International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Classification Techniques: A Survey

Sneha Pradeep Vanjari[1], V. D. Thombre[2]

[1]Department of Computer Engineering, SKN-Sinhagad Institute of Technology & Science, Lonavala, India

[2]Head of Department, Computer Engineering, SKN-Sinhagad Institute of Technology & Science, Lonavala, India

Abstract: ***An ongoing field of research in the topic of natural language processing is called as Sentiment Analysis. Sentiment analysis (SM) is the process of extraction of information from a set of documents that are related to a specific entity.Depending on text classification and processing, user identify the polarity of given text. There are many classification techniques. Use of particular algorithm depends on the kind of input provided. Understanding and analyzing when to use which algorithm is an important task and it helps in improving accuracy of results.The main contribution of this paper to give brief description about sentimental analysis techniques and base for our future roadmap.***

Keywords: sentimental classification, sentimental analysis, feature selection, opinion mining

1. Introduction

Recently many researchers are working on the fields of natural language processing. This topic is a combination of computer science and artificial intelligence. But, the most interesting topic of natural language processing is sentimental analysis. Sentiment analysis (SM) is the process of extraction of information from a set of documents that are related to a specific entity. Opinion Mining (OM) is somewhat similar to SM but with a slight difference. Opinion Mining is used to extract and analyze people's opinions whereas SM identifies the sentiments expressed in a text document that analysis it. Sentiment Analysis is used to know the people's opinion about a specific entity. The use of sentimental analysis is done for marketing and consumer research purpose. When the company launches a product, it needs a feedback related to the product from the end users. Here, sentimental analysis plays its role. It gets the customer's feedbacks about the product launched, the political campaigns and even of the financial markets [1]. Sentimental analysis is used to determine the attitude of aindividual who speaks or writes its opinion related to a specific topic. The authors in [2], [3] have studied on this topic earlier, where they used many different methods to detect the antithesis of a specific product and movie reviews.

2. Classification of Sentimental Analysis

Sentimental Analysis is been classified in three levels. These levels are: document-level sentimental analysis, sentence-level sentimental analysis and aspect-level sentimental analysis. In a document-level sentiment analysis, an opinion document is analyzed and then classified into positive or negative or sentiments. The opinion document is considered as a basic information unit which is giving information related to one specific entity. In sentence-level SA, each sentence is checked and the sentiments in these sentences are classified. The very first step in this is to identify whether the sentence is subjective or objective. We then analyze whether the sentence gives a positive opinion or a negative opinion if it is a subjective sentence. But it is not always the case where the sentences are subjective [4]. The document-level SA and sentence-level SA are very little different from each other because sentences are just the short documents[5]. In many applications only the document-level SA and sentence-level SA is not enough in order to gain the detailed opinions on each and every aspect of the entity. Hence, an aspect-level SA is needed.In the aspect-level SA, the sentiment classification is done with respect to the specific aspects and specific properties of an entity. Therefore, we first need to recognize the entity and spot the different aspects of the entity. Then the opinion givers and then give their opinions, where there may be different opinions for different aspects of the entity.

3. Sentiment Classification Technique

There are roughly three different approaches which can be used for dividing the sentiment classification techniques [6]. Approaches namely, machine learning approach, lexicon based approach and the hybrid approach. In the Machine Learning Approach (ML), the linguistic features are used along with the well known ML algorithm. Sentiment lexicons are used in the Lexicon based Approach. These sentiment lexicons are nothing but the group of known and precompiled sentiments. This approach is further divided into two methods, dictionary-based approach and corpus-based approach. In these methods, statistical or semantic methods are used to identify the sentiment duality. As the name indicates, hybrid approach is a combination of first two approaches. The term sentiment lexicons play a vital role in majority of methods.

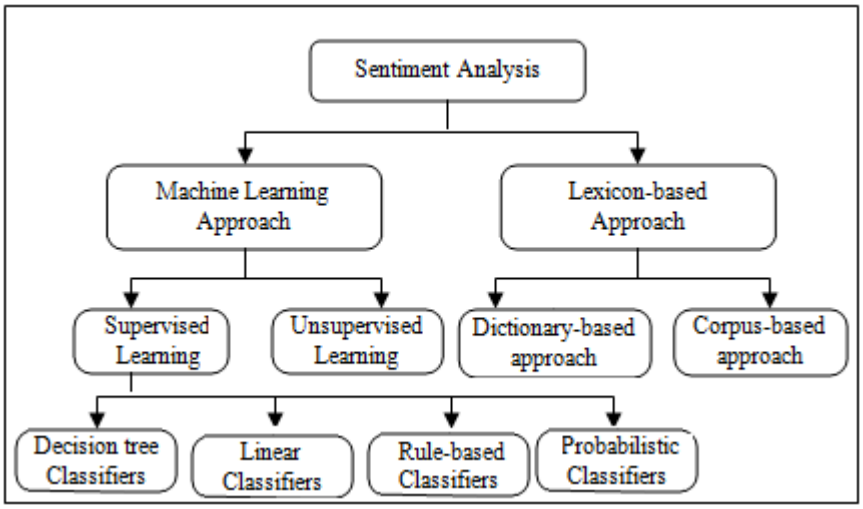

Figure 1.1: Techniques of sentiment classification

The text classification process which uses the machine learning approach can also be divided into two different methods, supervised and unsupervised learning approaches. The first method uses a very large number of labeled training documents. The second method is used when the labeled training documents are hard to find.

As we have discussed, sentiment lexicons are important in lexicon based sentiment classification. Finding an opinion lexicon which is perhaps used to analyze the text is used in lexicon-based approach. This approach further consists of two methods namely the dictionary-based approach and corpus-based approach. In the dictionary-based approach, the analysis depends on finding the basic opinion seed word, and then searching the dictionary containing the related antonyms and the synonyms.

In the corpus-based approach, the process starts with a list known as seed list which consists of some opinions words. And then search other opinion words which are context specific inclination. This later search is done in a large corpus. The summary of the techniques are given in figure 1.1.

A. Machine Learning Approach

The prominent ML algorithm is used for the machine learning approach. It is used to sentiment analysis for a regular text classification challenge that results into proper syntax and lingual trait.

*Text Classification Problem Definition:*Let there be a set of training records as D. This record can be illustrated as D = $\{X_1,X_2...X_n\}$. In this set each record is tagged to a class. The sentiment analysis is done on record consisting of features of the class to which the record is labeled or tagged. Let's consider a situation where the analysis encounters an unknown class, then the classification model is used to anticipate a class label. In this technique, there are two important terms, hard classification problem and soft classification problem. When there is only a label assigned to an instance then it is called as hard classification problem. And when there is a probabilistic values of labels assigned to an instance, then it is called as soft classification problem.

1. Supervised Learning

The extent of labeled training documents has made the supervised learning approach dependable on it. There are varieties of supervised classifiers. Some of the used classifiers in SA are described in the next section.

1.1 Probabilistic classifiers
The probabilistic classifiers make use of the mixture models for classification. These classifiers are also known as generative classifiers. The class is a component of the mixture, this is an assumption made by the mixture model. Each mixture component supplies a sampling of a particular term for the component.

1.2 Linear classifiers
Let there be a normalized document of word frequency as $X = \{x_1,x_2,...,x_n\}$, a vector of linear coefficients given as $A= \{a_1,....,a_n\}$, and a scalar as b. The linear classifiers give the output as the results of the linear predictor $p = X . A + b$. Support Vector Machines (SVM) [7,8] are one of the kinds of linear classifiers. The predictor is a separating hyperplane between different classes. SVM is the classifiers that attempts to decide the good linear separators between different classes. The two of the most well known linear classifiers are SVM and Neural networks.

1.3 Decision Tree classifiers
The training data spaces are hierarchically decomposed by the decision tree classifiers. In this decomposition process, the conditions on the attribute values are used to split the data [9]. The presence and absence of one or more words is on the conditions or the predicates. The splitting of the data space or the decomposition of the data is always carrier out in recursive manner to the stage when the leaf nodes may contain at least some number of records which are then used for classification. The splitting can be done in various ways like single attribute split, similarity-based multi-attribute split, and discriminate-based multi-attribute split. In a tree, the presence and absence of a particular word or a phrase at a particular node decides the splitting of the data [10]. The splitting done on documents where the similarity between the frequently used words clusters and the documents are used for the process of splitting, is called as similarity-based multi-attribute split.In discriminate-based multi-attribute

split, there are different discriminants used for the process of splitting, and one of the discriminant are fisher discriminate [11].

1.4 Rule-based classifiers

In this method, the data space is shaped with a set of rules. There are two sides of representation; the left side represents the conditions on the feature set while the right side represents the class label.All the rules are generated based on some criteria. The training phase constructs all the rules depending on some of the criteria. Support and confidence are the two most known criteria's [12].

2. Unsupervised Learning

Classification of documents into a random number of predefined categories is the main purpose of text classification. As discussed, for text classification, a supervised learning is carried out on a large number of labeled training documents. But sometimes it is easier to create unlabeled documents than creating the labeled documents. This is dealt by the unsupervised learning method. The author of [13] has presented the work in this field, which includes the division of the documents into the sentences and then categorizing each sentence using the keywords list of different categories. The unsupervised approach is used in [14], in order to automatically locate the facet discussed in Chinese social reviews as well as to recognize the sentiments expressed in different facets. They used LDA model to discover multi-aspect global topics of social reviews, then they extracted the local topic and associated sentiment based on a sliding window context over the review text. Another approach for unsupervised classification depends on semantic orientation using PMI [15] or lexical association using PMI.

B. Lexicon Based Approach

Many sentiment classification tasks make use of opinion words. There are always two kinds of opinions, positive and negative. Positive opinion words are used to express some desire, while the negative opinions are used to express anything undesired. There are also opinion lexicons that are combination of opinion phrases and idioms. To compile and correct the opinion word list there are three different approaches: manual approach, and two automated approaches. The manual approach is never used singly and it is time consuming. It is always conjunct with two automated approaches.

1. Dictionary-based approach.

The main strategy of this approach is presented in [16,17]. With the known orientation a set of opinion words is composed manually. Then this set is grown by searching in the well known corpora WordNet [18] or thesaurus [19] for their synonyms and antonyms. The seed list is always added with the newly founded words and then the next iteration starts. The iteration stops when no new word is found. To remove or correcting the errors manual inspection can be carried out.

2. Corpus-base approach

For finding the opinion words with context specific orientations, corpus-based approaches are used. Its methods include synthetic patterns or the patterns that occur together along with a seed list of opinion words to find otheropinion words in a large corpus. The author in [20] represents one of these methods. In this work a list of seed opinion adjectives is used initially, and then using the list with a set of linguistic constraints. This combinational use is done for recognizing more adjective opinion words along with their orientations. The connectives like AND, BUT, OR, etc are considered as the constrains, because for instance AND represents the conjoined adjectives that in fact has the same orientations. This proposition is called as sentiment consistency, as it is not always consistent. In order todetermine if two conjoined adjectives are of the same or differentorientations, learning is applied to a large corpus. Then,the links between adjectives form a graph and clustering is performedon the graph to produce two sets of words: positiveand negative words.

4. Conclusion

Text mining field is growing prominently and sentiment analysis has become the most famed topic. This paper gives a comprehensive overview on the classification techniques of sentiment analysis. Every type of classification model always has its own pros and cons. The very important benefaction of this paper includes the brief discussion on the semantic analysis classification techniques.

References

[1] Devitt, Ann, Khurshid Ahmad. "Sentiment polarity identification in financial news: A cohesion-based approach", Annual Meeting-Association for Computational Linguistics, Vol. 45, No. 1, 2007.
[2] Manning, Christopher D., PrabhakarRaghavan, HinrichSchütze, "Introduction to information retrieval", Vol. 1. Cambridge: Cambridge University Press, 2008.
[3] Kennedy, Alistair, Diana Inkpen, "Sentiment classification of movie reviews using contextual valence shifters", Computational Intelligence 22.2: 110-125, 2006.
[4] Wilson T, Wiebe J, Hoffman P.,"Recognizing contextual polarityin phrase-level sentiment analysis",in: proceedings of HLT/EMNLP; 2005.
[5] Liu B.,"Sentiment analysis and opinion mining", Synth LectHuman Lang Technol 2012.
[6] Diana Maynard, Adam Funk,"Automatic detection of politicalopinions in tweets", In: Proceedings of the 8th internationalconference on the semantic web, ESWC'11; p 88–99, 2011.
[7] Cortes C, Vapnik V. Support-vector networks, presented at theMachine Learning; 1995.
[8] Vapnik V.,"The nature of statistical learning theory", New York;1995.
[9] Quinlan JR. Induction of decision trees. Machine Learn, 1:81–106, 1986.
[10] Lewis David D, Ringuette Marc,"A comparison of two learningalgorithms for text categorization", SDAIR 1994.
[11] ChakrabartiSoumen, Roy Shourya, Soundalgekar Mahesh V., "Fast and accurate text classification via multiple linear discriminant projections", VLDB J 2003;2:172–85.

[12] Liu Bing, Hsu Wynne, Ma Yiming. Integrating classification andassociation rule mining. In: Presented at the ACM KDDconference; 1998.
[13] KoYoungjoong, SeoJungyun, "Automatic text categorization by unsupervised learning", In: Proceedings of COLING-00, the 18th International conference on computational linguistics; 2000.
[14] Xianghua Fu, Guo Liu, YanyanGuo, Zhiqiang Wang,"Multiaspectsentiment analysis for Chinese online social reviews basedon topic modeling and HowNet lexicon", Knowl-Based Syst2013;37:186–95.
[15] Turney P.,"Thumbs up or thumbs down?: semantic orientationapplied to unsupervised classification of reviews", In: Proceedingsof annual meeting of the Association for ComputationalLinguistics (ACL'02); 2002.
[16] Hu Minging, Liu Bing,"Mining and summarizing customerreviews", In: Proceedings of ACM SIGKDD international conferenceon Knowledge Discovery and Data Mining (KDD'04);2004.
[17] Kim S, Hovy E, "Determining the sentiment of opinions" In:Proceedings of interntional conference on Computational Linguistics(COLING'04); 2004.
[18] Miller G, Beckwith R, Fellbaum C, Gross D, Miller K, "WordNet: an on-line lexical database", Oxford Univ. Press; 1990.
[19] Mohammad S, Dunne C, Dorr B,"Generating high-coveragesemantic orientation lexicons from overly marked words anda thesaurus", In: Proceedings of the conference on EmpiricalMethods in Natural Language Processing (EMNLP'09);2009.
[20] Hatzivassiloglou V, McKeown K,"Predicting the semantic orientation of adjectives", In: Proceedings of annual meeting ofthe Association for Computational Linguistics (ACL'97); 1997.