ISSN:0975-9646

A Survey on Feature Level Sentiment Analysis

Neha S. Joshi[#1], Suhasini A. Itkat[*2]

[#1]*Department of Computer Engineering,Pune University*
PES's Modern College of Engineering ,Pune ,India.

[*2]*Department of Computer Engineering,Pune University*
PES's Modern College of Engineering ,Pune ,India.

***Abstract* – In recent years we highly consider opinions of friends, domain experts for decision making in day today's life. For example, which brand is best for certain product, whether the current movie is good, whether product gives better performance or not etc. Opinion mining, also known as Sentiment analysis plays an important role in this process. It is the study of emotions i.e. Sentiments, Expressions that are stated in natural language. Natural language techniques are applied to extract emotions from unstructured data. There are several techniques used to analyse data such as supervised learning, unsupervised learning and hybrid techniques. The binary classification classifies data into two subclasses viz. positive and negative. The positive class shows a good opinion and negative class shows the bad decision opinion. This paper gives an overview of the proposed methods and recent advances in these areas, and tries to layout the future research directions in the field.**

Keywords— Machine learning, Polarity, Classification, Natural Language Processing (NLP), Sentiment Classification, Semi-Supervised learning, Support Vector Machine (SVM), Opinion Mining

1. INTRODUCTION

A vast amount of massive information is available in on-line documents or websites as well as on-line discussion groups, review sites. These are the sites where users post some crucial information about products or any other subjects. This information is nothing but sentiment or overall opinion about that subject- for example whether movie review is positive or negative. Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

With the explosive growth of social media (i.e., reviews, forum discussions, blogs and social networks) on the Web, individuals and organizations are increasingly using public opinions in these media for their decision making. However, finding and monitoring opinion sites on the Web and distilling the information contained in them remains a formidable task because of the proliferation of diverse sites. Each site typically contains a huge volume of opinionated text that is not always easily deciphered in long forum postings and blogs. The average human reader will have difficulty identifying relevant sites and accurately summarizing the information and opinions contained in them. Moreover, it is also known that human analysis of text information is subject to considerable biases, e.g., people often pay greater attention to opinions that are consistent with their own preferences. People also have difficulty, owing to their mental and physical limitations, producing consistent results when the amount of information to be processed is large. Automated opinion mining and summarization systems are thus needed, as subjective biases and mental limitations can be overcome with an objective sentiment analysis system.

This paper is organized as follows: Section-II different levels of sentiment analysis, section-III gives literature survey, section-IV describes analysis of some techniques proposed in literature and section-V will conclude the paper.

2. LITERATURE SURVEY

2.1 LEVELS OF ANALYSIS:

The analysis levels can be done at three levels namely document level, sentence level and Feature level analysis.

A. Document Level Analysis:

The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity. Thus, it is not applicable to documents which evaluate or compare multiple entities [1].

B. Sentence Level Analysis:

The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions. However, we should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions.

C. Feature Level Analysis:

Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier called feature level (feature-based opinion mining and summarization).Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level 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).

Following different types of machine learning techniques can be applied to analyze sentiment in documents.

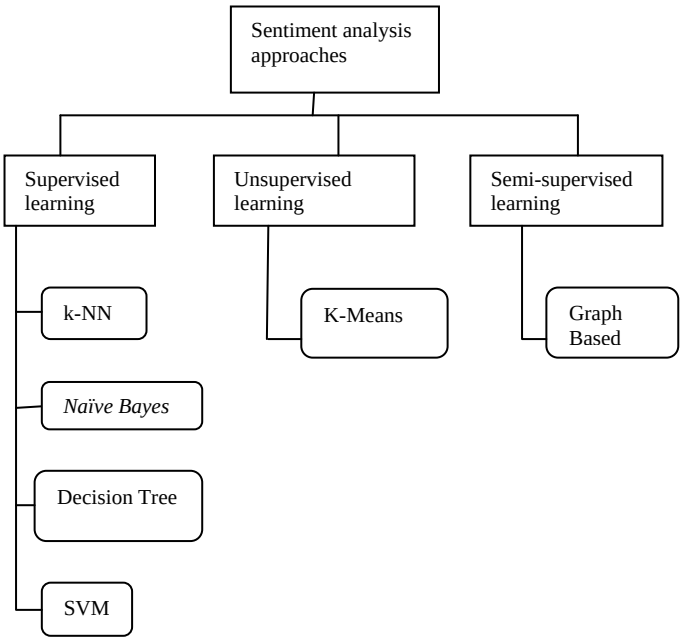

Fig.1 Approaches in Sentiment Analysis.

2.2 LEARNING ALGORITHMS

Mainly there are three categories of learning algorithms which are further divided into different sub-categories as shown in fig.1

2.2.1 Supervised Learning Algorithms

Some of the most predominant Supervised Learning techniques in Sentiment Analysis have been SVM, Naive Bayesian Classifiers and other Decision Trees [2].

2.2.1.1 Naïve Bayes

A Naïve Bayes classifier is a simple probabilistic model based on the Bayes rule along with a strong independence assumption. The Naïve Bayes model involves a simplifying conditional independence assumption [3]. That is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem. In this case, the maximum likelihood probability of a word belonging to a particular class is given by the expression:

$$p(Xi/C) = \frac{\mathrm{n(xi/C)}}{\mathrm{n(C)}}$$

Where

$\mathrm{n(xi/C)}$ = count of Xi in documents of class c

$\mathrm{n(C)}$ = no of words in documents of class c

The frequency counts of the words are stored in hash tables during the training phase. According to the Bayes Rule, the probability of a particular document belonging to a class Ci is given by,

$$p(Ci/d) = \frac{p(d/Ci)*p(Ci)}{p(d)}$$

2.2.1.2 Maximum Entropy

It is conditional exponential classifier. It maps each pair of feature set and its label to a vector. It is also called as log-linear classifier because they work by extracting some set of features from the input, combining them linearly (each feature is multiplied by its weight and added up) and using this sum as exponent. It is parameterized by a set of weights that are used to combine the joint-features that are generated from a set of features by an encoding19 [3].

2.2.1.3 Decision Tree

It is a tree in which internal nodes are represented by features, edges represent tests to be done at feature weights and leaf nodes represent categories which results from above tests. It categorizes a document by starting at the tree root and moving successfully downward via the branches (whose conditions are satisfied by the document) until a leaf node is reached. The document is then classified in the category that labels the leaf node. Decision Trees have been used in many applications in speech and language processing.

2.2.1.4 Support Vector Machines

In comparisons, SVM has outperformed other classifiers such as Naïve Bayes. While SVM has become a dominant technique for text classification [4], other algorithms such as Winnow and AdaBoost have also been used in previous sentiment classification studies. SVM gives highest accuracy results in text classification problems. SVM represents example as points in a space which are mapped to a high dimensional space where mapped examples of separate classes are divided by as wide as possible tangential possible distance to the hyper plane. New examples are mapped into this same space and depending on which side of the hyper plane they are positioned, they are predicted to belong to a certain class. SVM hyper planes are fully determined by a relatively small subset of the training instances, which are called support vectors. The rest of the training data have no influence on the trained classifier. SVMs have been employed successfully in text classification and in a variety of sequence processing applications [6].

Pang et al. [1] compared the performance of three classifiers Naïve Bayes, Maximum Entropy and Support Vector Machines in Sentiment classification at document level on different features like considering only unigrams, bigrams, combination of both, combining unigrams and parts of speech, taking only adjectives and combining unigrams and position information. The result has shown that feature presence is more important than feature frequency and when the feature set is small, Naïve Bayes performs better than SVM. But SVM's perform better when feature space is increased. When feature space is increased, Maximum Entropy may perform better than Naïve Bayes but it may also suffer from over fitting.

2.2.2 Unsupervised Learning Algorithms

These are also known as lexicon based techniques. This involve learning patterns in the input when no specific output values are supplied, this means that the learner only receives an unlabelled set of examples [7].Unsupervised methods can also be used to label a corpus that can later be used for supervised learning. An agent purely based on unsupervised

learning cannot lean what to do, because it has no information as to what constitutes a correct action or a desirable state. Examples of unsupervised learning methods are (k means) clustering or cluster analysis, the problem of discerning multiple categories in a collection of objects and the expectation-maximization algorithm, an algorithm for finding the maximum likelihood of examples. In Unsupervised techniques classification is done by comparing the features of a given text against sentiment lexicons whose sentiment values are determined prior to their use. Sentiment lexicon contains lists of words and expressions used to express people's subjective feelings and opinions. For example, start with positive and negative word lexicons, analyze the document for which sentiment need to find. Then if the document has more positive word lexicons, it is positive, otherwise it is negative. The lexicon based techniques to Sentiment analysis is unsupervised learning because it does not require prior training in order to classify the data [10].

There are three approaches to construct a sentiment lexicon: manual construction, corpus-base construction and dictionary based construction. In manual construction sentiments and its corresponding features are manually constructed. But it is difficult, time consuming and impractical task.

Corpus-base construction depends on pattern in large documents or corpora. It has major advantage over dictionary based construction that it can help find domain specific opinion words and their orientation. In dictionary base construction, initially a small set of opinion words is collected with their associated polarity and then it is grown by searching in WordNet dictionary for their synonyms and antonyms.

2.2.3. Hybrid Techniques

In Hybrid Techniques both combination of machine learning and lexicon base approaches ate used. Researchers have proved that this combination gives improved performance of classification.

Mudinas et al. [15] proposed a concept-level sentiment analysis system, called pSenti, which is developed by combining lexicon based and learning-based approaches. The main advantage of their hybrid approach using a lexicon/learning symbiosis is to obtain the best of both worlds-stability as well as readability from a carefully designed lexicon, and the high accuracy from a powerful supervised learning algorithm. Their system uses a lexicon from public resources for initial sentiment detection. They used sentiment words as features in machine learning method. The weight of such a feature is the sum of the sentiment value in the given review. For those adjectives which are not in sentiment lexicon, their occurring frequencies are used as their initial values. Their hybrid approach pSenti achieved 82.30% accuracy.

Fang et al. [16] incorporates a general purpose lexicon as well as domain specific lexicon into SVM learning. Later, they used this method for identifying both product aspects and their corresponding polarities. Experiment results show that a general purpose sentiment lexicon gives less accuracy improvement than domain specific dictionaries. Their system performed a two step classification. In step 1, a classifier is trained to predict the camera aspect being discussed. In step 2, a classifier is trained to predict the sentiment associated with that camera aspect. Finally, the two step prediction results are aggregated together to produce the final prediction. In both steps, the lexicon knowledge is incorporated into conventional SVM learning. They achieved 66.8% polarity accuracy.

Zhang et al. [14] employ an augmented lexicon-based method for entity level sentiment analysis. First extract some additional opinionated indicators (e.g. words and tokens) through the Chi-square test on the results of the lexicon-based method. With the help of the new opinionated indicators, additional opinionated tweets can be identified. Afterwards, a sentiment classifier is trained to assign sentiment polarities for entities in the newly identified tweets. The training data for the classifier is the result of the lexicon-based method. Thus, the whole process has no manual labelling. They achieved accuracy of 85.4%.

2.3 PROCESSING STEPS

There are standard methods involved in above techniques. Those are as follows:

2.3.1 Data Collection

The data collection (data RAW) from twitter, blogs or any website is collected for analysis.

2.3.2Preprocessing

Data cleaning and filtering: Since data contain several syntactic features that may not be useful for machine learning, the data needs to be cleaned such as @ (at) for link to username, url or link website (http, url, www), (hashtag), RT(for retweet). A module that allows option of different cleaning operations is designed.

2.3.3 Feature Selection and Extraction

2.3.3.1 Case Normalization

Most English texts (and other Romance languages) are published in combined case that is, published text contains both higher and lowercase characters. The process is to turn the entire document or sentences into lowercase one.

2.3.3.2Tokenization

Tokenization is splitting up the systems of text into personal terms or tokens. This procedure can take many types, with regards to the terminology being examined. For English, an uncomplicated and effective tokenization technique is to use white space and punctuation as token delimiters.

2.3.3.3 Stemming

Stemming is the procedure of decreasing relevant tokens into a single type. Typically the stemming procedure contains the recognition and elimination of prefixes, suffixes, and unsuitable pluralisation.

2.3.3.4 Generate n-Grams

Character n-grams are n nearby figures from a given feedback sequence. For example, a 3-gram of phrase TERM can be {T,-TE, TER, ERM etc. N-grams of single dimension is known as unigram, 2 dimension is known as bigrams and so on.

2.3.4 Feature Weighting

Term Frequency and Inverse Document Frequency (TF-IDF)-TF-IDF is a typical measurement used in text classification projects, but its use in opinion mining has been less extensive, and amazingly it does not look to have been used as a unigram feature weight. TF-IDF is consisting of two ratings, phrase regularity and inverse papers regularity. Term frequency is discovered by basically keeping track of frequent that a given phrase has took place in a given document, and inverse document frequency is discovered by splitting the amount of records that given term seems to be in. When these principles are increased together we get a ranking that is maximum for terms that appear regularly in a few records, and low for conditions that appear regularly in every document, enabling us to discover conditions that are essential in a document.

ANALYSIS AND CONCLUSION

Supervised machine learning techniques have shown better performance than unsupervised machine, earning techniques. However, the unsupervised methods is important too because supervised methods demand large amounts of labelled training data that are very expensive whereas acquisition of unlabelled

data is easy. Most domains except movie reviews lack labelled training data in this case unsupervised methods are very useful for developing applications.

Most of the researchers reported that Support Vector Machines (SVM) has high accuracy than other algorithms. The NB classifier produces the best results, followed by the DT classifier. The ME classifier performs poorly The main limitation of supervised learning is that it generally requires large expert annotated training corpora to be created from scratch, specifically for the application at hand, and may fail when training data are insufficient.

If the opinion words are included in dictionary then it must contain all words. It is important for lexicon based approach. Because it will reduce the performance if there are fewer words present in dictionary. Another significant challenge to this approach is that the polarity of many words is domain and context dependent. For example, ‘funny movie’ is positive in movie domain and ‘funny taste’ is negative in food domain. Such words are associated with sentiment in a particular domain.

REFERENCES

[1] Pang B., Lee, L., And Vaithyanathain. S. 2002.Thumbs up? Sentiment classification using machine learning techniques, In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* , 79-86.

[2] Sowmya Kamath S, Anusha Bagalkotkar, Ashesh Khandelwal, Shivam Pandey, Kumari Poornima,” Sentiment Analysis Based Approaches for Understanding User Context in Web Content”, 978-0-7695-4958-3/13, 2013 IEEE.

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

[4] Abd. Samad Hasan Basaria, Burairah Hussina, I. Gede Pramudya Anantaa, Junta Zeniarjab,” Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization”, Procedia Engineering 53 (2013) 453 – 462

[5] Rudy Prabowol, Mike Thelwall,” Sentiment Analysis: A Combined Approach”

[6] Michelle Annett and Grzegorz Kondrak, "A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs"

[7] B. Liu. Web Data Mining: Exploring hyperlinks, contents, and usage data," Opinion Mining. Springer, 2007

[8] B. Pang & L. Lee. “Opinion Mining and Sentiment Analysis." Foundations and Trends in Information Retrieval. Vol. 2, Nos. 1-2. pp.1-135, 2008

[9] Hogenboom, A.; van Iterson, P.; Heerschop, B.; Frasincar, F.Kaymak, U. , "Determining negation scope and strength in sentiment analysis," Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on , vol., no., pp.2589-2594, 9-12

[10] Kechaou, Z.; Ben Ammar, M.; Alimi, A.M.; , "Improving e-learning with sentiment analysis of users' opinions," Global Engineering Education Conference (EDUCON), 2011 IEEE , vol., no., pp.1032-1038, 4-6 April 2011

[11] Wenying ZHENG, Qiang YE. "Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm". Third International Symposium on Intelligent Information Technology Application,2009

[12] YuanbinWu, Qi Zhang, Xuanjing Huang, LideWu,” Phrase Dependency Parsing for Opinion Mining”. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 15331541,Singapore, 6-7 August 2009. c 2009 ACL and AFNLP

[13] Rudy Prabowo, Mike Thelwall. "Sentiment Analysis: A Combined Approach". White paper

[14] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B.Liu, combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis”, Technical report, HP Laboratories, 2011.

[15] A. Mudinas, D. Zhang, M. Levene, “Combining lexicon and learning based approaches for conceptlevel sentiment analysis”, Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY, USA, Article 5, pp. 1-8, 2012.

[16] Ji Fang and Bi Chen, “Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classication”, In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011.