International Conference On Emanations in Modern Technology and Engineering (ICEMTE-2017) ISSN: 2321-8169
Volume: 5 Issue: 3 151 - 155

A Survey on Sentiment Analysis Algorithms and Techniques

Prajakta Gosavi[1]
Department of Computer Engineering,
Shree L.R. Tiwari College of Engineering,
Mira Road, Thane, India
prajakta777@yahoo.co.in

Vaishali Shirsath[2]
Department of Information Technology,
Vidyavardhini College of Engineering and Technology,
Vasai, India
vaishalishirsath1@gmail.com

Abstract **-** Sentiment Analysis (SA) is an ongoing field of research in text mining. SA is the computational linguistic treatment of opinions, sentiments and subjectivity of text. This survey paper tackles a comprehensive overview of the last update in this field.There are various articles which categorized according to their contributions in the various SA techniques. This fields includes transfer learning, emotion detection, and building resources and applied to reviews and social media for different applications ranging from marketing to customer service.The main goal of this survey is to give nearly big image of SA techniques and the related fields with brief details. The main contributions of this paper include the sophisticated categorizations of a large number of recent articles and the illustration of the recent trend of research in the sentiment analysis or opinion mining and its related areas.

Index Terms **-** *Emotion detection, Feature selection, Sentiment analysis, Sentiment classification*

*****

1 INTRODUCTION

Sentiment Analysis or Opinion Mining is the computational study of people's opinions, their attitudes and emotions toward an entity. The entity can represent individuals, events or topics. These events are likely to be covered by reviews. Sentiment analysis (SA) and Opinion mining (OM) expresses a mutual meaning. However, some experts stated that OM and SA have slightly different notions [1]. Opinion Mining extracts and analyzes indiuals opinion about an entity while Sentiment Analysis identifies subjective information from source material. Therefore, the main aim of Sentiment analysis(SA) is to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of given document. The target of SA is to find opinions, identify the sentiments they express, and then classify their polarity as shown in Fig. 1.

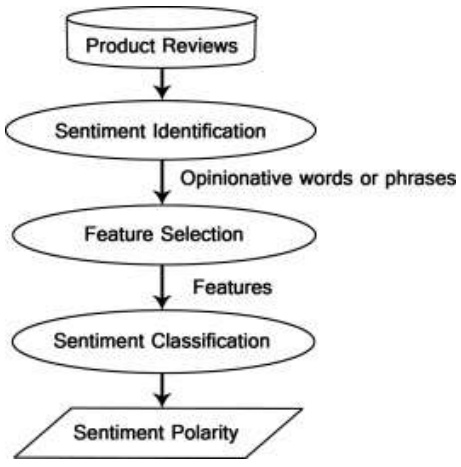

Figure 1. Sentiment analysis process on product reviews.

Sentiment Analysis has a classification process as illustrated in Fig.1. There are three classification levels in SA: i) document-level ii) sentence-level and iii) aspect-level SA. Aim of document-level is to classify an opinion document as illustrating a positive or negative opinion or sentiment. Sentence-level SA aims to classify sentiment expressed in each sentence. The initial stage is to identify whether the sentence is subjective or objective. If the given sentence is subjective then sentence-level SA will determine whether the sentence expresses positive or negative opinions. Wilson et al. [2] have mentioned that sentiment expressions are not necessarily subjective in nature. However, there is no such difference between document and sentence level classifications because sentences are just short documents [3]. Aspect-level SA aims to classify the sentiment with respect to the specific aspects of entities. The first step is to identify
the entities and their aspects.Indiuals can give different opinions for different aspects of the same entity like this sentence "*The voice quality of this phone is not good, but the battery life is long"*. This survey tackles the first two kinds of SA.

The data sets used in SA are the major issue in respective field. The main sources of data are from the product reviews. These reviews are important to the business holders as they can take business decisions with the help of analysis results of users opinions about their products. The sources are mainly review sites. SA is not only applied on product reviews but can also be applied on stock markets [4] and [5], news articles, [6] or political debates [7]. In political debates for example, we could figure out people's opinions on a certain election candidates or political parties. The election results can also be predicted from political posts. The social network sites and micro-blogging sites are considered a very

good source of information because number of people share and discuss their opinions about a certain topic freely.

There are varieties of applications and enhancements on SA algorithms that were proposed in the last few years.This study aims to give a closer look on these enhancements and to summarize and categorize some articles presented in same field according to the various SA techniquesThey are categorized with respect to the target of the article illustrating the algorithms and information used in their work. According to Fig. 1, the authors have discussed the Feature Selection (FS) techniques in details along with their related articles referring to some originating references. The Sentiment Classification (SC) techniques, as shown in Fig. 2, are discussed with more details illustrating related articles and originating references as well.

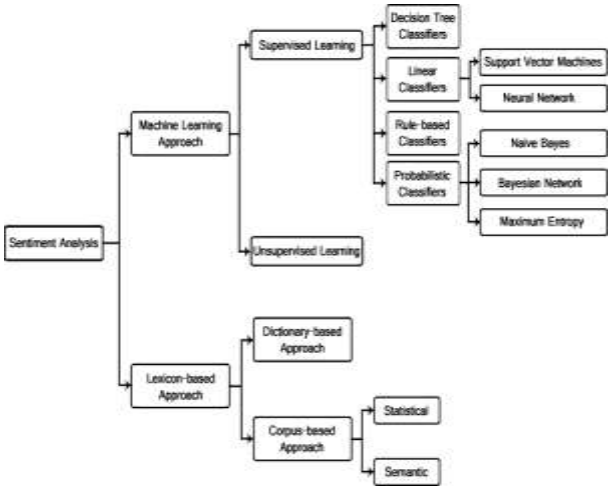

Figure 2. Sentiment classification techniques.

This survey can be useful for upcoming researchers in this field as it covers the most famous SA techniques and applications in one research paper. This survey uniquely gives categorization to the various SA techniques which is difficult to found in other surveys. It discusses also new related fields in SA .These fields include Emotion Detection (ED), Building Resources (BR) and Transfer Learning (TL). The main aim of emotion detection is to extract and analyze emotions, while the emotions could be explicit or implicit in the sentences. Transfer learning is concerned with analyzing data from one domain and then using the results in a target domain. Building Resources aims at creating lexica, corpora in which opinion expressions are annotated according to their polarity, and sometimes dictionaries.

2 Related Work

There are number of articles presented every year in the Sentiment analysis fields. These articles is increasing through years. This creates a need to have survey papers that summarize the recent research trends and directions of SA. The reader can find some easy way and detailed surveys including [1], [3], [8], [9], [10] and [11]. Those surveys have discussed the problem of SA from the applications point of view not from the SA techniques point of view.

We have two long and detailed surveys presented by Pang and Lee [8] and Liu [3]. They focused on the applications and challenges in SA. They explained the techniques used to solve each problem in SA. Cambria and Schuller et al. [9], Feldman [10] and Montoyo and Martínez-Barco [11] have given short surveys illustrating the upcoming trends in SA. Tsytsarau and Palpanas [1] have presented a survey which discussed the main topics of SA in details. For each topic they have illustrated its definition, problems and development and categorized the articles with the aid of tables and graphs. The analysis of the articles presented in this survey is similar to what was given by [1] but with another perspective and different categorization of the articles.

3 Feature Selection

SA task is considered a sentiment classification problem. The first step in the SC problem is to extract and select text features. Some of the features are[12]:

Terms presence and frequency: These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word exists, or one if otherwise) or uses term frequency weights to indicate the relative importance of features.

Parts of speech (POS): Finding adjectives.They are important indicators of opinions.

Opinion words and phrases: These are words normally used to express opinions including good or bad, like or hate. Some of them express opinions without using opinion words. For example: cost me an arm and a leg.

Negations: The appearance of negative words may change the opinion orientation like not good is nothing but bad.

4 Sentiment Classification Techniques

SC techniques can be divided into machine learning approach, lexicon based approach and hybrid approach. The Machine Learning Approach (ML) applies the famous ML algorithms and uses linguistic features. The Lexicon-based Approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. It is divided into dictionary-based approach and corpus-based approach which use statistical or semantic methods to find sentiment polarity. The hybrid Approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.

4.1 Machine learning approach

ML approach can be roughly divided into supervised and unsupervised learning methods.This approach relies on the famous ML algorithms to solve the SA as a regular text classification problem that makes use of syntactic and/or linguistic features.

Text Classification Problem Definition: We have a set of training records $D = \{X_1, X_2, \ldots, X_n\}$ where each record is

labeled to a class. Then for a given instance of unknown class, the model is used to predict a class label for it. The hard classification problem is when only one label is assigned to an instance. The soft classification problem is when a probabilistic value of labels is assigned to an instance.

4.1.1Supervised learning

This methods depend on the existence of labeled training documents. There are many kinds of supervised classifiers in literature.

4.1.1.1.Probabilistic classifiers

This type of classifier use mixture models for classification. The mixture model assumes that each class is a component of the mixture. Each mixture component is a generative model that provides the probability of sampling a particular term for that component.

4.1.1.1.1. Naive Bayes Classifier (NB)

The Naive Bayes classifier is the simplest and commonly used classifier.This classification model computes the probability of a class, based on the distribution of the words in the document.It uses Bayes Theorem to predict the probability that a given feature set belongs to a particular label.

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)} \quad (1)$$

P(label) is the prior probability of a label. P(features|label) is the prior probability that a given feature set is being classified as a label. P(features) is the prior probability that a given feature set is occurred.

4.1.1.1.2. Bayesian Network (BN)

The main assumption of the NB classifier is the independence of the features. The other main assumption is to assume that all the features are fully dependent. BN was used by Hernández and Rodríguez [13] to consider a real-world problem in which the attitude of the author is characterized by three different target variables. They proposed the use of multi-dimensional Bayesian network classifiers. They extended the multi-dimensional classification framework to the semi-supervised domain in order to take advantage of the huge amount of unlabeled information available in this context.

4.1.1.1.3. Maximum Entropy Classifier (ME)

The Maxent Classifier converts labeled feature sets to vectors using encoding. This vector is then used to calculate weights for each feature.This classifier is parameterized by a set of X{weights}, which is used to combine the joint features that are generated from a feature-set by an X{encoding}. In particular, the encoding maps each C{(featureset, label)} pair to a vector. The probability of each label is then computed using the following equation:

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l)) for l in labels)} \quad (2)$$

ME classifier was used by Kaufmann [14] to detect parallel sentences between any language pairs with small amounts of training data. Their results showed that this classifiers can produce useful results for any language pair. This can allow the creation of parallel corpora for many new languages.

4.1.1.2. Linear classifiers

There are many kinds of linear classifiers; among them is Support Vector Machines (SVM)[15]which is a form of classifiers that attempt to determine good linear separators between different classes. Two of the most famous linear classifiers are discussed in the following subsections.

4.1.1.2.1. Support Vector Machines Classifiers (SVM)

This classifiers works on digital cameras and MP3 reviews.The clustering algorithm which provides an improvement to the support vector machines is called **support vector clustering** and is used in industrial applications when data are not labeled or when only part of data are labeled as a preprocessing for a classification pass.

4.1.1.2.2. Neural Network (NN)

Multilayer NN are used for non-linear boundaries. These multiple layers are used to induce multiple piece-wise linear boundaries, which are used to approximate enclosed regions belonging to a particular class. The outputs of the neurons in the earlier layers feed into the neurons in the later layers.

4.1.1.3. Decision tree classifiers

This classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data. The predicate is the presence or absence of one or more words. The division of the data space is done continuously until the leaf nodes contain certain minimum numbers of records which are used for the purpose of classification.

4.1.1.4. Rule-based classifiers

In this classifiers, the data space is modeled with a set of rules. The LHS represents a condition on the feature set expressed in disjunctive normal form while the RHS is the class label. The conditions are on the term presence. Term absence is rarely used because it is not informative in sparse data.

4.2. Lexicon-based approach

Opinion words are used in many sentiment classification tasks. Positive opinion words are used to express some desired states, while negative opinion words are used to express some undesired states. There are three approaches in

order to compile or collect the opinion word list. Manual approach is very time consuming and it is usually combined with the other two automated approaches as a final check to avoid the mistakes that resulted from automated methods. The two automated approaches are as follows.

4.2.1. Dictionary-based approach

A small set of opinion words is collected manually with known orientations. Then, this set is grown by searching in the well known method WordNet [16]. The newly found words are added to the seed list then the next iteration starts. The iterative process stops when no new words are found. After the process is completed, manual inspection can be carried out to remove or correct errors.

This approach has a major disadvantage which is the inability to find opinion words with domain and context specific orientations. Qiu and He [12] used dictionary-based approach to identify sentiment sentences in contextual advertising. They proposed some advertising strategy to improve ad relevance and user experience. Their results shows the effectiveness of the proposed approach on advertising keyword extraction and ad selection.

4.2.2. Corpus-based approach

This approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus.

Using the corpus-based approach alone is not as effective as the dictionary-based approach because it is difficult to prepare a huge corpus to cover all English words, but this approach has a major advantage that can help to find domain and context specific opinion words and their orientations using a domain corpus. The corpus-based approach is performed using statistical approach or semantic approach as illustrated in the following subsections:

4.2.2.1. Statistical approach

Finding co-occurrence patterns or seed opinion words can be done using statistical techniques. This could be done by deriving posterior polarities using the co-occurrence of adjectives in a corpus, as proposed by Fahrni and Klenner [17]. It is possible to use the entire set of indexed documents on the web as the corpus for the dictionary construction. This overcomes the problem of the unavailability of some words if the used corpus is not large enough [18]. Latent Semantic Analysis (LSA) is a statistical approach which is used to analyze the relationships between a set of documents in order to produce a set of meaningful patterns.

4.2.2.2. Semantic approach

The Semantic approach gives sentiment values directly and relies on different principles for computing the similarity between words. WordNet for example provides different kinds of semantic relationships between words used to calculate sentiment polarities. WordNet could be used too for obtaining a list of sentiment words by iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word [16].

The Semantic approach is used in many applications to build a lexicon model for the description of verbs, nouns and adjectives to be used in SA as the work presented by Maks and Vossen [7]. Their model described the detailed subjectivity relations among the actors in a sentence expressing separate attitudes for each actor.

5 Conclusion

This survey paper presented an overview on the recent updates in SA algorithms and applications. Mentioned articles give contributions to many SA related fields that use SA techniques for various real-world applications. After analyzing these articles, it is clear that the enhancements of Sentiment classification and Feature selection algorithms are still an open field for research. Naïve Bayes and Support Vector Machines are the most frequently used ML algorithms for solving SC problem.

References

[1] Tsytsarau, Mikalai, and Themis Palpanas. "Survey on mining subjective data on the web." Data Mining and Knowledge Discovery 24.3 (2012): 478-514.

[2] Wilson,Theresa,JanyceWiebe,and Paul Hoffmann."Recognizing contextual polarity in phrase-level sentimentanalysis." Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005.

[3] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.

[4] Yu, Liang-Chih, et al. "Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news." Knowledge-Based Systems 41 (2013): 89-97.

[5] Hagenau, Michael, Michael Liebmann, and Dirk Neumann. "Automated news reading:Stock price prediction based on financial news using context-capturing features." Decision Support Systems 55.3 (2013): 685-697.

[6] Xu,Tao,Qinke Peng, and Yinzhao Cheng. "Identifying the Semantic Orientation of terms using S-HAL for

sentiment analysis." Knowledge-Based Systems 35 (2012): 279-289.

[7] Maks,Isa, and Piek Vossen. "A lexicon model for deep sentimet analysis and opinion mining applications." Decision Support Systems 53.4 (2012): 680-688.

[8] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval 2.1– 2 (2008): 1-135.

[9] Cambria, Erik, et al. "New avenues in opinion mining and sentiment analysis." IEEE Intelligent Systems 28.2 (2013): 15-21.

[10] Feldman, Ronen. "Techniques and applications for sentiment analysis." Communications of the ACM 56.4 (2013): 82-89.

[11] Montoyo, Andrés, Patricio MartíNez-Barco, and AlexandraBalahur."Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments." (2012): 675-679.

[12] Bhonde, Swati B., and Jayashree R. Prasad. "Sentiment Analysis-Methods,Applications & Challenges." International Journal of Electronics Communication and Computer Engineering 6.6 (2015): 634.

[13] Ortigosa-Hernández, Jonathan, et al. "Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers." Neurocomputing 92 (2012): 98-115.

[14] Kaufmann, Max. "JMaxAlign: A Maximum Entropy Parallel Sentence Alignment Tool." COLING (Demos). 2012.

[15] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273-297.

[16] Miller, George A., et al. "Introduction to WordNet: An on-line Lexical database." International journal of lexicography 3.4 (1990): 235-244.

[17] Fahrni, Angela, and Manfred Klenner. "Old wine or warm beer:Target-specific sentiment analysis of adjectives." Proc. of the Symposium on Affective Language in Human and Machine, AISB. 2008.

[18] Turney, Peter D. "Thumbs up or thumbs down?: semantic Orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics(2002).