Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 87 (2016) 44 – 49

Procedia Computer Science

Fourth International Conference on Recent Trends in Computer Science & Engineering.

Chennai, Tamil Nadu, India

Sentiment Analysis:A Comparative Study On Different Approaches

Devika M D[a]*, Sunitha C[a], Amal Ganesh[a]

[a]*Department of CSE, Vidya Academy of Science and Technology, Thrissur 680501, India*

Abstract

Sentiment analysis (SA) is an intellectual process of extricating user's feelings and emotions. It is one of the pursued field of Natural Language Processing (NLP). The evolution of Internet based applications has steered massive amount of personalized reviews for various related information on the Web. These reviews exist in different forms like social Medias, blogs, Wiki or forum websites. Both travelers and customers find the information in these reviews to be beneficial for their understanding and planning processes. The boom of search engines like Yahoo and Google has flooded users with copious amount of relevant reviews about specific destinations, which is still beyond human comprehension. Sentiment Analysis poses as a powerful tool for users to extract the needful information, as well as to aggregate the collective sentiments of the reviews. Several methods have come to the limelight in recent years for accomplishing this task. In this paper we compare the various techniques used for Sentiment Analysis by analyzing various methodologies.

Keywords: Sentiments*;* Lexicons*;* Polarity

1. **Introduction**

Sentiment analysis is a kind of text classification that catalogs texts based on the sentiment orientation of opinions they contain. It thus plays an important part of Natural Language Processing. NLP is a field of computer science and artificial intelligence that mainly deals with human-computer language interaction. This field is particularly of use to merchants, stock traders, and in election works.

Sentiment analysis is the process of detecting the contextual polarity of the text. It determines whether given text is positive, negative or neutral. It is otherwise called as opinion mining too, since it derives the opinion or attitude of the speaker. For this analysis, the opinions are collected from the users, which can be employed for further

* Corresponding author. Tel.: 0091-9846947813
E-mail address: md.devika@gmail.com

Peer-review under responsibility of the Organizing Committee of ICRTCSE 2016
doi:10.1016/j.procs.2016.05.124

improvements. The social networks act as a medium where the users can post many opinions a day and these blogs are used for classification. A lot of research work is being held in the field of sentiment analysis due to its significance in the marketing level competition and the changing needs of the people. Sentiment analysis requires the usage of a training set for its performance, and its quality plays a great role in the accurate evaluation of the text. The semantic analysis of the sentence also increases the meaning and accuracy of the result. POS tagging will be helpful to users for understanding whether the review or comment corresponds to the relevant subject searched for.

2. Levels of Analysis

In general, sentiment analysis has been investigated mainly at three levels [1]. In document level the main task is to classify whether a whole opinion document expresses a positive or negative sentiment. This level of analysis assumes that each document expresses opinions on a single entity. In sentence level the main task is to check whether each sentence expressed a positive, negative, or neutral opinion. This level of analysis is closely related to subjectivity classification, which distinguishes objective sentences that express factual information from subjective sentences that express subjective views and opinion. Document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself.

3. Sentiment Analysis Methods

Sentiment analysis played a great role in the area of researches done by many, there are many methods to carry out sentiment analysis. Still many researches are going on to find out better alternatives due to its importance in this scenario. Some of the methods are discussed in this paper.

3.1. Machine learning Approach

Machine learning strategies work by training an algorithm with a training data set before applying it to the actual data set. Machine learning techniques first trains the algorithm with some particular inputs with known outputs so that later it can work with new unknown data [2]. Some of the most renowned works based on machine learning are as follows:

3.1.1 Support Vector Machine

It is a non-probabilistic classifier in which a large amount of training set is required. It is done by classifying points using a (d-1)-dimensional hyper plane. SVM finds a hyper plane with largest possible margin [3]. Support Vector Machines make use of the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class membership. An illustration is given in Fig. 1a. In this the objects belong to either class red or green, and the separating line defines the boundary. Here the original objects are (left side of Fig. 1b) mapped or rearranged using a mathematical function known as kernel and this is known as mapping or transformation. After transformation, the mapped objects are linearly separable and as a result the complex structures having curves to separate the objects can be avoided.

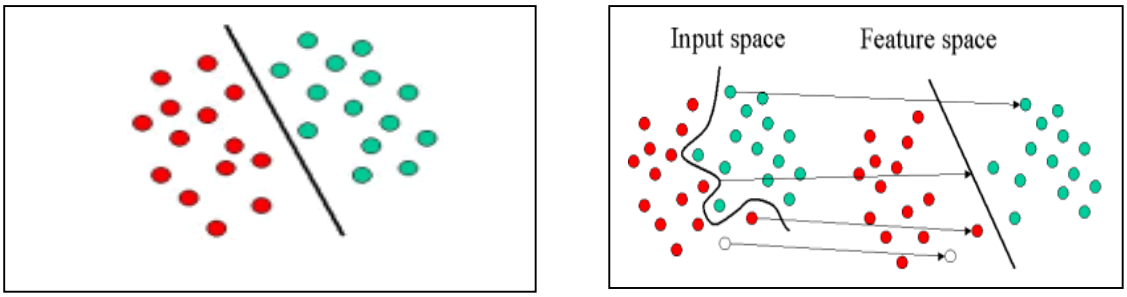

Fig 1. (a). Linear classifier (b). SVM illustration.

3.1.2 N-gram Sentiment Analysis

In the fields of linguistics and probability, an n-gram is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles. In this they are considering the sentence as a whole [5]. They are making use of four types of lexicons namely sentiment phrase lexicon, sentiment strength lexicon, lexicon with aspects and exception lexicons.

3.1.3 Naïve Bayes Method

It is a probabilistic classifier and is mainly used when the size of the training set is less. In machine learning it is in family of sample probabilistic classifier based on Bayes theorem. The conditional probability that an event X occurs given the evidence Y is determined by Bayes rule by the (1).

$$P(X/Y) = P(X)\,P(Y/X) / P(Y) \quad (1)$$

So for finding the sentiment the equation is transformed into the below (2) [6].

$$P(Sentiment/Sentence) = P(Sentiment)P(Sentence/Sentiment)/P(Sentence) \quad (2)$$

P(sentence/sentiment) is calculated as the product of P (token /sentiment) [6], which is formulated by the (3).

$$Count(Thistokeninclass)+1/Count(Alltokensinclass)+Count(Alltokens) \quad (3)$$

Here 1 and count of all tokens is called add one or Laplace smoothing.

3.1.4 Maximum Entropy Classifier

A Maximum Entropy (ME) classifier, or conditional exponential classifier, 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 encoding. The encoding maps each pair of feature set and label to a vector. ME classifiers belong to the set of classifiers known as the exponential or log-linear classifiers, because they work by extracting some set of features from the input, combining them linearly and then using this sum as exponent. If this method is done in an unsupervised manner, then Point wise Mutual Information (PMI) is made use in order to find the co-occurrence of a word with positive and negative words. The ME Classifier is one of the models which do not assume the independent features [7]. The uncertainty is

maximum for a uniform distribution. The measure of uncertainty is known as entropy. So model in this paper should be uniform as possible, still obeying the constraints that are imposed.

3.1.5 K-NN and Weighted K-NN

K-Nearest Neighbour method is based on the fact that the classification of an instance will be somewhat similar to those nearby it in the vector space. Further some group researched on weighted k-Nearest Neighbour method, in which they provided weightage to those elements in the training set and they used these weights for their calculation of sentiment of text in word by word manner [8]. Here the score is calculated by using the (4).

$$\text{Positivity Score} = ({}_1\textstyle\sum^{j} \text{score (pos)} + {}_1\textstyle\sum^{k} \text{score (neg)})/\ {}_1\textstyle\sum^{s} \text{maximum score} \quad (4)$$

Here s=j+k, ie. Count of both positive and negative together. In weighted k-NN method they first of all tokenise the sentences and removed the stop words from the tweets they have fetched. The algorithm proposed by the authors of [8] is carried out in two parses. A positive score is assigned to each reviews after the first parse. This is passed for second parsing and an input of neutral review is given. Using this the score is modified if required. It is done for better positivist determination and an output file consisting of review ID and its positive score is determined.

3.1.6 Multilingual Sentiment Analysis

Now a days customers are having options to express their views in various language of choice, to yield better result researcher should consider the posts in different language. It is elaborated in [9], which explained a method, within multilingual framework to carry out the task of determining the polarity of the text. It is done using several Natural Language Tool Kits. In this language is identified first using language models. After identification, the language is translated to English using standard translation software. In [9] they are making use of PROMT eXcellent Translation (XT) Technology, for the purpose of translation. After that they are going on to the process of sentiment classification [10].

3.1.7 Feature Driven Sentiment Analysis

The product feature extraction plays a key role in the evaluation of the products, since we can see the importance of the knowledge of the features and their relationships for the enhanced marketing plan. In [11], it is done by Fuzzy Domain Ontology Sentiment Tree (FDOST). In FDOST, the root node represents the product, the leaf nodes represent the polarity and the non-leaf nodes represent the sub features of corresponding parent features.

3.2. Rule Based Approach

Rule based approach is used by defining various rules for getting the opinion, created by tokenizing each sentence in every document and then testing each token, or word, for its presence. If the word is there and has with a positive sentiment, a +1 rating was applied to it. Each post starts with a neutral score of zero, and was considered positive. If the final polarity score was greater than zero, or negative if the overall score was less than zero [12] After the output of rule based approach it will check or ask whether the output is correct or not. If the input sentence contains any word which is not present in the database which may help in the analysis of movie review, then such words are to be added to the database. This is supervised learning in which the system is trained to learn if any new input is given.

3.3. Lexical Based Approach

Lexicon Based techniques work on an assumption that the collective polarity of a sentence or documents is the sum of polarities of the individual phrases or words. In the seminar ROMIP 2012 the lexicon based method proposed in [14] was used. This method is based on emotional research for sentiment analysis dictionaries for each domains. Next, each domain dictionary was replenished with appraisal words of appropriate training collection that have the highest weight, calculated by the method of RF (Relevance Frequency) [15]. The word-modifier changes

(increases or decreases) the weight of the following appraisal word by a certain percentage. Word-negation shifts the weight of the following appraisal word by a certain offset: for positive words to decrease, for negative to increase. The procedure of the text sentiment classification was carried out as follows. First weights of all training texts the classified text is calculated. All the texts are placed into a one dimensional emotional space. The proportion of deletions was determined by the cross-validation method. Then the average weights of training texts for each sentiment class were found. The classified text was referred to the class which was located closer in the one-dimensional emotional space.

4. Comparison And Consolidation

The comparison and consolidation of the three main approaches used in sentiment analysis is shown in Table 1. Performing sentiment analysis by various approaches will produce different results. Each approach has its own pros and cons. By considering the key factors like performance, efficiency, and accuracy, the machine learning approach yields the best result and most of the work has been done in this approach. Several methods are evolved for doing this task which are described in Table 2.

Table 1. Comparison of Three Approaches

Approaches	Classification	Advantages	Disadvantages
Machine Learning Approach	• Supervised and Unsupervised learning.	• Dictionary is not necessary. • Demonstrate the high accuracy of classification.	• Classifier trained on the texts in one domain in most cases does not work with other domains.
Rule Based Approach	• Supervised and Unsupervised learning.	• Performance accuracy of 91% at the review level and 86% at the sentence level. • Sentence level sentiment classification performs better than the word level.	• Efficiency and accuracy depend the defining rules.
Lexicon Based Approach	• Unsupervised learning.	• Labelled data and the procedure of learning is not required.	• Requires powerful linguistic resources which is not always available.

Table 1. Comparison of Different Machine Learning Methods

Methods	Advantages	Disadvantages
SVM Method	• High-dimensional input space. • Few irrelevant features. • Document vectors are sparse.	• A large amount of training set is required. • Data collection is tedious.
N gram SA	• Usage of 1- and 2-grams as features for sentiment prediction can increase the accuracy of the model in comparison with only single word feature.	• Long range dependencies are not captured. • Dependent on having a corpus of data to train from.
NB Method	• Simple and intuitive method. • It combines efficiency with reasonable accuracy.	• Mainly used when the size of the training set is less. • It assumes conditional independence among the linguistic features.
ME Classifier	• This method do not assume the	• Simplicity is hard.

	independent features like NB method.	
	• Can handle large amount of data.	
KNN Method	• Based on the fact that the classification of an instance will be somewhat similar to those nearby it in the vector space. • It is considered computationally efficient.	• Large storage required. • Computationally intensive recall.
Multilingual SA	• The texts of different languages are evaluated without translation. • Deals with 15 different languages.	• Training corpus for different language is needed.
Feature Driven SA	• Adaptable to large projects. • It is a concise process.	• Not a powerful on smaller projects.

5. Conclusion

Various sentiment analysis methods and its different levels of analysing sentiments have been studied in this paper. Our ultimate aim is to come up with Sentiment Analysis which will efficiently categorize various reviews. Machine learning methods like SVM, NB, Maximum Entropy methods were discussed here in brief, along with some other interesting methods that can improve the analysis process in one or the other way. Semantic analysis of the text is of great consideration. Research work is carried out for better analysis methods in this area, including the semantics by considering n-gram evaluation instead of word by word analysis. We have also come across some other methods like rule based and lexicon based methods. In the world of Internet majority of people depend on social networking sites to get their valued information, analysing the reviews from these blogs will yield a better understanding and help in their decision-making.

References

1. Neha S. Joshi, Suhasini A. Itkat, " A Survey on Feature Level Sentiment Analysis" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5422-5425.
2. He Y., "Incorporating sentiment prior knowledge for weakly supervised sentiment analysis ", ACM Transactions on Asian Language Information Processing, Vol. 11(2).
3. N. Veeranjaneyulu, Akkineni Raghunath, B, Jyostna Devi, Venkata Naresh Mandhala, "Scene Classification Using Support Vector Machines With Lda " journal of theoretical and applied information technology 31 may 2014. Vol. 63 No.3
4. Ankush Sharma, Aakanksha, "A Comparative Study of Sentiments Analysis Using Rule Based and Support Vector Machine", IJRCCE Vol. 3, Issue 3, March 2014.
5. P. Saloun, M. Hruzk and I. Zelinka, "Sentiment Analysis e-Bussines an e- Learning Common Issue," ICETA 2013 ,11th IEEE International Conference on Emerging eLearning Technologies and Applications, Stary Smokovec, The High Tatras, Slovakia, October 24-25, 2013.
6. A. Tamilselvi, M. ParveenTaj, "Sentiment Analysis of Micro blogs using Opinion Mining Classification Algorithm " International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Volume 2 Issue 10, October 2013.
7. Pablo Gamallo, Marcos Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets" Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 171–175, Dublin, Ireland, August 23-24 2014.
8. Ankitha Srivastava, Dr.M.P. Singh, "Supervised SA of product reviews usin Weighted k-NN Algorithm," 2014 11th International Conference on Information Technology.
9. Kerstin Denecke, "Using SentiWordNet for Multilingual Sentiment Analysis," ICDE Workshop 2008, 978-1-4244-2162-6/08/ 2008 IEEE.
10. Brett W. Bader, W. Philip Kegelmeyer, and Peter A. Chew "Multilingual Sentiment Analysis Using Latent Semantic Indexing and Machine Learning," 2011 11th IEEE International Conference on Data Mining Workshops.
11. Lizhen Liu, Xinhui Nie, Hanshi Wang, "Toward a Fuzzy Domain Sentiment Ontology Tree for Sentiment Analysis," 5th Internationa Congress on Image and Signal Processing (CISP 2012).) 2012.
12. Swati A. Kawathekar1, Dr. Manali M. Kshirsagar, "Movie Review analysis using Rule-Based &Support Vector Machines methods", IOSR Journal of Engineering Mar. 2012, Vol. 2(3), March. 2012, pp: 389-391.
13. Blinov P. D. ,Klekovkina M. V. , Kotelnikov E. V. , Pestov O. A." Research of lexical approach and machine learning methods for sentiment analysis", Vyatka State Humanities University, Kirov, Russia.
14. Klekovkina M. V., Kotelnikov E. V., "The automatic sentiment text classification method based on emotional vocabulary" , Digital libraries: advanced methods and technologies, digital collections (RCDL-2012) , pp. 118–123.
15. Lan M., Tan C. L., Su J., Lu Y. (2009), "Supervised and traditional term weighting methods for automatic text categorization", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31(4), pp. 721–735.