Business intelligence analytics using sentiment analysis-a survey

Prakash P. Rokade, Aruna Kumari D.
Department of Computer Science and Engineering, KLEF Deemed University Vaddeswaram, India

Article Info

Article history:

Received Dec 22, 2017
Revised Jan 16, 2018
Accepted Sep 10, 2018

Keywords:

Feature extraction
Machine learning
Sentiment analysis

ABSTRACT

Sentiment analysis (SA) is the study and analysis of sentiments, appraisals and impressions by people about entities, person, happening, topics and services. SA uses text analysis techniques and natural language processing methods to locate and extract information from big data. As most of the people are networked themselves through social websites, they use to express their sentiments through these websites.These sentiments are proved fruitful to an individual, business, government for making decisions. The impressions posted on different available sources are being used by organization to know the market mood about the services they are providing. Analyzing huge moods expressed with different features, style have raised challenge for users. This paper focuses on understanding the fundamentals of sentiment analysis, the techniques used for sentiment extraction and analysis. These techniques are then compared for accuracy, advantages and limitations. Based on the accuracy for expexted approach, we may use the suitable technique.

Corresponding Author:

Prakash P. Rokade,
Department of Computer Science and Engineering,
KLEF Deemed University Vaddeswaram,
Guntur, A.P, India.
Email: prakashrokade2005@gmail.com

1. INTRODUCTION

Natural language is playing key role for direct or indirect communication. Any information can be partitioned in to facts and opinions. Sentiments are the subjective statements that reflect the sentiments of an individual about an event or an object. The technique to extract subjective information from text and determining the overall contextual polarity of opinion is called as sentiment analysis. As stated by Liu [1, 2], it is always useful for businesses to know impressions by their customers about their product or services; so as to know the reasons of their profits or losses. Also, the customer who wants to buy a product would like to know the impressions about that product by existing users [3]. Users may comment on overall product or feature of the product. To find the sentiment of an object, first the sentiment words for features of given object are identified. These features assign weights either on the basis of their polarity or on the basis of polarity and importance. The polarity weights of all the features of the object are aggregated in order to get the aggregate sentiment about object [4]. The techniques for measuring sentiments are broadly classified into machine learning methods and lexicon-based approaches.

Huge number of sentiments, their analysis and impact of parts of speech on sentiment polarity raised great challenges to the researches. A comparative study of different machine learning algorithms, linguistic features, natural language processing will definitely provide a proper direction to all researchers for better business intelligence decisions [5].

2. DATA SOURCES

2.1. Review sites

Reviews about products or services are available on review site. Review sites are helpful to both manufacturers and customers of the product. The manufacturers will get the sentiments through these sites [6]. Reviews on website can not be restricted. Positive reviews may be written by businesses or individuals being reviewed, while negative reviews may be fake.

2.2. Blogs

It is a periodically updated web page, run by a person or group. Blogs are written in an informal language. The bloggers can post blogs hourly, daily or weekly and make conversation fast and real time.

2.3. Micro blogging

It is a website where a user makes short, frequent posts. Micro blogging is a broadcast that is available in the form of blogs. A micro blog is smaller in its contents than traditional blog. Using micro blogs users exchange audio, video links, images, short texts which are used to comment their reputation. These micro blogs are also called as micro posts.

2.4. News articles

Most of the websites of daily news papers allow users to comment on ongoing event or issue. Rich site summary is also helpful to get the sentiments posted by readers.

2.5. Social aetworks

Social networks like Facebook, Twitter are lifeline of conversation for today's world. People are sharing their opinion through these websites.

a. Twitter

It is online micro blogging service helpful for reading or sending textual posts. The post is called as 'Tweet'.

b. Facebook

One can create his/het profile on Facebook. Images, text message, video can be uploaded on Facebook. Based upon privilege provided to a person he/she can see the profiles of their friends if added, to exchange text messages.

3. SENTIMENT CLASSIFICATION OUTLINE

SA architecture shows that preprocessing, feature extraction, sentiment classifications are the steps involved in it.

3.1. Basic terminologies

a. Opinion: Based on knowledge or experience about any object or event, one can express opinion about it. Mathematical expression for opinion is (o, po, f, ho, t), where o is object, po is the polarity of the opinion for particular feature of object or event, f is feature of object or event, ho is the person or group posted opinion t is the time at which opinion is impressed.

b. Opinion Holder: A person who expresses their views about any object or event is called as opinion holder.

c. Object: Object is any entity or event say A. It is a pair, A: (C, P), where C is the part of A, and P is the feature or attribute of A. An opinion is expressed as "I like Red Mi mobile phone", or based on attribute of Red Mi mobile phone, opinion can be given as "The display quality of Red Mi mobile phone is nice" [7]. Also, the component of an object or feature of an object can be used to express the opinion.

d. Object to Feature Form Simplification: The general opinion about an object can be expressed as "I like Red Mi Phone". Whereas the opinion is expressed on feature of the object as "Red Mi phone has a very good picture quality", here Red Mi phone has a feature 'picture quality'. Let, an object O has feature f in sentiment text [8]. Let d be a document, containing opinions about the products. The opinionated document is collection of opinions by users in statement s. Now, document d is expressed with the help of sentence series.

d = <s1, s2. . . sm>

e. Opinion Polarity: Opinion polarity expresses whether the opinion is positive, negative or neutral. The intensity of opinion can vary from strong to weak. A positive opinion can vary from good to excellent.

3.2. Preprocessing

We are interested in features of an object. For this preprocessing of input data are preprocessed using following steps.

a. Tokenization: White spaces, special characters, symbols are removed; remaining words are called as tokens.
b. Removal of Stop Words: The articles and common words like “a, an, the, this, that am, is”, etc.
c. Stemming: Reduces the tokens or words to its root form.
d. Case Normalization: It changes the whole document either in lower case letters or upper case letters [9].

3.3. Feature extraction

This step deals with the following scenario.

a. Feature Types: It deals with finding of types of features used for sentiments viz. term frequency, term co-occurrence, sentiment word, negation, syntactic dependency.
b. Feature Selection: It deals with finding good features for sentiment classification viz. information gain, odd ratio, document frequency, mutual information.
c. Feature Weighting Mechanism: It calculates weight for ranking the features using term frequency and inverse document frequency.
d. Feature Reduction: The dimensionality of features is reduced for better performance.Opinion posted is classified as positive opinion, negative opinion and neutral opinion. The 3 levels of sentiments analysis are as follows.

3.4. Levels in sentiment analysis

Opinion posted is classified as positive opinion, negative opinion and neutral opinion. The 3 levels of sentiments analysis are as follows.

3.4.1. Document level

The whole document is considered for impressing the opinion as positive, negative or neutral. The opinion about an object may be expressed without using any opinion word. In this case natural language processing plays a vital role to mine the correct sentiments. The main challenge is to extract subjective text for inferring the overall sentiment of the whole document.

3.4.2. Sentence level

The documents in collection are divided into sentences and then the sentences are classified as per positive, negative or neutral polarity. A document is a combination of subjective and objective sentences. First the subjective sentences are determined and then the opinion in those subjective sentences will be calculated.

The sentence level polarity identification can be done in either of the two ways: a grammatical syntactic approach or a semantic approach. The grammatical syntactic approach takes grammatical structure of the sentence into account by considering parts of speech tags. [10].

3.4.3.Word or phrase level

When product feature is considered for sentiment analysis, it is word or phrase level sentiment analysis. It uses adjective, adverb as features .Word level sentiment can be attained by ‘Dictionary Based Approach’ or ‘Corpus Based Approach’.

a. Dictionary based approach

Sometimes the opinion is not expressed by a popular keyword. Some jargons may be used to express the sentiments. Here WorldNet containing the synonyms and antonyms is considered for finding out the polarity of a word.

b. Corpus based approach

In this method, occurrence of any word with other word whose polarity is known is taken in to account. Adjectives joined by ‘and’ show the same impression and if joined by ‘but’ show opposite impression.

4. ROUTES TO ACHIEVE SENTIMENT ANALYSIS

Several algorithms, techniques are used to achieve sentiment analysis. Still many researchers are striving for improvement of existing methods and developing new effective methods. Types of sentiment analysis approaches is shown in Figure 1.

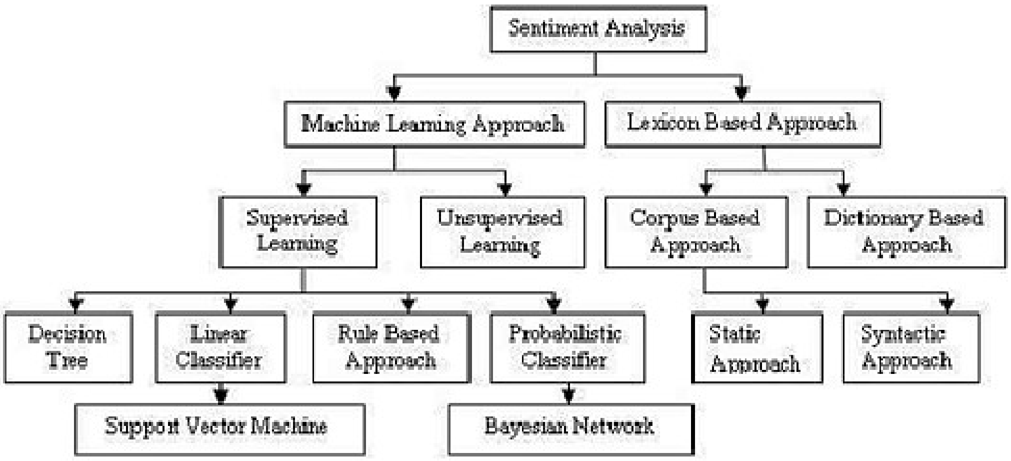

Figure 1. Types of sentiment analysis approaches

4.1. Machine learning strategy

Using Machine learning the performance P of current task T using past experience E is improved. Machine learning techniques first train the algorithm with some particular inputs to formulate a model. Lateron this model is used to test the new data for categorization. Some of the techniques are as follows:

4.1.1. Support vector machines (SVM)

SVM are supervised learning methods used for classification. SVM uses hyper plane to separate multidimensional data. All the data points on these planes are called as supports. The supports which are closer to hyper plane boundaries are called as support vectors, Figure 2. Number of hyper planes is possible for given data. The hyper plane with highest margin is selected for classification. With greater margin, less classification errors can be attained. SVM is performing well in text classification and in a variety of sequence processing applications [11].

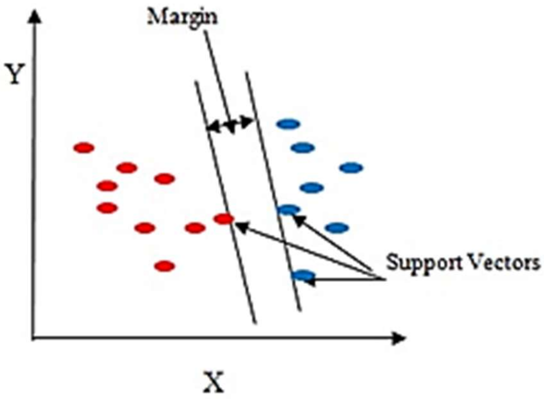

Figure 2. Vector space model

4.1.2. Naive bayes

It is simple in implementation with high accuracy. The algorithm will take every word in the training set and calculate the probability of it being in each class (positive or negative). It assumes that a feature value does not depend on other feature value. A naive Bayes classifier describes that the correlation between features is not considered to exist any object. For example Red color, round shape, 5 to 8 cm in diameter object is tomato. If a single feature is present, it will be considered as tomato.

In machine learning we are often interested in selecting the best hypothesis (h) given data (m).In a classification problem, our hypothesis (h) may be the class to assign for a new data instance (m). One of the easiest ways of selecting the most probable hypothesis given the data that we have that we can use as our prior knowledge about the problem Bayes [11].

Bayes' Theorem is given as:
P(h|m) = (P(m|h) * P(h)) / P(m)
Where
P(h|m)-It is posterior probability of hypothesis h when data m is given.
P(m|h) –It finds probability of data m considering that hypothesis h is true.
P(h) It is the prior probability of hypothesis h being true.
P(m) –The data probability.

4.1.3. Decision trees (DT)

DT divides input training data into subset and pattern is found from the subset. These patterns are used further for classification. The knowledge gained from patterns is used to construct a tree. A decision tree consists of a root node, interior nodes and leaf nodes. Root node consists of whole dataset, interior nodes split the incoming data set into two or more parts based on dissimilarity and leaf nodes represents classification [12].

The in record form data is as follows.

$(x, Y)=\{x_1, x_2, x_3........, x_k, Y)$

Y is dependent variable. The vector **x** is combination of the features, x_1, x_2, $x_{3..}$ $x_{k.}$
At each node best attributes to partition data into individual classes are chosen. The best partitioning attribute is selected using information gain measured are used in speech and language processing.

4.1.4. Rule based approach (RBA)

Relational rules represents the knowledge is the backbone of RBA. This classification technique is based on IF-THEN rules. It has very poor generalization, but provides better performance within a narrow domain. This approach includes learning classifier systems, association rule learning, and artificial immune systems. [12]. Rules can be generated by different methods. Support and confidence are the two common methods for generating rules.

The number of instances present in the training data set and which are relevant to the rules are called as support. The Confidence represents the conditional probability that the right hand side of the rule is proved if the left hand side is proved. A word with positive sentiment is assigned a score +1 and if the word is giving negative sentiment -1 score is assigned to it. A rule is analyzed by a target word. If two mobiles are compared based on their prices, the word 'price' is the target word to analyze the sentiment by RBA.

4.2. Lexicon based approach

It is unsupervised technique for SA. This approach can work with small amount of training data. Positive opinion deals with desired states and negative opinion deals with undesired states. Rich, good, beautiful are the words expressing positive sentiment. Poor; bad, dirty are the words expressing negative opinion. The opinion phrases and idioms together are called as opinion lexicon. There are 3 approaches for collecting the opinion word list: manual approach, dictionary-based approach, and corpus-based approach [8].

4.2.1. Dictionary based approach

Opinion words and their synonyms are found and saved as the part of dictionary. Then the errors from such dictionary are removed manually. This approach is has limited applicability. Using a small set of testing opinion words and an online dictionary this approach can be used e.g., Word Net or thesaurus [12].

4.2.2. Corpus based approach

A word may express different meaning in the same domain. [13, 14]. The word “small" gives opposite opinions in two sentences: “The mobile phone is small” (positive) and “The storage capacity of mobile phone is small" (negative). If two sentences or two words are joined by 'AND', both of them are impressing the common sentiment. Sometimes 'AND 'may gives us opposite sentiments like “The Red Mi phone battery is excellent and takes long time to charge”. It is expected that on both the sides of 'AND' same opinion is expressed. For the connectors 'OR', 'BUT', etc rules are designed. This idea is called sentiment consistency.

A large corpus is studied to find out the similar or different impression of words conjoined. Same and different opinion flow between adjectives form a graph. Any clustering algorithm is applied on this graph to form two clusters, say positive and negative.

4.3. Comparison of three major techniques 0f sentimental analysis

The three major techniques used in sentimental analysis are analyzed based on their performance and accuracy. The major advantages and disadvantages of using any approach are also discussed. The comparison of all these techniques is shown in Table 1.

Table 1. Comparison of Various Sentimental Analysis Approaches

Technique for SA	Classification	Advantages	Limitations
Decision Tree	Supervised Learning	1. It is easy, fast and simple to interpret 2. Data containing errors can be handled.	1. Poor performance if so many complex interactions are present. 2. The decision tree has greedy characteristic. 3. Over-sensitivity to the training set, to irrelevant attributes and to noise.
Linear Classifier (SVM)	Supervised Learning	1. High accuracy 2. SVM's can model non-linear decision boundaries. 3. They are fairly robust against over fitting, especially in high dimensional space.	1. SVM's are memory intensive 2. Do not scale well to larger datasets. 3. Hard to interpret.
Rule Based Approach	Supervised Learning	1. High accuracy 2. Require lesser data but need expert human labour.	Rules must need to define accuracy as performance is highly rule dependent
Probabilistic Classifier (Bayesian Network)	Supervised Learning	1. Easy to undersatand and Implement. 2. Easily updateable if new training data is received. 3. Small memory footprint.	1. Don't use when features are correlated. 2. Data set is large so usually it is not preferred
Corpus Based Method	Unsupervised Learning	1. It finds domain dependent opinions. 2. It can reflect the characteristics of the unstructured text data.	Big data preparation for containing all English words is challenging.
Dictionary Based Method	Unsupervised Learning	1. Good to find a lot of such words 2. It is a simple technique uses dictionary	Excessively rely on emotional dictionary.

5. APPLICATIONS

As large opinionated data sources are available, it is being used in many applications. Some applications of SA are enlisted as follows [15, 16].

a. Business
Demand and supply decisions can be made based on the impressions by user of that product. Through these impressions, organization can check the quality of the service they are providing. To grow in the market decisions can be made based on sentiments available timely.

b. Politics
Political issues discussed on political blogs. The positive and negative popularity of public figure can be known by analyzing the sentiments of people available on social websites.Business

c. Recommender system
The product or service is will have positive sentiments if people are using it happily. These products can be recommended to a new user strongly, if user is viewing the ratings or sentiments. So sentiment analysis is having a vital role in recommender system.

d. Summarization
It is time consuming for a reader to read all opinions for a given entity and then judge.SA will give us the summarized opinion of any entity for a given tenure. Recommender System

e. Government intelligence
For monitoring the sources, the increase in aggressive communication can be tracked. For making policies, the sentiments of people can be studied. For analyzing mentality of people towards any controversy SA can be used.

6. CHALLENGES

Sentiment analysis is facing following challenges [17].

a. Negation handling

Negation alters the polarity of the associated adjective and hence the text. Negation words (like not, neither, nor, it is not the cases) are playing a vital role for polarity change. One possible solution to handle negations is to reverse the polarity of the adjective occurring after negating word. However, this solution fails to deal with the cases like “No wonder the mobile is good" and "Not only the menu was delicious, the culture and service was also remarkable". The use of pure language processing techniques, mathematical models fails to deal with negations completely.

b. Domain generalization

Polarity of adjective changes when is used in different domains. For example, "The play was inspired from Shakespeare’s play" (negative orientation), "I got inspired from the autobiography" (positive orientation). Here, the word 'inspired' exhibits two different polarities for two different contexts. A generalized sentiment analyzer still remains a challenge because of difference in the meaning of a word/sentence in different domains.

c. Pronoun resolution

Generally the opinion for an object is expressed in the first sentence of the complete text. Remaining portions may contain opinionated text expressed using pronoun. To resolve the pronoun i.e. for what it has been used for is a challenging job.

d. Language generalization

A separate dictionary is required for separate domain and language. Most of the sentiment analyzers are implemented in English language. A language general sentiment analyzer is fruitful as it gives a broad view of sentiments about a product.

e. World knowledge

One entity can be referred using other entity. For example, “He is as intelligent as Einstein". Here, to identify the sentiment orientation of the text, one has to know about 'Einstein''.

f. Detection of spam and fake reviews:

The spam and fake review can be eliminated before preprocessing by using outlier analysis and by considering reputation of reviewer.

g. Single sentence possessing multiple opinions

A sentence may contain multiple opinions for different features. In this case the strength of opinion for a feature should be computed properly.

h. Comparative sentences

The order of words in a sentence expresses different meaning. The sentence, “Red Mi mobile is better than Nokia mobile” gives opposite opinion from “Nokia mobile is better than Red Mi mobile”.

i. Synonym gathering

Two words may express same feature of an object. These synonym words should be grouped together. In this example, “cute” and “excellent” both refer to the same feature.

7. CONCLUSION

In this paper, basic concepts related to sentiment analysis are discussed thoroughly. The techniques for sentiment analysis are discussed and compared on the basis of their advantages and limitations. The contribution of current research study is to distinguish positive, negative or neutral nature of sentiments based using explicit opinions. It is needed to deal with implicit opinion with its feature. In future implicit and explicit opinions and their features may be considered to find better sentiment classification.

REFERENCES

[1] Liu, B., “Sentiment Analysis and Subjectivity”, *Handbook of Natural Language Processing*, 2nd edition, pp. 1–38, 2012.

[2] Liu, B.” Sentiment Analysis: A Multi-Faceted Problem,” *IEEE Intelligent Systems*, pp. 1–5, 2010.

[3] Dalal, M.K., Zave, M.A.” Automatic Text Classification: A Technical Review,”*Inernational Journal. Compuer Appications,* (0975–8887) ,Vol 28,No. 2,pp. 37–40,2011.

[4] Ayesha Rashid, Naveed Anwer, Dr. Muddaser Iqbal, Dr. Muhammad “A Survey Paper:Areas, Techniques and Challenges of Opinion Mining”, *IJCSI International Journal of Computer Science Issues*, Vol. 10, Issue 6, No 2, November 2013.

[5] N. Anitha, B. Anitha, S. Pradeepa, “ Sentiment Classification Approaches – A Review”, *International Journal of Innovations in Engineering and Technology*, Volume 3, Issue 1, October 2013

[6] Kumar P K, Nandagopalan S.**,”** Insights to Problems, Research Trend and Progress in Techniques of Sentiment Analysis**”,** *International journal of Electrical and Computer Engineering***,** Vol-7, No-5, Oct. 2017

[7] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", *Proceedings of EMNLP*, pp. 79--86, 2002

[8] Castillo, Carlos, Marcelo Mendoza, Barbara Poblete, *"*Information Credibility On Twitter*", AMC, Proceedings of the 20th international conference on World Wide Web, 2011.*

[9] K. Denecke, "Using Sentiwordnet For Multilingual Sentiment Analysis", *Data Engineering 10 Workshop, 2008, ICDEW 2008*, pp. 507– 512, 2009.

[10] Blessy Selvam, S.Abirami, "A Survey On Opinion Mining Framework", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol 2, Issue 9, September 2013.

[11] Madina Hamiane, Fatema Saeed**," *SVM*** Classification of MRI Brain Images for Computer-Assisted Diagnosis**",** *International journal of Electrical and Computer Engineering*,Vol-7, No-5, Oct.2017.

[12] Liu Bing, Hsu Wynne, Ma Yiming,"Integrating Classification And Association Rule Mining": *Presented at the ACM KDD conference;* 1998.

[13] Hu Minging, Liu Bing" Mining And Summarizing Customer Reviews*", Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'04);*2004.

[14] Kim S, Hovy E.,"Determining The Sentiment Of Opinions", *Proceedings of interntional conference on Computational Linguistics (COLING'04)*; 2004.

[15] Hatzivassiloglou, V. and K. McKeown,"Predicting The Semantic Orientation Of Adjectives", *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-1997)*, 1997.

[16] Kanayama, H. and T. Nasukawa,"Fully Automatic Lexicon Expansion For Domain-Oriented Sentiment Analysis", *In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, 2006.

[17] Lekha R. Nair, Sujala D. Shetty, Siddhant Deepak Shetty,"Streaming Big Data Analysis for Real-Time Sentiment based Targeted Advertising," International journal of Electrical and Computer Engineering,Vol-7, No-1, Feb.2017.

BIOGRAPHIES OF AUTHORS

Prakash P.Rokade has received his B.E.degree in Computer Pune University, Maharashtra; Indiain 2005.He has received his M.Tech. degree Computer Engineering from BhartiVidyapeerth, Pune, Maharashtra,India in 2011 and presently pursuing his Ph.D. in Computer Science and Engineering Koneru Lakshmaiah Education Foundation, formerly K L University, Vaddeswaram , Andhra Pradesh, India.Hisr research interest includes Sentiment Analysis, Opinion Mining,and Machine Learning.

Dr. Aruna Kumari D has received her Ph.D. degree in Computer Science and Engineering from the K L University, Vaddeswaram, Andhra Pradesh, India. Currently, She is Professor Koneru Lakshmaiah Education Foundation, formerly K L University. Her teaching and research areas include in data Mining, Machine Learning and has published more than 50 papers in many National, International journals.She is honoured by DST Young Scientist Award (Govt. of India).