products or market them more effectively. Recent approaches using machine learning techniques treat the task as a text classification problem, where they learn to classify sentiment based on labelled training data and have been providing accurate results.

References

1. Hu, Minqing, and Bing Liu, Mining and summarizing customer reviews, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

2. Michael Gamon, Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the 20th international conference on Computational Linguistics. 2004.

3 Pang, Lee and Vaithyanathan, Thumbs up?:Sentiment classification using machine learning techniques, Proceedings of ACL- 02 Conference, Empirical Methods Natural Lang. Process., 2002, pp. 79–86

4. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.

5. Bing Liu. 2010. Sentiment Analysis: A Multi-Faceted Problem, Invited talk at the 5th Annual Text Analytics Summit (2009).

6. https://en.wikipedia.org/wiki/Machine_learning

7. Rish, Irina. ,An empirical study of the naive Bayes classifier, IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. IBM New York, 2001.

8. Dinu, Liviu P., and Iulia Iuga. "The Naive Bayes classifier in opinion mining: in search of the best feature set." *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer Berlin Heidelberg, 2012.

9. Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009)