Sentiment Analysis Levels and Techniques: A Survey

Priyanka Patil
Department of Computer Science and Engineering
Walchand Institute of Technology, Solapur, Maharashtra, India

Pratibha Yalagi
Department of Computer Science and Engineering
Walchand Institute of Technology, Solapur, Maharashtra, India

Abstract- The growth of social media contributes a large amount of user generated content such as comments, customer reviews and opinions. This user generated content is very useful for businesses, individuals and governments. But analysis of this content is difficult and time consuming. So there is a need to develop an intelligent system which automatically mines and classify into positive, negative and neutral category. Sentiment analysis is the automated mining of opinions and emotions from text, speech, and database sources. The objective of this work is to discover the concept of Sentiment Analysis, and describes a comparative study of its techniques in this field.

Keywords – Sentiment, Opinion, Machine learning, Semantic score

I. INTRODUCTION

Today, Internet has a huge amount of users and textual data, which is growing every day. The social media is now a major part of the Web. The statistics show that every four out of five users on the Internet use some form of social media. With the increase in accessibility of opinion resource such as movie reviews, product reviews, blog reviews, social network tweets, and the new challenging task is to mine a large volume of texts and devise suitable algorithms. The algorithm understands the opinion of others. This information is of massive potential to companies which try to know the sentiment about their products or services [15]. This sentiment helps them in taking informed decisions. In addition to being useful for companies, the reviews and opinion mined from them, is helpful for users as well. For example, reviews about restaurant in a city may guide a user visiting that city. Similarly, movie reviews help other users in deciding whether the movie is worth watching or not.

Sentiments are actually emotions of users. It can be good, excellent, bad or neutral. The analysis of such emotions of users is sentiment analysis. It is a natural language processing task that uses computational approach to identify the opinion and classify it as positive, negative or neutral. The web contains the structured and unstructured textual information that often carries opinion or sentiments of the user. The sentiment analysis identifies the mood of writers and expressions of opinion.

The rest of this paper is structured as follows. Section II describes the different levels of analysis; Section III describes the classification and the paper concludes with key observations in Section IV.

II. LEVELS OF ANALYSIS

Sentiment analysis has been investigated mainly at three levels:

A. Document level –

The document level sentiment analysis classifies the entire document opinion into different sentiment, for a product or service. This level classifies opinion document into a positive, negative or neutral sentiment.

B. Sentence level –

The sentence level sentiment analysis determines whether each sentence expresses a positive, negative or neutral opinion, for a product or service. This type is used for reviews and comments that contain one sentence and written by the user [12].

C. *Entity and Aspect level –*

Aspect level is the opinion mining and summarization based on feature. The classification concerns by identifying and extracting product features from the source data. This type is used when we need sentiments about desired aspect/feature in a review.

III. CLASSIFICATION

Basically two main approaches for classification are: supervised and unsupervised. In supervised classification, the classifier is trained on labeled examples that are similar to the test examples. The unsupervised learning techniques assign labels based only on internal differences between the data points.

There are broadly three types for sentiment classification: (a) using a machine learning based text classifier -such as Naïve Bayes, SVM or KNN- with suitable feature selection scheme, Maximum Entropy, and Decision Trees; (b) using the unsupervised semantic orientation scheme such as K-Means clustering (c) using the SentiWordNet based publicly available library [2].

A. *Machine Learning Algorithms –*

The machibe learning algorithm is a branch of artificial intelligence. It focuses on building models that have the ability to learn from data.

"Machine Learning is a field of study, which gives computers the ability to learn without being explicitly programmed".

A supervised learning algorithm learns to map the input examples to expected target. The machine learning algorithm should be able to generalize the training data after the correct implementation of the training process, So that it can accurately map new data that it has never seen before.

Naïve Bayes

The Naïve Bayes classifier is a simple probabilistic model which relies on the assumption of feature independent in order to classify input data. The algorithm is commonly used for text classification [1] [16]. It has simpler implementation, low computational cost and its relatively high accuracy.

The algorithm will take every word in the training set and calculate the probability of it being in each class (positive or negative). Then the algorithm is ready to classify new data. When a new sentence is being classified it will split it into single word features. The model will use the probabilities, which computed in the training phase to calculate the condition probabilities of the combined features in order to predict its class [14]. The advantage of the Naïve Bayes classifier is that it utilizes all the evidence that is available to it in order to make a classification. Using this approach it takes into account that many weak features which may have relativistic minor effect individuals may have a much larger influence on the overall classification when combined [1][6].

Support Vector Machine

The support vector machine is considered as a non-probabilistic binary linear classifier. It works by plotting the training data in multidimensional space; it then tries to separate the classes with a hyperplane. If the classes are not immediately linearly separable in the multidimensional space the algorithm will add a new dimension in an attempt to further separate the classes. The SVM algorithm chooses the hyperplane which provides the maximum separation between the classes has the greatest margin or the maximal margin hyperplane which minimizes the upper bound of the classification errors. A standard method for finding the optimum way of separating the classes is to plot two hyperplanes in a way that there are no data points between them, and then by using these planes the final hyperplane can be calculated. The data points that fall on these planes are known as the supports.

A major problem with the SVM is that by adding extra dimensions the size of the feature space increases. From a processing point of view the SVM algorithm counteracts this by using dot products in the original space. This method hugely reduces processing as all the calculations are performed in the original space and then mapped to the feature space [1] [6].

Decision Tree

Decision trees on one of the most widely used machine learning algorithms which can be adapted to almost any type of data. It divides its training data into smaller parts in order to identify patterns that can be used for classification. Then, the knowledge is represented in the form of logical structure similar to a flow chart that can be

easily understood without any statistical knowledge. The algorithm is particularly used where many hierarchical categorical distinctions can be made. The structure of a decision tree consists of a root node which represents the entire dataset, decision nodes, which perform the computation and leaf nodes which produce the classification. In the training phase the algorithm learns what decisions have to be made in order to split the labeled training data into its classes [5] [7].

By passing data through the tree, unknown instance is classified. At each decision node a specific feature of the input data is compared with a constant that was identified in the training phase. The computation which takes place in each decision node usually compares the selected feature with this predetermined constant, the decision will be based on whether the feature is greater than or less than the constant, creating a two way split in the tree. The data will eventually pass through these decision nodes until it reaches a leaf node which represents its assigned class [1] [6].

B. Semantic Orientation –

The Semantic orientation approach to be unsupervised learning because it does not require prior training in order to mine the data. But, it measures how far a word is inclined towards positive and negative.

Many researches in the unsupervised sentiment classification make use of lexical resources available. Kamps et al, [11] used lexical relations in sentiment classification. Andrea Esuli and Fabrizio Sebastiani, [3] proposed semi-supervised learning method. It started from expanding an initial seed set using WordNet. Their basic assumption is termed with similar orientation tend to have similar glosses.

Sometimes the review cannot provide enough contextual information to determine the orientation of opinion. So Chunxu Wu, [9] proposed an approach which resort to other feedbacks discussing the same topic to mine useful contextual information. Then semantic similarity measures are used to check the orientation of opinion. By getting the orientation of context independent opinions, they attempted to handle this problem. Then consider the context dependent opinions using linguistic rules to assume orientation of context distinct-dependent opinion. And extract contextual information from other reviews which are on the same product feature to judge the context indistinct-dependent opinions. Ting-Chun Peng and Chia-Chun Shih, [17] investigated an unsupervised learning algorithm extracts the sentiment phrases of each review by rules of part-of-speech (POS) patterns. They used it as a query term to get top-N relevant snippets for each unknown sentiment phrase. The predictive sentiments of unknown sentiment phrases are computed after gathering sentiment lexicon. The predictive sentiments are based on the sentiments of nearby known sentiment words inside the snippets. They consider only opinionated sentences containing at least one detected the sentimental phrase for opinion extraction. The opinion extraction is performed using POS pattern. Gang Li & Fei Liu, [10] developed an approach based on the k-means clustering algorithm. The technique of TF-IDF (term frequency – inverse document frequency) weighting is applied to the raw data. Then, to extract more stable clustering result voting mechanism is used. The result is based on multiple implementations of the clustering process. Documents are combined into positive group and negative group. Chaovalit and Zhou, [8] compared the Semantic Orientation approach with other machine learning approach such as the N-gram model by applying to movie reviews. The result obtained by the machine learning approach is more accurate. But it requires a significant amount of time to train the model. The semantic orientation approach is less accurate, but is more efficient to use in real-time applications. The performance of semantic orientation depends on the performance of the underlying POS tagger.

C. SentiWordNet based approaches –

To make use of SentiWordNet these methods first extracts relevant opinionated terms and then look up for their scores in the SentiWordNet. There are four scoring schemes have implemented with the two feature selection variants, namely using adjectives only and using “adverb+adjective” combine. In order to evaluate the accuracy and performance of different variants of the SentiWordNet based approaches, they computed the standard performance metrics of Accuracy, F-measure and Entropy. They computed results of four SentiWordNet based approaches for two movie reviews and two blog post datasets. They have also compared results for movie review datasets with NB and SVM based machine learning classifiers The ease of implementation of SentiWordNet allows not only allows to perform sentiment analysis, but it also makes a very reasonable case of using it as an added level of filtering for movie recommendations.

SentiWordNet is used with document level sentiment classification combined with two linguistic features. SentiWordNet is a publicly available library that contains scores of each word and based on the score we classify the reviews as positive, negative or neutral opinion. The two linguistic features are: i) adverb and adjective combination and ii) adjective adverb and adverb verb combination [18]. This is used for producing better results. Aspect level is used when we consider specific feature of a movie like, direction, acting, cast, music, etc.

The adverb+adjective combination used for better result as compared to using only adjectives. Because adverbs increase the score and we can say that they play the role modifier. When we combine scores of adverb verb combined

with scores of adjective adverb combination than it improves the accuracy of sentiment classification. The more accurate or focused sentiment summary of particular movie is produced by aspect-level sentiment profile. Limitation of aspect level sentiment classification is that it is domain specific [4].

Martin Wollmer et al., [13] proposed method of sentiment classification for audio and video reviews of user. A movie review is given in 2 minute YouTube video. The automatic speech recognition system and video recognition system is used for sentiment classification of reviews. And vocal and face expression give better classification of reviews.

IV. COMPARATIVE STUDY

Sentiment analysis Techniques	Properties	Advantages	Disadvantages	Application
Naïve Bayes	- Figure out categorical class labels - Organizes data based on the training set and values in classifying attributes and uses it in classifying new data	- Easy to implement - Requires a few training data to estimate the parameters - Good result obtained in most of the cases	- Assumption :independence class conditions, so loss of accuracy - Practically, dependencies exist among variables	- approval of credit - medical diagnosis - target marketing - treatment effectiveness analysis
Support vector machine	- simultaneously minimize empirical classification error - maximize geometric margin(maximum margin cla ssifier)	- High prediction accuracy - Robust, although training examples contain errors, it works - Fast evaluation of the learned target function	- Understanding the learned function (weights) is difficult - Requires long training time - not easy to incorporate domain knowledge - expensive in both memory and computational time	- handwritten digit recognition - object recognition, speaker identification -benchmarking time-series prediction tests
Decision Tree	- Beneficial when the complexity of the problem grows - Used as visual aids to structure - Solve sequential decision problems	-simple to understand and interpret - can be combined with other decision techniques - have value even with little hard data - Addition of new possible scenarios can be allowed	- Information gain in decision trees are biased in support of those attributes with more levels. - When many values are uncertain and/or many outcomes are linked, the calculations can get very complex.	- Classifying cardiovascular outcomes
Semantic Orientation	- based on fixed syntactic patterns - computing sentiment polarity of a text	- Atomically identify antonyms - distinguish near synonyms	-does not readily extend beyond isolated adjectives to adverbs or longer phrases	- Summary statistics for search engines - Summarization of reviews - Filtering “flames” for newsgroups
SentiWordNet based approaches	- Lexical resource for	-Fast -No training data necessary	-Unable to work for multiple word phrases	-Review-related analysis (movies, social issues, hotels,

	sentiment analysis - Built on the top of WordNet synsets - With synsets, sentiment-related information attached	-good initial accuracy	-Unable to deal with multiple word senses	etc.) -Question-answering (Opinion-oriented questions may involve different treatment) - Developing 'hate mail filters' analogous to 'spam mail filters'

Supervised machine learning techniques have relatively better performance than the unsupervised methods. But, the unsupervised methods are also important because supervised methods demand large amounts of labeled training data. And that are very expensive whereas acquisition of unlabelled data is easy. Most domains lack labeled training data in this case unsupervised method is very useful for developing applications. Most of the researchers concluded that SVM has high accuracy than other algorithms. The limitation of supervised learning is that it generally requires large expert annotated training corpora which are created from scratch. And it specifically for the application at hand, and may fail when training data are insufficient.

From a machine learning point of view, while the predictive performance of the different algorithms SVM and Naïve Bayes comparable, the performance of the Decision Tree algorithm is far below the others.

The semantic orientation classifies reviews as recommended (thumbs up) or not recommended (thumbs down) where a review is predicted by the average semantic orientation of the phrases in the review that contain adjectives or adverbs. The limitations of this work compare to SentiWordNet, include the time required for queries and, for some applications, the level of accuracy that was achieved.

The recent sentiment classification is based on applying the SentiWordNet. The SentiWordNet lexical resource applied to the problem of automatic sentiment classification of reviews. It counts positive and negative term scores which determine sentiment orientation. And an improvement is presented by building a dataset of relevant features using SentiWordNet as source.

V.CONCLUSION

In this paper, it observed that in the decision making process about product, service, movie, social issues, sentiment analysis or opinion mining plays a very important role. Opinion mining is not only consisting of the concepts of text mining, but also the concepts of information retrieval. For good classification, feature weighting which plays a crucial role is one of the major challenges in opinion mining. Over the Web, social media is one of the major parts of it. Calculation says that every 9 users out of 10 use one form social media. Now a day's user over internet creates a large amount of data. So, for web content users become co-creators. Over social media, user contribution ranges from photo and video uploads, reviews, blog posts and tweets. On internet the data that is available is unstructured text.

Over social media, views or opinion is expressed through user reviews or posts. With demand growth about the accessibility of opinion resources such blog reviews, movie reviews, social network tweets, product reviews, results in new challenge is to mining large volume of data/text and required suitable algorithm for analysis of the opinion of others. This is important for the organizations because these help them to improve their services or goods and it also helps them for making decisions for the future.

REFERENCES

[1] Alec Go, Lei Huang, Richa Bhayani, "Twitter Sentiment analysis", s.l.: The Stanford Natural Language Processing Group, 2009.

[2] Alexander Pak, Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", s.l.: LREC, 2010.

[3] Andrea Esuli and Fabrizio Sebastiani, "Determining the semantic orientation of terms through gloss classification",Proceedings of 14th ACM International Conference on Information and Knowledge Management, Bremen, Germany, 2005.

[4] Asha S Manek, Pallavi R P, Veena H Bhat, P Deepa Shenoy, M Chandra Mohan, Veenugopal K R and L M Patnaik, "SentReP: Sentiment Classification of Movie Reviews using Efficient Repetitive Pre-Processing", 978-1-4799-2827-9/13 IEEE, 2013.

[5] Bifet, Albert, Eibe Frank, "Sentiment knowledge discovery in twitter streaming data". In: *Discovery Science*. s.l.:Springer Berlin Heidelberg, 2010.

[6] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques". Philadelphia, Association for Computer Linguistics, 2002.

[7] Castillo, Carlos, Marcelo Mendoza, Barbara Poblete, "Information credibility on twitter". AMC, Proceedings of the 20th international conference on World Wide Web, 2011.

[8] Chaovalit,Lina Zhou, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[9] Chunxu Wu, Lingfeng Shen , "A New Method of Using Contextual Information to Infer the Semantic Orientations of Context Dependent Opinions", International Conference on Artificial Intelligence and Computational Intelligence, 2009.

[10] Gang Li, Fei Liu, "A Clustering-based Approach on Sentiment Analysis", 978-1-4244-6793-8/10 IEEE, 2010.

[11] Kamps, Maarten Marx, Robert J. Mokken and Maarten De Rijke, "Using wordnet to measure semantic orientation of adjectives", Proceedings of 4th International Conference on Language Resources and Evaluation, pp. 1115-1118, Lisbon, Portugal, 2004.

[12] Kang Wu, Bofeng Zhang, Jianxing Zheng and Haidong Yao, "Sentiment Classification for Topical Chinese Microblog Based on Sentences Relations", IEEE International Conference on Green Computing and Communication and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, 2013.

[13] Martin Wollmer, Felix Weninger, Tobias Knaup, Bjorn Schuller, Congkai Sun, Kenji Sagae and Louis-Philippe Morency, "YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context", IEEE Computer Society,1541-1672, 2013.

[14] McCallum, Andrew, and Kamal Nigam, "A comparison of event models for naive bayes text classification", AAAI98 workshop on learning for text categorization, Volume 752, 1998.

[15] Mostafa Karamibekr and Ali A Ghorbani, "Sentiment analysis of Social Issues", ASE International Conference on Social Informatics, 978-0-7695-5038-2/12 IEEE, 2012.

[16] Prem Melville, Wojciech Gryc, Richard D. Lawrence, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification". s.l.,Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009.

[17] Ting-Chun Peng and Chia-Chun Shih , "An Unsupervised Snippet-based Sentiment Classification Method for Chinese Unknown Phrases without using Reference Word Pairs", IEEE/WIC/ACM International Conference on Web Intelligence and intelligent Agent Technology JOURNAL, 2010.

[18] V.K. Singh, R.Priyani, A. Uddin and P.Waila, "Sentiment Analysis of Movie Reviews and Blog Posts Evaluating SentiWordNet with different Linguistic Features and scoring schemes", 3rd IEEE International Advance Computing Conference (IACC), 978-1-4673-4529-3/12, 2012.