International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume V, Issue X, October - 2018

A Survey on Sentiment Analysis

Shylaja. P
Research Scholar, Department of Information Technology, Kannur University
shylup2002@gmail.com

N. K. Narayanan
Principal
College of Engineering, Vadakara
nknarayanan@gmail.com

Abstract

A thought, view / opinion or attitude based on emotion instead of reason is called sentiment. Human opinion is always a part of decision making irrespective of its impact in the short term or long term. Opinion Mining is used for automatic extraction of knowledge from the opinion about a particular topic or problem and is used to classify sentiments into positive, negative or neutral. The purpose of sentiment analysis is to determine the attitude or inclination of a communicator through the contextual polarity of their speaking or writing. Their attitude may be reflected in their own judgment, emotional state, or the state of any emotional communication they are using to affect a reader or listener. Simply, it is trying to determine a person's state of mind on the subject they are communicating about. This information can be mined from internet data: texts, tweets, blogs, social media, news articles, or comments. The basic task of sentiment analysis is emotion recognition and polarity detection. Most promising development in sentiment analysis is the application of deep learning. The aspect level sentiment analysis is performed in two steps: viz aspect term extraction and sentiment classification. This provides an overview of the existing sentiment analysis methods.

1. Introduction

Human opinion is always a part of decision making process; even if no critical thinking is involved as some people are tend to give their opinion based on merely what they believe in. Opinion mining is used for automatic extraction of knowledge from the opinion about a particular topic or problem. A thought, View or Attitude based on emotion instead of reason is called Sentiment. Sentiment Analysis is used to extract individual person's sentiment. Sentiment Analysis classifies sentiments into mainly three groups: Positive, Negative or Neutral. Sentiment Analysis is used to determine the attitude or inclination of a particular individual through the contextual polarity of his / her speaking or writing. Their attitude may be reflected in their own judgment, emotional state, or the state of any emotional communication they are using to affect a reader or listener. In fact, it is trying to determine a person's state of mind on the subject they are communicating about. This information can be mined from
texts, tweets, blogs, social media, news articles, or any such data: written, electronic or print.

These days, Social media is the best tool to know about the opinion, advice, comment of people and their perception about any product, services, event, thought, policy, politics, government etc.. Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, and emotions toward entities, events and their attributes.

1.1. Levels of Classification

Sentiment classification can be done at Document level, Sentence level and Aspect or Feature level.

1.1.1 Document Level

By using Document Level classification we can clearly point out whether a whole opinion document expresses a positive or negative sentiment [1]. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about that product. Commonly, Sentiment Analysis and Opinion Mining are also known as document-level sentiment classification. This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single topic or product). Therefore, it is not applicable to documents which evaluate or compare multiple entities.

1.1.2 Sentence Level

In this type of classification, the whole document is classified as positive or negative. As the name suggests, the task at this level goes to the sentences and determines

whether each sentence expresses a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to subjectivity classification [2], which distinguishes sentences (called objective sentences) that express factual information from sentences (called subjective sentences) that express subjective views and opinions. Yet, it should be noted that subjectivity is not equivalent to sentiment as many objective sentences can imply different opinions, e.g., "The new house we bought has an elevated open terrace which is too big." Researchers have also analyzed clauses [3], but the clause level is still not enough, e.g., "Inflation is too high even though GDP growth is better than last year."

1.1.3 Entity/Aspect Level

Both the document level and the sentence level analyses were not able to pin point what exactly people liked and did not like. Aspect level performs more accurate level analysis. Initially, it was termed as feature level analysis [4]. Aspect level tries to identify the opinion of the customer which consists of an idea. Identifying the relevance of opinion targets helps to understand the sentiment analysis in a better manner.

For example, "even though the price is not that low, I still like this Jacket" clearly has a positive tone, even though one cannot say that this sentence is entirely positive. In fact, the sentence is positive about the jacket (emphasized), but negative about its price (not emphasized). In many applications, opinion targets are described by entities and/or their different aspects. Thus, the aim of this level of analysis is to discover sentiments on entities and/or their aspects. For example, the sentence "The Mercedes Benz new Model car is futuristic in design, but its price is too high" evaluates two aspects, style and cost, of the car (entity). The sentiment on Benz's design is positive, but the sentiment on its price is negative. The design and price of the car are the opinion targets. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses.

Pimpalkar et al.[5] developed the system which shows the comments and feedbacks/reviews for products. They used Sentiwordnet and smiley's dictionary for determining the scores of words.

Lertsuksakda et al. [6] developed a model called 'Hourglass of Emotion' using Plutchik's wheel of emotions.

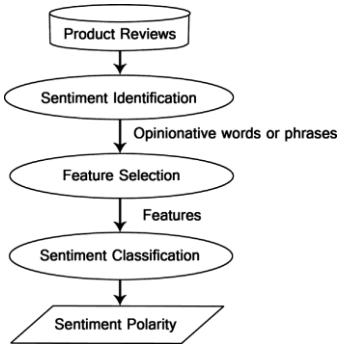
Finding sentiment can be formally defined as finding the quadruple (s,g,h,t) [7], where s represents the sentiment, g represents the target object for which the sentiment is expressed, h represents the holder (i.e., the one expressing the sentiment),and t represents the time at which the sentiment was expressed. Note that most approaches focus only on finding the pair (s, g).The target can be an entity, such as the overall topic of the review, or an aspect of an entity, which can be any characteristic or property of that entity. This decision is made based on the application domain at hand. For example, in product reviews, the product itself is usually the entity, while all things related to that product (e.g., price, quality, etc.) are aspects of that product. Aspect-level sentiment analysis is concerned, not just with finding the overall sentiment associated with an entity, but also with finding the sentiment for the aspects of that entity that are discussed. Some approaches use a fixed, predetermined list of aspects, while others freely discover aspects from the text.

There are basically three processing steps to be followed while performing aspect-level sentiment analysis: identification, classification, and aggregation [8]. While in practice, not every method implements all three steps or in this exact order, they represent major issues for aspect-level sentiment analysis. The first step is concerned with the identification of sentiment-target pairs in the text. The next step is the classification of the sentiment-target pairs. The expressed sentiment is classified according to a predefined set of sentiment values, for instance positive and negative. Sometimes the target is classified according to a predefined set of aspects as well. At the end, the sentiment values are aggregated for each aspect to provide a concise overview. The actual presentation depends on the specific needs and requirements of the application.

Besides these core elements of aspect-level sentiment analysis, there are additional concerns: robustness, flexibility, and speed. Robustness is needed in order to cope with the in formal writing style found in most user-generated content. People often make lots of errors in spelling and grammar, not to mention the slang language, emotions, and other constructions that are used to voice a certain sentiment. Flexibility is the ability to deal with multiple domains. An application may be performing very well on a certain domain, but very poorly on another, or just mediocre on all domains. Last, an aspect-level sentiment analysis solution ideally is

accessible using a Web interface due to Web ubiquity, underlining the need for high speed performance.

1.2. Sentiment analysis process on product reviews



Preprocessing techniques must be applied to the intended data set for removing unnecessary and irrelevant words such as articles, stop-words etc. thereby facilitating feature extraction. Several techniques are used for pre-processing such as Parts of Speech (POS) tagging, stemming and stop word removal.

Parts of Speech (POS) tagging is a linguistic technique used for product feature extraction from a data set as product aspects are generally nouns or noun phrases. POS tagging gives a tag to each word and classifies it to a specific and distinctive morphological category such as noun, verb, adjective etc. The most popular among this is Hidden Markova Models. This is widely used for developing POS taggers due to accuracy compared to rule based, statistical and machine learning.

Stemming and Lemmatization are two essential morphological processes of pre-processing module during feature extraction. Stemming converts all the inflected words present in the text into a root form called a 'stem'. For example, 'dramatic', 'drama' and 'dramatization' are each converted into the stem 'drama'. Stemming removes word inflections. The lemma of a word includes its base form plus inflected forms. By applying Lemmatization various inflected forms of a word can be clearly grouped into a single one. For example, the words 'study', 'studied' and 'studying' have 'study' as their lemma. Lemmatization replaces words with their base form. Lemmatization is more accurate and needs additional dictionary support for searching and indexing. Word Net Lemmatizer and Word Net Database can be used to lookup for lemmas.

Stop Word Removal is the process by which removal of common and high frequency words like 'a', 'the', 'of', 'an' etc are filtered out. It reduces the dimensionality of the datasets. Stop word list is available in Savoy [9]

2 Types of Features

2.1 Morphological Types

Based on its structure, there are three types of morphological features: semantic, syntactic and lexicon. Semantic type of feature works on contextual information and syntactic feature use POS tagging, chunk labels and lexicon structural feature consists of special symbol frequencies and word level lexical features.

2.2 Frequent Features

Frequent or hot features are the features which have more interest. Frequent Pattern Mining is used in text mining [10].

2.3 Implicit Features

Implicit features are the features which are not very clear in its review.

3 Sentiment Analysis Methods

Sentiment analysis methods can be mainly categorized into lexicon based methods, machine learning based methods and hybrid methods.

3.1 Lexicon Based Techniques

This is an unsupervised learning technique and therefore it does not require prior training. Normally Sentiment Lexicon contains list of words and expressions which are used to express people's subjective feelings and opinions. In this two main methods are corpus based methods which includes statistical methods and semantic methods, and the dictionary based methods. Statistical method uses the probability of each word used in the text. Hyeoncheol et al.[11] built the sentiment lexicon, and computed the probability score of each word using a term frequency. Using the probability score and the threshold, sample tweets were categorized. Dictionary-based method (also known as lexicon-based approach) selects a small set of "opinion words" as seed words, and then expands a word set from the seed words using an online dictionary. These expanded words and the seed words are used for sentiment analysis as

features. In [12], the author used mobile phone reviews and using the dictionary Wordnet, the polarity is calculated on the basis of majority of opinion words. In [13], Richa Sharma et al. developed Aspect based Sentiment System which extracted the features and opinions from sentences and found out whether it is positive, negative or neutral for each feature. They adopted a dictionary based technique of the unsupervised approach. Wordnet is used as a dictionary for determining the opinion words and their synonyms and antonyms. All the features of the product were clearly identified and the orientation is determined. Polarity of the sentence is determined on the basis of the majority of the opinion words. The system generated the feature wise summary of positive, negative and neutral sentences may have helped customers to take a decision whether the product is to be purchased or not. The Basic steps includes pre-process the text, initialize the total sentiment score, tokenize text and find out if the token is present in the sentiment dictionary and get the total sentiment score.

3.2 Machine Learning Based Techniques

Sentiment analysis in Twitter is a different paradigm compared to other researches that attempt sentiment analysis through machine learning. This is due to constraints which are present in identifying the sentiments expressed in tweets. Most commonly used machine learning techniques are Naïve Bayes Classification, Random Forest, Support Vector Machines, SMO Sequential mining optimization algorithm, J48, Maximum Entropy Rule, Decision Induction Tree, Neural Network, Probability Latent Semantic and Latent Dirichlet Allocation.

3.2.1 Naive Bayes Classification

The Naive Bayes classifier is a probabilistic model based on the Bayes' theorem, which calculates the probability of a tweet belonging to a specific class such as neutral, positive or negative. In this method, it is assumed that all the features are conditionally independent [14]. Even though Naive Bayes classifier has yielded better results in [15], it is failed to show superior results compared to some other classifiers.

3.2.2 Random Forest

Random Forest method involves many classification trees known as tree classifiers, which are used to predict the class based on the categorical dependent variable [16]. Each tree gives a class for the input vector and the class with highest turns will be chosen. This classifier's error rate depends on the correlation between any two trees in the forest and the strength of each individual tree in the forest. In order to minimize the error rate the trees should be strong and independent of each other [17].

3.2.3 Support Vector Machines (SVM)

Support vector machines have proved to be highly effective at traditional text categorization, especially compared to Naive Bayes. They are large-margin, rather than probabilistic, classifiers, in contrast to Naive Bayes and Maximum Entropy that are based on kernel substitution [18]. They can be defined as systems which use hypothesis space of linear functions in a high dimensional feature space. This will be trained with a learning algorithm that implements a learning bias derived from statistical learning theory. In the two-category case, the basic idea behind the training procedure is to find a hyper plane, represented by a vector, that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible.

3.2.4 SMO Sequential Mining Optimization Algorithm

It efficiently solves the optimization problem when training support vector machines. SMO takes an iterative approach to solve the optimization problem where it breaks it into a series of smallest possible sub-problems and solve them analytically. Each small problem involves two Lagrange multipliers because of the linear equality constraint.

3.2.5 J48 (pruned or un-pruned C4.5 decision tree)

C4.5 is an algorithm developed by Ross Quinlan to generate decision trees from a set of training data. J48 is an open source java implementation of C4.5 algorithm which is used in WEKA data mining tool. The training data set is already classified where each sample is a vector which represents the attributes of the samples. C4.5 splits the sample at each node by choosing a suitable attribute of the data based on information gain [19].

3.2.6 Maximum Entropy

It is an alternative technique which has proven effective in a number of natural language processing applications. It sometimes, but not always, outperforms Naive Bayes at standard text classification. The Maximum entropy Classifier converts labeled feature sets to vectors using encoding. The encoded vector is used to calculate weights for each feature. In [20], a novel method is used to collect various twitter messages. After pre-processing, the dataset is used to build user's emotional state classification and the SVM, ME and Naïve Bayes Classifiers are applied and the result was efficient.

3.2.7 Decision Induction Tree

Decision Tree learning is one of the most widely used and practical methods for inductive inference over supervised data. Decision tree represents a procedure for classifying categorical data based on their attributes. It is also efficient for processing large amount of data, therefore often used in data mining applications.

3.2.8 Neural Network

Recently Convolutional Neural Networks (CNNs) models have proven remarkable results for text classification and sentiment analysis [21]. They classified business reviews using word embedding's on a large-scale dataset provided by Yelp: Yelp 2017 challenge dataset. They compared word-based CNN using several pre-trained word embedding's and end-to-end vector representations for text reviews classification and used deep learning techniques.

3.2.9 Probabilistic latent semantic

Probabilistic latent semantic analysis (PLSA), also known as probabilistic latent semantic indexing is a statistical technique for the analysis of two-mode and co-occurrence data, which has applications in information retrieval and filtering, natural language processing, machine learning from text, and in related areas.

3.2.10 Latent Dirichlet Allocation

It is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

3.3 Hybrid Techniques

Hybrid Techniques indicates the combination of both the Machine Learning and the Lexicon Based Approaches to improve the sentiment classification performance. Mudinas et al. [22] presented concept-level sentiment analysis system by combining Lexicon Based and Learning Based approaches. The sentiment lexicon consists of 7048 sentiment words with vales from -3 to +3 and the method achieved 82.3% accuracy. Fang et al. [23] incorporate the Domain Specific Sentiment Lexicons into SVM learning and got significant improvement in the accuracy. They achieved 66.8% polarity accuracy. Zhang et al. [24] combined Machine Language Technique and Lexicon Based Technique on Twitter data and achieved an accuracy of 85.4%. They proposed a hybrid model to reduce the classification work and label the reviews. They used SVM baseline classifier in the form of feature vectors and it achieved 91% accuracy. Rudy Prabowo et al(2009) combines rule-based classification, supervised learning and machine learning into a hybrid method. This method is tested on movie reviews, product reviews and MySpace comments. The results show that a hybrid classification can improve the classification effectiveness than any individual classifier. Albornoz [25] proposed a hybrid approach to determine the polarity using the combination of NLP and ML methods. They transformed the given sentence into a vector of emotional occurrences (VEO) and that is given as input to the ML algorithms like j48, SVM etc. Pang and Lee et al. [26] proposed a machine-learning method that applies text-categorization techniques to extract the subjective portions of the document and found out the minimum cut in graphs which facilitates incorporation of cross-sentence contextual constraints. Pedro P. BalageFilho et al. [27] proposed a hybrid classification system for Semeval-2014: Sentiment Analysis in Twitter. They have noticed that the improvement of the lexicon and machine learning modules, the overall score tends towards the machine learning score.

4 Conclusion

This survey paper presented an overview on the Sentiment Analysis work. It has been noticed that the hybrid classification approach on Sentiment Analysis gives better result compared to all other individual classification approach.

References

[1] Bo Pang and Lillian Lee, ShivakumarVaithyanathan(2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

[2] StevenBethard, Hong Yu, Ashley Thornton, VasileiosHatzivassiloglou, and Dan Jurafsky(2017). Extracting Opinion Propositions and Opinion Holders using Syntactic and Lexical Cues. Book Computing Attitude and Affect in Text: Theory and Applications (pp.125-141) DOI 10.1007/1-4020-4102-0_11, ISSN 1387-5264.

[3] AnyceWiebe1, Theresa Wilson2 and Claire Cardie3, Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation (2005) 39: 165–210 Springer 2006 DOI 10.1007/s10579-005-7880-9.

[4] Konstantin Bauman, Bing Liu, and Alexander Tuzhlin. Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. Proceedings of SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2017).

[5] Pimpalkar, A., Wandhe, T., Kene, M., &Rao, M. S. (2014). Review of online product using rule based and fuzzy logic with Smiley's. Internationl Journal of Computing and Technology. I(1), 39-44.

[6] Lertsuksakda, R. Pasupa, K., &Netisopakul, P. (2014). Thai sentiment terms construction Using the Hourglass of emotions. International Proceedings of 6th International Conference on Knowledge and Smart Technology (KST, pp. 46-50)

[7] B. Liu, Sentiment Analysis and Opinion Mining, ser. Synthesis Lectures on Human LanguageTechnologies.Morgan&Claypool,2012,vol.16.

[8] M.Tsytsarauand T.Palpanas,"Survey on Mining Subjective Data on the web," Data Mining and Knowledge Discovery, vol.24,no.3,pp. 478–514,2012.

[9] Savoy, J. (2005). IR Multilingual Resources at Uni NE Available at http://members.unine.ch/jacques.savoy/clef/index.html.

[10] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. Proceedings of 20th International Conference Very Large Databases, VLDB, 1215:487-499, 1994.

[11] Lee, Hyeoncheol, Youngsub Han, and K. Kim. "Sentiment Analysis on Online Social Network Using Probability Model." Proceedings of the The Sixth International Conference on Advances in Future Internet. 2014.

[12] Richa Sharma, Shweta Nigam, Rekha Jain, "Polarity Detection at Sentence Level", International Journal of Computer Applications, Volume 86- No 11, 2014.

[13] Richa Sharma, Sheta Nigam and Rekha Jain, " Mining Of Product Reviews At Aspect Level", International Journal in Foundations of Computer Science &Technology(IJFCST), Vol.4, No. 3, May 2014.

[14] K.P.Murphy. (2006). Naive Bayes classifiers.[Online].Available:http://www.cs.ubc.cal-murphyk/TeachinglCS340- F all06/readingiNB. pdf.

[15] A. Pak and P.Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", in Proc. i h conference on International Language Resources and Evaluation LREC'lO ,May 2010.

[16] L.Breiman."Random Forests.", Machine Learning, vol. 45, pp.5-32, Jan. 2001.

[17] L.Breiman and A. Cutler. "RandomForests." Internet: www.stat.berkeley.edu/-breimanlRandomForests/cc_home.htm. [Apr. 1, 2012].

[18] K.P.Bennet and C.Campbell. "Support Vector Machines: Hype or Hallelujah?" in Proc.SIGKDD Explorations, 2000, vol. 2, no. 2, pp 1-13.

[19] J.Ross Quinlan. Programs for machines learning. Edward Brothers, 1993, pp.17-32

[20] JasakaranKaur, ShevetaVashisht, "Analysis And Identifying Variation In Human Emotion Through Data Mining", Int.J.Computer Technology & Applications, 2012.

[21] AndreenaSalinca, "Convolutional Neural Networks for Sentiment Classification on Business Reviews",Proceedings of IJCAI Workshop on Semantic Machine Learning (SML 2017): 35-39

[22] A. Mudinas, D. Zhang, M. Levene, " Combining Lexicon and Learning Based Approaches for concept level 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.

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

[24] 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.

[25] Albornoz, Laura Plaza, Pablo Gervás, "hybrid approach to emotional sentence polarity and intensity classification", Proceedings of the Fourteenth Conference on Computational Natural Language Learning Pages 153-161,Upsalla, Sweden, 2010.

[26] Bo Pang, Lillian Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics Article No. 271, Barcelona, Spain, 2004.

[27] Pedro P. BalageFilho, Lucas Avanc¸o, Thiago A. S. Pardo, Maria G. V. Nunes, "NILC USP: An Improved Hybrid System for Sentiment Analysis in Twitter Messages", International Workshop on Semantic Evaluation, 8th, 2014, Dublin.