International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 8, October - 2012

A Conceptual Study Of Sentiment Mining

Pranali Yenkar
P.G. Student,
Department of Computer Engg.
Datta Meghe College of Engg
Sector-3,Airoli, Navi Mumbai, .
India,400708

Dr. S.D. Sawarkar
Principal,
Datta Meghe College of Engg,
Sector 3-Airoli, Navi Mumbai,
India, 400708

Abstract:

Information about people's opinions plays important role for more accurate decision making in a number of domains. There are number of opinion-rich resources available on the internet e.g. personal blogs, reviews, specialized websites which helps in understanding the opinions of others. Thus the Sentiment Mining has become the exciting field of research deals with extracting opinion, sentiment and polarity from textual data sources. Sentiment Mining has interesting applications number of fields including in commerce, social and academic areas. This study covers the concepts of sentiment mining and approaches, applications and algorithms in the field of sentiment mining are discussed in more details.

Keywords: Sentiment Mining, Opinion Mining, Text Mining, SVM

I. Introduction

Internet has become ever increasing source of information with more and more people share their reviews on the online forums. The opinions shared within comments, reviews, feedback or recommendation provides useful information for many different purposes. Most of the available information in computer systems is in the form of the text and according to survey around 85% of information found in organizations is in text format. So before the Sentiment Mining process commence, data mining on the text needs to perform which is called as Text Mining[3]. Knowledge discovery methods are applied in Text mining to unstructured textual data. Other research areas like natural language processing, artificial intelligence and machine learning can be used to tackle the complexities of fetching the information from unstructured textual data. A text document can be consider as a collection of objective and subjective statements, where objective statements refer to factual information present in text, and subjectivity relates to the expression of opinions. Sentiment mining deals with the application of automatic methods and algorithms for predicting the orientation and polarity of opinions mention in text documents .It detect and extract this information from textual data, and has a number of potential applications on building very efficient recommender systems, financial analysis, product engineering and market research.

Sentiments can be classified as positive or negative sentiments[2] with varying degree like very good, good, satisfactory, bad, and very bad. Mining opinions from the any review is the complicated procedure. First the data needs to be crawled from the web using web crawler then the data needs to prepared by cleaning it and removing some unwanted tags and non review data and then the data will be mined to summarize the opinion of the users in terms of positive or negative votes or the other way by

categorizing as recommended or not-recommended. Sentiment Mining has three main tasks: determining subjectivity, determining sentiment orientation, and determining the strength of the sentiment orientation.

II. Literature survey

Opinion mining and detection often called as sentiment analysis, sentiment classification, or sentiment mining automatically identifies emotions in textual data present on the web which is mostly in the unstructured format and extracts sentiment by rating a segment of text as either positive (favorable) or negative (unfavorable). Sentiment detection helps in business intelligence applications by perceiving how a user thinks about a certain product, service, tourism location, movie or political party. Some of the past work includes mining reviews of automobiles, banks, travel destinations [6], electronics [5,7] and mobile devices[7]. Pang et al.[8] applied machine learning approaches such as Naïve Bayes, Support Vector Machines and Maximum Entropy Modeling on movie reviews and obtained considerable results. Turney [6] performs binary classification on product reviews. Like Hatzivassiloglou and McKeon [9] he uses a lexicon containing a set of known sentiment terms which he extends by applying Point wise Mutual Information (PMI) and Latent Semantic Analysis (LSA).A more fine-grained approach presented by Pang and Lee determines the exact number of stars provided by the review author. Beineke et al. [10] refine Turney's work [6] by applying a Naïve Bayes model which they train on a labeled and an unlabeled corpus. Like Turney, they use a list of seed terms for the classification of new words, which only contains five positive and negative sentiment terms, as well as a larger list which they assemble from the WordNet synonyms of the terms like good, best, bad, boring. Nicholls and Song[11] examine the impact of different part-of-speech tags by employing a Maximum Entropy classifier by considering only adverbs, adjectives, verbs and nouns as relevant for sentiment detection and assign these categories different weights. According to the results adjectives and adverb carries the strongest sentiments as compared to verbs and nouns.

In [12], phrase patterns are used to explain a sentiment classification application which classifies opinions. At the phase of document classification, the tags are added to certain words in the text, and then the tags are matched within a sentence with predefined phrase patterns to get the sentiment orientation of the sentence under study. Next, the sentiment orientation of each sentence is considered and the text is classified according to the sentiment of the most repeated sentiment. A sentiment miner is described in [13] which uses Natural language processing (NLP) to analyze grammatical sentence structures and phrases and to detect the sentiment of the topic. This method has achieved high quality results (~90% of accuracy) on various datasets including online review articles and the general Web pages and news articles. An application on sentiment classification with review extraction is described in [14]. This uses the sentiment tags and weight approach. Here the review on any particular subject is extracted and a sentiment tag and weight is attached to each expression. Then, it calculates the sentiment indicator of each tag by accumulating the weights of all the expressions corresponding to a tag. Next, it uses a classifier to predict the sentiment label of the text. In this study, the authors use online documents covering two domains i.e. politics and religion are used to test the performance of the proposed application. The experiments within those domains achieve accuracy between %85 and %95. Opinion mining is studied for the e-learning system in[15].In this study, opinion mining is used to know the users'

opinions on the course-wares and teachers of the e-learning system and to help improve the services. The authors achieve following precisions for these subtasks respectively: %94, %84.2, %80.9 and %92.6. A sentiment mining and retrieval system called Amazing is described in [16]. In this system, the authors incorporate the temporal dimension information into the ranking mechanism, and make use of temporal opinion quality and relevance in ranking review sentences. This study monitors the changing trends of customer reviews in time and visualizes the changing trends of positive and negative opinion respectively. The authors conduct experiments using the customer reviews of four kinds of electronic products including digital cameras, cell phones, laptops, and MP3 players. The evaluation results indicate that the proposed approach achieves a precision of %85 approximately. The approach proposed in [17] uses a WorldNet, a large lexical database of English, for statistical analysis and movie knowledge. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms. WorldNet is used to generate a keyword list for finding features and opinions. Valid feature-opinion pairs are identified and finally, the sentences are recognized according to the extracted feature-opinion pairs to generate the summary. Experimental results show that this method has an average precision of %65 approximately. An opinion mining application is introduced in [18] which extracts and classifies people's opinions and emotions (or sentiment) from the contents of weblogs about movie reviews to calculate movie scores used unsupervised approach for sentiment mining. In [19] online hotel reviews are mined with supervised machine learning approach using unigram feature to realize polarity classification of documents.

III. Applications

Sentiment mining has applications in number of areas including customer intelligence, advertising systems, review and recommendation system and information retrieval. Companies are interested in finding out what are their customers' opinions about a new product launched on a marketing campaign. Sentiment analysis provides companies with a means to estimate the extent of product acceptance and to determine strategies to improve product quality .Consumers would also benefit from accessing other people's opinions and reviews on a given product they are intending to purchase, as recommendations from other users influence their purchasing decisions. Knowledge and polarity of common people's opinions are also important in the political scenario, where one could find out the sentiment towards an individual such as a politician or activist, any political party or a new updations in legislation.. It also facilitates policy makers or politicians to analyze public sentiments with respect to policies, public services or political issues and take proper and timely actions based on the information. Opinion Mining is also very useful in understanding the point of views of a web community about movie. By Mining the movie reviews having comments about different elements (e.g. screen- play, special effects, music, dance), as well as movie-related people (e.g. director, screenwriter, actor),user come to know the overall rating in terms of star.

IV. Sentiment Mining Approaches

Two approaches by which the sentiment mining can be done are Supervised Learning Method Unsupervised Learning Method and Semi-Supervised Learning Method. Some of the algorithms in each approach are mentioned below[1].

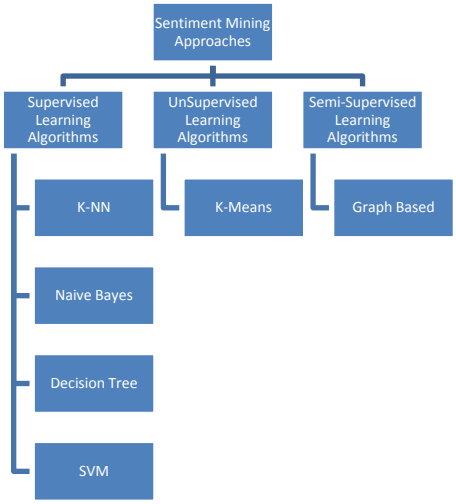

Fig 1:Sentiment Mining Approaches

1. Supervised Learning Method:

Supervised learning algorithm attempts to predict future values based on information contained on an already present data used for training. The training data set contains instances of the variable we wish to predict, and it is assumed that future values will have certain similarity to already observed values stored as a training data, which can be "learned" by a supervised learning algorithm. If future values do not retain any similarity to already seen data, prediction results will not be reliable . Thus, for the successful execution of good supervised learning algorithms, training data should be rich.

1.1. K-Nearest Neighbor classifier

It is a well known pattern recognition algorithm. Given a test document, the kNN algorithm finds the k nearest neighbors among the training documents and uses the categories of the k nearest neighbors to weight the category candidates. The similarity score of each neighbor document to the test document is used as the weight of the categories of the neighbor document. This algorithm is based on the assumption that the characteristics of members of the same class should be similar. Thus observations located close together in covariate space are members of the same class. It is suitable for data streams.

Merits: This method is effective, simple and easy to implement.

Demerits: This method becomes slow for vast training set .Its accuracy get affected by irrelevant features.

1.2 Naïve Bayes Method

Naïve Bayes classifier predicts by reading a set of examples in attribute value-representation and then by using the Bayes theorem to estimate the posterior probabilities of all qualifications. The independence assumptions of features make the features order irrelevant and presence of one feature does not affect other features in classification task.

Merits: The size of the training data set needs to be smaller.

Demerits: This method works well if assumed features are independent; when dependency arises then it gives low performance.

1.3 Decision Trees

The decision tree categorizes the training documents by constructing well-defined true/false queries in the form of tree structure. In this leaves represent the corresponding category of the text documents and branches represent conjunctions of features that lead to these categories.

Merits: This method works on any data. It is fastest even in the presence of large amounts of attributes.

1.4 Support Vector Machines

SVMs have been shown to be highly effective at traditional text categorization, which generally outperform Naive Bayes .SVMs has a hyper plane represented by vector w which separates

the positive and negative training vectors of documents with maximum margin[4].

Let y equal +1(_1), if document d is in class + (_). The solution can be written as

$$\vec{W} = \sum_{i=1}^{n} \alpha_i^* y_j \vec{d}_i \quad \alpha_i \geq 0$$

where α_i^* are obtained by solving a dual optimization problem. Above equation shows that the resulting weight vector of the hyper plane is constructed as a linear combination of $\vec{d}i$. Only those examples that I contribute to which the coefficient α_i is greater than zero.Those vectors are called support vectors, since they are the only document vectors contributing to $\vec{W}$.

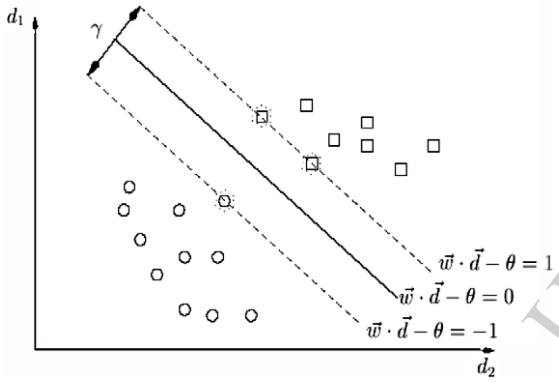

Fig 2. Classifier of SVM

The SVM need both positive and negative training set which are uncommon for other classification methods. The performance of the SVM classification remains unchanged even if documents that do not belong to the support vectors are removed from the set of training data.

Merits: SVM has been emerged as one of the most effective text classification methods as it is able to manage large spaces of features and high generalization ability.

Demerits: SVM algorithm is relatively more complex which in turn demands high time and memory consumption.

2. UnSupervised learning metod:

In unsupervised clustering, there are unlabelled collection of documents. The aim is to cluster the documents without additional knowledge or intervention such that documents within a cluster are more similar than documents between clusters.

2.1 K-Means Clustering:

In this each of *k* clusters can be represented by the mean of the documents assigned to that cluster, which is called the *centroid* of that cluster. There are two versions of *k*-means algorithm. The first version is the *batch* version and is also known as Forgy's algorithm . It consists of the following two-step major iterations[1]:

- Reassign all the documents to their nearest centroids
- Recompute centroids of newly assembled groups

(i) Before the iterations start, firstly *k* documents are selected as the initial centroids.

(ii) Iterations continue until a stopping criterion such as no reassignments occur is Achieved.

Merits: The main advantage of k-means is its speediness and simplicity.

Demerits: Its random process makes this an indeterminate method.

3. Semi-Supervised Learning Method

These algorithms make use of unlabeled data along with few labeled data to classify new unlabeled text document. In text classification most of the times there is limited labeled data, and in most cases it can be expensive to generate that labeled data so semi-supervised algorithms gives good solution in such a situations. Its framework is applicable to both classification and clustering .Some of the important algorithms discussed here are as:

3.1 Graph based

This learning algorithm works on a closed data set and test set is revealed at the time of training. Here one assumes that the data is embedded within a low-dimensional manifold which is expressed by a graph. Each data sample is represented by vertex within a weighted graph with the weights providing a measure of similarity between vertices. But many graph based SSL algorithms assume binary classification tasks and require the use of sub optimal approaches. Modified versions of Graph based SSL are based on optimizing a loss function composed of KL-divergence terms between probability distributions defined for each graph vertex.

Merits: These methods not only provide solution to binary TC but also for multiclass TC.
Demerits: These methods require excessive computation.

V. Conclusion

Due the immense growth of the available websites which has become the major source of the information, the consumer often overwhelmed with the information. As a consequence, he find it extremely difficult to obtain any useful comments to make a decision regarding the product to purchase or the service to avail. Hence the technique of Sentiment analysis is very useful to classify the comments on any particular issue, product or service. Supervised learning approach is the most efficient approach for Sentiment mining as compared to unsupervised and Semi-Supervised learning approach.

References

[1] Shweta C. Dharmadhikari, Maya Ingle ,Parag Kulkarni "Empirical Studies on Machine Learning Based Text Classification Algorithms" Advanced Computing: An International Journal (ACIJ), Vol.2, No.6, November 2011

[2] Bo Pang and Lillian Lee "Opinion Mining and Sentiment Analysis" Foundations and Trends in Information Retrieval Vol. 2, Nos. 1–2 (2008) 1–135.

[3] Ohana, Bruno, "Opinion mining with the SentWordNet lexical resource" (2009). *Dissertations*. Paper 25.

[4] Lucas Carstens" Sentiment Analysis : A multimodal approach" September 2011

[5] Kuat Yessenov, Sasa Misailovic"Sentiment Analysis of Movie Review Comments" Spring 2009

[6]Peter D.Turney, "Thumbs up or Thumbs Down? Semantic orientation applied to Unsupervised Classification of Reviews"presented at the Association for Computational Linguistics 40th Anniversary Meeting,New Brunswick,N,J,2002

[7]Satoshi Morinaga,Kenji Yamanishi ,Kenji Tateishi and Toshikazu Fukushima,"Mining product Reputations on the web"presented at the 8th ACM SIGKDD international conference on Knowledge discovery and data mining Edmonton,Alberta,Canada,2002

[8] Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up? Sentiment Classification using Machine Learning Techniques.In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Morristown, USA, pp. 79–86,2002.

[9] Hatzivassiloglou, V. and McKeown, K. R. Predicting the Semantic Orientation of Adjectives. In Proceedings of the European Conference of the Association for Computational Linguistics. Morristown, USA, pp. 174–181, 1997.

[10]Beineke, P., Hastie, T., and Vaithyanathan, S. The Sentimental Factor: Improving Review Classification via Human-providedInformation.In Proceedings of the Annual Meeting on Association for Computational Linguistics.Morristown, USA, pp. 263–269, 2004.

[11] Nicholls, C. and Song, F. Improving Sentiment Analysis with Part-of-Speech Weighting.InProceedingsof the International Conference on Machine Learning and Cybernetics. Baoding, China, pp. 1592–1597, 2009.
[12] Zhongchao Fei, et al., Sentiment Classification Using Phrase Patterns Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04), 2004.
[13] Jeonghee Yi, et al., Sentiment Mining in WebFountain,Proceedings of the 21st International Conference on Data Engineering (ICDE 2005), 2005
[14] Jian Liu, et al., Super Parsing: Sentiment Classification with Review Extraction, Proceedings of the Fifth International Conference on Computer and Information Technology (CIT'05), 2005.
[15] Yun-Qing Xia, et al., The Unified collocation Framework for Opinion Mining,
[16] Qingliang Miao, et al., AMAZING: A sentiment mining and retrieval system,Expert Systems with Applications (2008) doi:10.1016/j.eswa.2008.09.035.
[17] Li Zhuang, et al., Movie review mining and summarization, Proceedings of the 15th ACM international conference on Information and knowledge management,2006
[18] Arzu Baloglu, Mehmet S. Aktas" BlogMiner: Web Blog Mining Application for Classification of Movie Reviews" Fifth International Conference on Internet and Web Applications and Services-2010
[19] Han-Xiao Shi, Xiao-Jun Li" A Sentiment Analysis Model For Hotel Reviews Based On Supervised Learning" International Conference on Machine Learning and Cybernetics, Guilin, 10-13 July, 2011