Date of Publication: June 30, 2017 | Volume 5, Issue 2 | Pages: 38- 42

Techniques of Opinion Mining: A Review

Gurkamalpreet Kaur, Research Scholar, Department of Computer Science, Lovely Professional University, Jalandhar, India, gurkamalpreetkaur4@gmail.com
Arjinder Singh, Assistant Professor, Department of Computer Science, Lovely Professional University, Jalandhar, India, arjinder.20858@lpu.co.in

ABSTRACT: Opinion or Sentiment mining allude to identify and analyze the opinion or viewpoint of the internet user expressed toward a particular topic. Opinion mining focuses on what people think and feels. Analyses of viewpoint provide useful information regarding people interest to the business analyst and other interested parties. There are various unsupervised and supervised learning techniques available for opinion mining for e.g. naïve bayes, support vector machine (SVM), wordnet etc. This paper focuses on the various opinion mining techniques available and related work done in opinion mining area.

Keywords: Opinion mining, unsupervised learning, supervised learning, classification.

1. INTRODUCTION

Opinion mining is the method of studying or detecting the viewpoints or perspective of the writer from any text and classifies the text into different categories. It has been utilized in various fields such as marketing, e-commerce, education, elections etc. In this modern era, Internet has become utmost important part in everyone's life. Every person is using internet for communicating or sharing the information over the web and it is free to express the opinion, feelings related to any topic. Everyday huge amount of data is being generating over the web. By analysing the text, interesting patterns of human behaviour can be determined. As a result, opinion mining has become a necessary task. Recently, Social media is enormously used in exhibiting personal opinion on any object that can be any product, service, news, issue or any celebrity. Social media provides a platform for expressing personal opinions where traditional data collecting techniques like surveys are more time consuming. Opinion can be positive or negative. In order to understand the satisfaction of a person regarding any object, opinions plays a major role. Accurate opinion mining can help in decision making.

There are three types of opinion mining [1]

Document level opinion mining: - It is one of the easiest types of opinion mining. In this type, it is assumed that whole document contains only single opinion. Thus whole document is used to analyze the subjectivity whether it is of positive, negative or neutral polarity. It provides brief information in terms of total number of positive or negative documents. In document level, opinion words are extracted from the whole document and analyze the polarity of the document.

Sentence level opinion mining: - A document may contain multiple opinionated sentences thus sentence level opinion mining is used to classify the complete sentence into positive, negative or neutral polarity. In this type, it is assumed that each and every sentence contains a single opinion. As single document can have several opinions regarding the same entity thus sentence level opinion mining is used to get more detailed information of the different views conveyed in a document.

Feature level opinion mining:-Now a days, feature based opinion mining is becoming more popular because people are more focusing on the specific features of a particular entity than the complete entity. Every entity has multiple features like speed, quality, price, size etc. and different people have different opinions about these features. In this type, sentence is further broken into nouns, pronouns, adjectives etc. and all the features of a sentence are analyzed and classification is performed based on relevant features.

2. PHASES OF OPINION MINING

2.1 Collection of data

In this phase, first step is to collect the data from the various social networking sites such as twitter, Facebook, blogs, any economic site or review sites etc.

2.2 Preprocessing of data

Data collected from the various social networking sites are highly unstructured in nature because of written in informal language. Thus it needs to clean the data before processing it. It includes:-

2.2.1 Stemming

Stemming is performed on each word to get a root word i.e. common morphological endings of the related words. For example walking, walked, walks are stemmed to walk.

2.2.2 Removal of stop words

Dataset collected contains huge amount of data and while processing the data it would give inaccurate results. Thus it needs to remove the stop words such as "is, from, are, across, am etc." that are not useful in opinion mining. For example "this girl is beautiful" will be processed and gives the output "girl beautiful".

2.2.3 Part of speech tagging

POS tagging is a technique of applying a part of speech tag to every words present in the input data such as noun, verb, and adjective. Thus with POS tagging, words of the input data are grouped into categories of noun, adjective, verb. For example "the quality of the product is good" will be tagged as "the (determinant) quality (noun) of (preposition) the (determinant) product (noun) is (preposition) good (adjective)".

2.3 Feature Extraction

The purpose of this phase is to extract the features from the sentence which can be used for opinion classification. Features are basically tagged as nouns or noun phrases which are present as subject or object in the sentence. There are various approaches used for feature extraction such as n-gram, feature ranking algorithm, frequency count, syntactic rules etc.

2.4 Negation handling

In this step, the polarity of the negative words are swapped. If the word is identified with negation then its polarity will be change from positive to negative.

2.5 Opinion classification

There are various machine learning classification techniques used for opinion Classification which classifies the words into positive or negative words. There are two main steps in machine learning techniques. First step is to construct a crisp model to distribute the class labels of training set (with known class labels) and second step is to use the resulting classifier to predict the class labels to the testing set (unknown class labels). Machine learning is also called supervised learning.

3. OPINION MINING TECHNIQUES

3.1 Unsupervised learning

It is also called lexicon based approach. It is a classic approach for opinion mining to classify the lexicons into positive, negative and neutral words. For example pretty has positive polarity and horrible has negative polarity. There are various lexicon based methods to classify the opinions.

3.1.1 WorldNet

It is a largest online lexical database which contains English words, nouns, pronouns, verbs, adverbs, adjectives along with their set of synonyms. WorldNet has more than 118,000 unique words with different word senses. Each category of word is organized in semantic network of words i.e. synonymy, antonymy, hyponymy, meronymy, troponomy, entailment. [2]

3.1.2 Sentiwordnet

It is also a lexicon resource which is related with three polarity scores positive, negative, objective and classify the data into PN-polarity or SO-polarity. It can used for all parts of speech i.e. adjectives, noun, pronoun, verb, adverb and associates the polarity to words according to the sense rather than terms. Polarity scores can be positive, negative. For example the word "healthy" can have the scores of polarity as:

Positive=0.5, Negative= 0.0, objective= 0.5
(sense1 for healthy economy)

Positive=0.75, Negative= 0.0, objective= 0.25
(sense2 for a good health)

3.1.3 English Subjectivity lexicons

Lexicons are list of words that represents the subjectivity of the text. Subjectivity lexicons consist of 8221 words and each word has allocated the polarities along with 4 levels i.e. positive, negative, neutral and both.

3.2 Supervised learning

There are various machine learning classification techniques used for opinion Classification[3]. There are two main steps in machine learning techniques. First step is to construct a crisp model to distribute the class labels of training set (with known class labels) and second step is to use the resulting classifier to predict the class labels to the testing set(unknown class labels). Supervised learning techniques are as follows:

3.2.1 Naive Bayes

Naive Bayes classifiers are based on Bayesian networks. Bayesian networks are graphical model to predict the probability of relationship among different variables and are comprise of directed acyclic graph with only one parent node and multiple child nodes. Assure that child nodes are independent of their parent nodes. Naive bayes is based on formula:

$$R = \frac{P(i\,|X)}{P(j\,|X)} = \frac{P(i)\,P(X|\,i)}{P(j)\,P(X|\,j)} = \frac{P(i)\,\pi\,P(Xr\,|\,i)}{P(j)\,\pi\,P(Xr\,|\,j)}$$

After the comparison of two probabilities, the highest probability predicts the class level value of tuple X belongs to i if and only if R>1. If R<1 then class label value of tuple X belongs to j. Merits of using the Naive bayes classifier is its less computational time for training. Its product form can also convert into sum by using logarithm.

Major drawback of naive bayes is the assumption that all the child nodes are independent because of this naive bayes classifier is less accurate than other machine learning algorithms. In order to solve this problem extra edges are added to incorporate some dependencies between the variables.

3.2.1 Support vector machine

Support vector machine is a machine learning algorithm used to classify both linear and nonlinear data. SVM includes the concept of margin on the either side of hyperplane that is used to separate two data classes. Hyperplane with larger margin are more accurate for classifying the data than hyperplane with smaller margin. Therefore SVM always look around the hyperplane with maximum margin this is called maximum marginal hyerplane (MMH).

In case of linearly separable data after finding the optimized separating hyerplane, data points that are on its margin called support vector points and consider the linear combination of these points for solution discards the remaining points. Let D be the data set given as (x1, y1), (x2, y2)…….. (xn, yn), where xi is the set of training tuples with associated class labels yi. Each class can have one of the two values either +1or-1.

Date of Publication: June 30, 2017 | Volume 5, Issue 2 | Pages: 38- 42

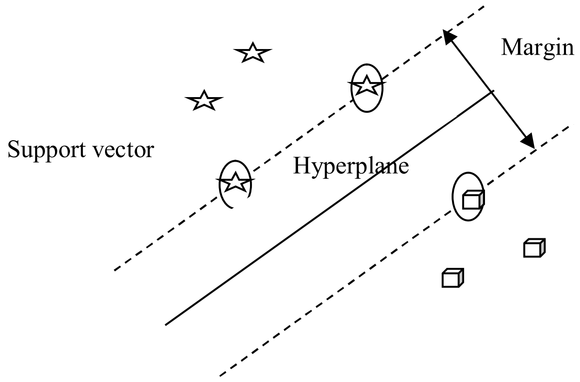

Figure 1: Linear separable hyperplane

In case of nonlinear separable data, there is no straight hyerplane exists to separate the classes. Benefit of using SVM is that, it can also find the nonlinear decision boundaries. SVM uses nonlinear mapping to convert the input data to higher dimensional and then define a separating hyperplane. Training tuples depends on the dot product, K (x_i, xj) = Φ (x_i).Φ (x_j) where Φ(x) for non-linear mapping of data to some other dimensions.

K is kernel function that allows dot product calculated directly in feature space. Once the maximal separating hyerplane gets created K will map new points to feature space for classification.

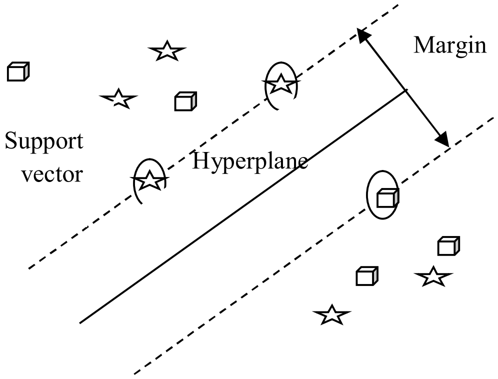

Figure 2: Non- Linear separable Hyperplane

4. LITERATUTE SURVEY

Minara P anto et al. [4] This paper proposed a technique which provides a automatic feedback of a product by collecting the data from the twitter. For this, first of all data from the twitter has been extracted using the twitter 4J API. After extracting the data, preprocessing is performed in which stop words like “is, am, but” are eliminated and the sentences are divided into smaller sentences and then POS tagging is performed on the text in which each word is assigned a parts of speech tag such as noun, verb, adjective etc then SVM classifier is used to classify into positive, negative or neutral words. After analyzing the efficiency, it has been analyzed that SVM classifier is more accurate among other classifiers i.e. Naïve bayes, Maximum Entropy etc. In order to classify the data more accurately dual prediction technique has been used which evaluates the sentence from both directions which gives the output of two distinct values and the mean among them has used to find the sentiments from it and provides feature based rating to a product and get the overall rating, unigram approach is used in which the frequency of the words are evaluated.

Tirath Prasad Sahu et al. [5] In this paper, movie reviews are classified on the scale of 0 to 4 i.e. highly disliked to highly liked. First of all, more than 50,000 reviews are collected from IMDB because of the collected data is highly unstructured, preprocessing is performed on it i.e. stemming (extracting the root word), stopping (removing the most commonly used stop words), part of speech tagging (words are tagged as noun, verb, adverb etc.). After preprocessing, features that affect the polarity of the document are analyzed by using the Sentiwordnet i.e. positive sentiment words, positive sentiment bigrams, negative sentiment words, negative sentiment bigrams etc. after analyzing the features, information gain and feature ranking algorithm is used to scale all the features and provide sentiment score. After providing the sentiment score, class labels for the document is determined i.e. strong negative, weak negative, neutral, weak positive, strong positive. After classification, this technique is compared with the other classifiers i.e. naïve bayes, decision tree, KNN etc and it is found that this technique is more accurate than other classifiers.

Mondher Bouazizi, *et al*.[6] Sarcasm is defined as opposite meaning of the sentence that any people speak. In this paper, tweets on any topic are classified into polarity of positive and negative. Sarcastic tweets are also detected to improve the accuracy of sentiment analysis. After collecting the tweets, textual and non -textual features are extracted from the tweets. Negative handling is performed on the text to convert the positive words into negative and vice versa. After extracting all the features, classification is performed on the test set by using naïve bayes, SVM and maximum entropy. It has been analyzed that some of the tweets are misclassified due to the presence of sarcasm. Sentiment related, punctuation related, syntactic, pattern related features are extracted for sarcasm detection. After comparing the results it has been analyzed that results are more accurate after taking sarcasm into consideration.

Mohamed Yassine, *et al*.[7] In this paper, emotion mining is performed on the text shared online in the form of comments and wall post of online social networking sites in order to identify whether the text is subjective or objective. Because most of the data shared online are written in informal language thus lexicons are developed for social acronyms, emotions and for foreign languages. After collecting the data from social networking sites , subjectivity features are extracted from the text based on correlation measure to avoid redundant attributes i.e. number of affective words, punctuation words, repeated words ,social acronyms, emotions etc. continuous

Date of Publication: June 30, 2017 | Volume 5, Issue 2 | Pages: 38- 42

attributes are mapped to discrete values by using K-Mean clustering with K=3 for subjective, objective and moderate subjective text . Min-max normalization is performed to map the attribute values between 0 and 1. In order to predict the relationship strength between two users SVM is used.

Eman M.G. Younis, *et al.*[8] In this research paper, text mining and sentiment mining is apply to analyze the twitter data about the products and services of two different UK stores (Tesco and Asda). This will help the business to check the views of customers about their products and services. Twitter messages are accessed using twitteR package. Next the data is cleaned from stop words, spaces and perform stemming. In this tm packages are used. It produces a structure representation of tweets. Next the data mining techniques are used such as association rules, finding most frequent words and sentiment mining (lexicon based approach). In lexicon based approach, there is a set of positive and negative words, which are combined with a scoring function to determine the polarity of sentiments. Finally wordcloud package and bar charts are used to show the frequency of words and sentiment score in the customer tweets.

Ms. Ashwini Rao et al.[9] In this paper, an algorithm has been proposed to select the relevant features in order to reduce the complexity of classification. An unsupervised and domain independent technique is used to extract the features from the data. As data collected from the web is highly unstructured thus preprocessing is required to remove the stop words, special words, tags. Boundary of the sentence is also determined. For feature generation, first of all Stanford parser is used to tag POS to identify the common features and then proposed algorithm is used to extract the more frequent features and then threshold value is used to filter out the rules which provides more relevant features and then some rules are defined to further refine the features. At the end it has been analyzed that there is reduction in the feature space.

Gautami Tripathi et al.[10] In this paper, both natural language processing and machine learning techniques are used to extract the sentiment from the movie reviews. First of all, preprocessing is performed on the data collected from the web because of its unstructured nature. Preprocessing includes various steps such as tokenization, pruning, filtering tokens and stemming. Next step is to extract the relevant features from the data. For this term occurrence, TF-IDF, term frequency and binary term occurrences are used and then n-gram method is used to extract the adjacent words i.e. unigram, bigram, trigram, four-gram and for classification both SVM and naïve bayes algorithms are used and their accuracy is measured and it has been analyzed that TF-IDF along with SVM classifier gives the maximum accuracy and term occurrence along with naïve bayes gives the maximum accuracy.

R.Nithya et al. [11] In this paper, combination of lexicon based and syntactical based approach has been proposed in order to detect the overall sentiment score, feature score and also the most relevant features of the product. In order to find the overall sentiment score, first of all data has been collected from the web and then preprocessing is performed to remove the tags, stop words and the text is divided into words and then POS tagger is used to tag every word. After the data cleaning, chunking of data is performed by the regular expression parser in order to extract and filter the adjective phrase accompanied by a noun phrase and then lexicon based dictionary is used to perform the sentiment classification. Thus in this way both syntactical rule and lexicon approach i.e. LPSA (Lexicon pattern sentiment analysis) is used to determine the overall sentiment score. In the second part to extract the most promising features from the text, term frequency method is used to avoid the irrelevant features. FBSC (Feature based sentiment score) algorithm is used to score the features of the product. In this way opinion classification is performed based on overall scores as well as feature scores.

HAN-XIAO SHI et al.[12] In this paper, a supervised machine learning and TF-IDF approach has been used to classify the hotel reviews from the web into positive or negative polarities. First of all, the reviews have been collected from the discussion forum and then TF-IDF technique is used to extract the important unigram features from the text. In TF-IDF, weight is assigned to every feature and evaluates how significant the words is in the corpus then classification is performed by using the SVM classifier which classifies the reviews into positive or negative reviews and then performance is evaluated by using the recall, precision and f-score. Results are compared with the frequency and it has been analyzed that TF-IDF is more accurate than frequency.
The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent sub- sections should be numbered and flush left. For a section head and a subsection head together (such as Section 4 and subsection 4.1), use no additional space above the subsection head.

5. CONCLUSION

Opinion mining is one of the popular research areas because many people make decisions on the basis of online reviews. Opinion mining has various applications in many areas as business, politics, ecommerce, elections etc. In this paper, we have illustrated the approach and techniques of opinion mining. There are various techniques for opinion classification we mainly focuses on machine learning techniques because they are more accurate.

REFERENCES

[1] R. Feldman, “Techniques and applications for sentiment analysis”, *Commun ACM*, pp. 82–89, 2013

[2] G. a. Miller, “WordNet: a lexical database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[3] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: A review of classification and combining techniques,” *Artif. Intell. Rev.*, vol. 26, no. 3, pp. 159–190, 2006.

[4] M. Antony, N. Johny, V. James, and A. Wilson, “PRODUCT RATING USING SENTIMENT ANALYSIS,” pp. 3458–3462, 2016.

[5] T. P. Sahu, “Sentiment Analysis of Movie Reviews : A study on Feature Selection & Classification Algorithms,” 2016.

Date of Publication: June 30, 2017 | Volume 5, Issue 2 | Pages: 38- 42

[6] M. Bouazizi and T. Ohtsuki, “Opinion Mining in Twitter How to Make Use of Sarcasm to Enhance Sentiment Analysis,” *Proc. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. 2015 - ASONAM ’15*, pp. 1594–1597, 2015.

[7] M. Yassine and H. Hajj, “A framework for emotion mining from text in online social networks,” *Proc. - IEEE Int. Conf. Data Mining, ICDM*, pp. 1136–1142, 2010.

[8] E. M. G. Younis, “Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools : An Empirical Study,” vol. 112, no. 5, pp. 44–48, 2015.

[9] M. A. Rao, “Model for Improving Relevant Feature Extraction for Opinion Summarization,” pp. 1–5, 2015.

[10] G. Tripathi, S. Naganna, G. Noida, and G. Noida, “FEATURE SELECTION AND CLASSIFICATION APPROACH FOR,” vol. 2, no. 2, 2015.

[11] “Correlation of Feature Score to Overall Sentiment Score for Identifying The Promising Features,” pp. 9–13, 2016.

[12] H. Shi and X. Li, “A SENTIMENT ANALYSIS MODEL FOR HOTEL REVIEWS BASED ON,” pp. 10–13, 2011.