A Survey of Sentiment Analysis techniques

Harpreet Kaur
UIET, Panjab university
Chandigarh, India

Veenu Mangat
UIET, Panjab university
Chandigarh, India

Nidhi
UIET, Panjab university
Chandigarh, India

***Abstract*: Sentiment analysis is an application of natural language processing. It is also known as emotion extraction or opinion mining. This is a very popular field of research in text mining. The basic idea is to find the polarity of the text and classify it into positive, negative or neutral. It helps in human decision making. To perform sentiment analysis, one has to perform various tasks like subjectivity detection, sentiment classification, aspect term extraction, feature extraction etc. This paper presents the survey of main approaches used for sentiment classification.**

Key words: sentiment analysis, sentiment classification, features selection, machine learning.

I. INTRODUCTION

Sentiment analysis is the process of extracting emotions or opinions from a piece of text for a given topic.[1] it allow us to understand the attitudes, opinions and emotions in the text. In it user's likes and dislikes are captured from web content. It involves predicting or analyzing the hidden information present in the text. This hidden information is very useful to get insights of user's likes and dislikes. The aim of sentiment analysis is to determine the attitudes of a writer or a speaker for a given topic. Sentiment analysis can also be applied to audio, images and videos.[2]
Today internet has become the major part of our life. Most of the people use online blogging sites or social networking sites to express their opinions on certain things. They also use these sites to know what other people's opinions are. Thus mining of this data and sentiment extraction has become an important field of research.

II. IMPORTANT NOTIONS

A) *Subjectivity/Objectivity*- To perform sentiment analysis we first need to identify the subjective and objective text. Only subjective text holds the sentiments. Objective text contains only factual information.
Example-
1.) Subjective**:** Titanic is a superb movie.
(this sentence has a sentiment(superb), thus it is subjective)
2.)Objective**:** James Cameron is the director of titanic.(this sentence has no sentiment, it is a fact ,thus it is objective)[3]

B) *Polarity*- Further subjective text can be classified into 3 categories based on the sentiments conveyed in the text.
1.) Positive: *I love new Samsung galaxy mobile.*

2.) Negative: *The picture quality of camera was awful.*

3.) Neutral: *I* usually get hungry by noon. (this sentence has user's views, feelings hence it is subjective but as it does not have any positive or negative polarity so it is neutral. This positive, negative and neutral nature of text is termed as polarity of text. There is a lot of debate whether to take two or three classes but it is found that by considering neutral class accuracy gets increased. There are two ways for it: either classify text into two classes positive/negative and neutral and then further handling positive/negative or classify text into three classes in first step only[3].

C) *Sentiment level*- sentiment analysis can be performed at various levels -

- Document Level- In it the whole document is given a single polarity positive, negative or objective[1] .
- Sentence Level – In it document is classified at sentence level. Each sentence is analyzed separately and classified as negative, positive or objective. Thus overall document has a number of sentences where each sentence has its own polarity.
- Phrase Level- It involves much deeper analysis of text and deals with identification of the phrases or aspects in a sentence and analyzing the phrases and classify them as positive, negative or objective. It is also called aspect based analysis.[3]

III. APPLICATIONS

A)Support in decision making: Decision making is a very important field of our life. Opinions extracted from reviews helps us in making various decisions like "which books to buy", "which hotel to go", "which movie to watch" etc.

B)Business application: In today's world of competition, every company wants to satisfy its customer's requirements by creating new innovative products. Assessments of individuals are an essential angle today with the goal that organizations can get an input from clients and can roll out sought improvements in their item. Google Product Search is one illustration.

C)Predictions and trend analysis: sentiment analysis enables one to predict market trends by tracking views of public. It is also helpful in elections where candidates wants to know the expectations of people from them.

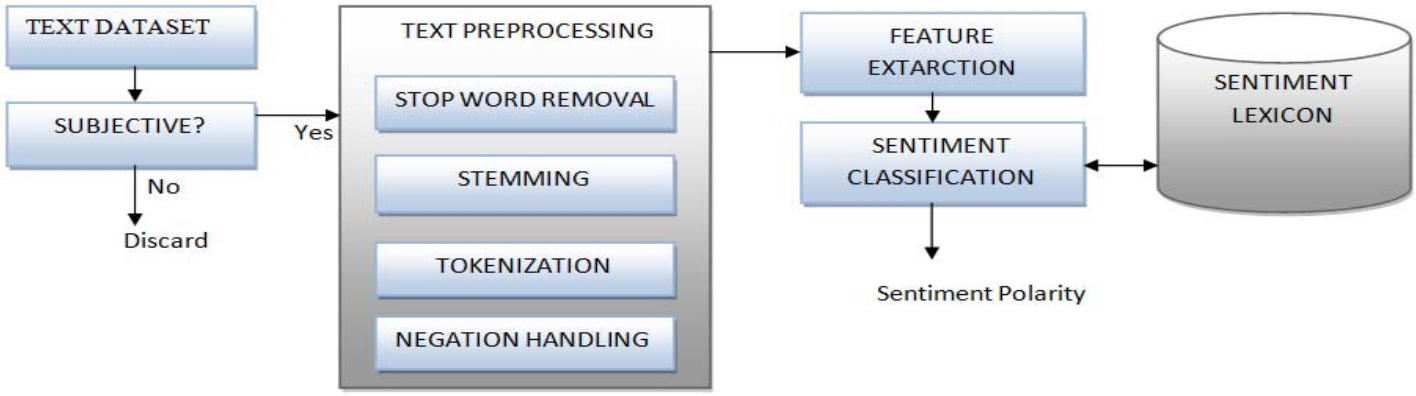

Figure 1: Sentiment Analysis Process

IV. METHODOLOGY

A)*Feature selection*: To perform sentiment classification, first task is to extract the features from text which are

1) N grams- n grams refers to consecutive n terms in text. One can take only one word at a time(unigram) or two words(bigram) up to n accordingly. Some sentiments can't be captured with unigram feature. For example this drink will knock your socks off. It is a positive comment if socks off is taken together and negative in case of only unigram model (off).

2) POS tagging- It is a way toward denoting a word in a content (corpus) as comparing to parts of speech in light of both its definition and its association with contiguous words. Nouns, pronouns, adjectives, adverbs etc are examples parts of speech. Adjectives and adverbs hold most of the sentiments in text.[11]

3) Stemming-It is the process of removing prefixes and suffixes. For example 'playing', 'played' can be stemmed to 'play'. It helps in classification but sometimes leads to decrease in classification accuracy.

4) Stop words- Pronouns (he/she, it), articles (a, the), prepositions (in, near, beside) are stop words. They provide no or little information about sentiments. There is a list of stop words available on the internet. It can be used to remove them in the pre-processing step.

5) Conjunction handling- In general, each sentence expresses only one meaning at a time. But certain conjunction words like but, while, although, however changes the whole meaning of the sentence. For example *although movie was good but it was not up to my expectations*. By using these rules throughput can be increased by 5%.[6]

6) Negation handling- Negation words like 'not' inverts the meaning of whole sentence. For example The movie was not good has 'good' in it which is positive but 'not' inverts the polarity to negative.

B)*Sentiment classification*

Two approaches are mainly used

1) Subjective lexicon: Subjective lexicons are collection of words where each word has a score indicating the positive, negative, neutral and objective nature of text. In this approach, for a given piece of text, aggregation of scores of subjective words is performed i.e. positive, negative, neutral and objective word scores are summed up separately. In the end there are four scores. Highest score gives the overall polarity of the text.[12]

a) Dictionary based approach- In this approach a set of opinion words are manually collected and a seed list is prepared. Then we search for dictionaries and thesaurus to find synonyms and antonyms of text. The newly found synonyms are added to the seed list. This process continues until no new words are found.

Disadvantage: difficulty in finding context or domain oriented opinion words

b) Corpus based approach- Corpus is collection of writings, often on a specific topic. In this approach, seed list is prepared and is expanded with the help of corpus text.[14] Thus is solves the problem of limited domain oriented text. It can be done in two ways

- Statistical approach: This approach is used to find co-occurrence words in the corpus. Idea is that if the word appears mostly in positive text, then its polarity is positive. If it mostly occur in negative text, then its polarity is negative.
- Semantic approach: This approach calculates sentiment values by using the principal of similarity between words. Wordnet can be used for this purpose. Synonyms and antonyms of given word can be found using this and sentiment value can be calculated. [2]

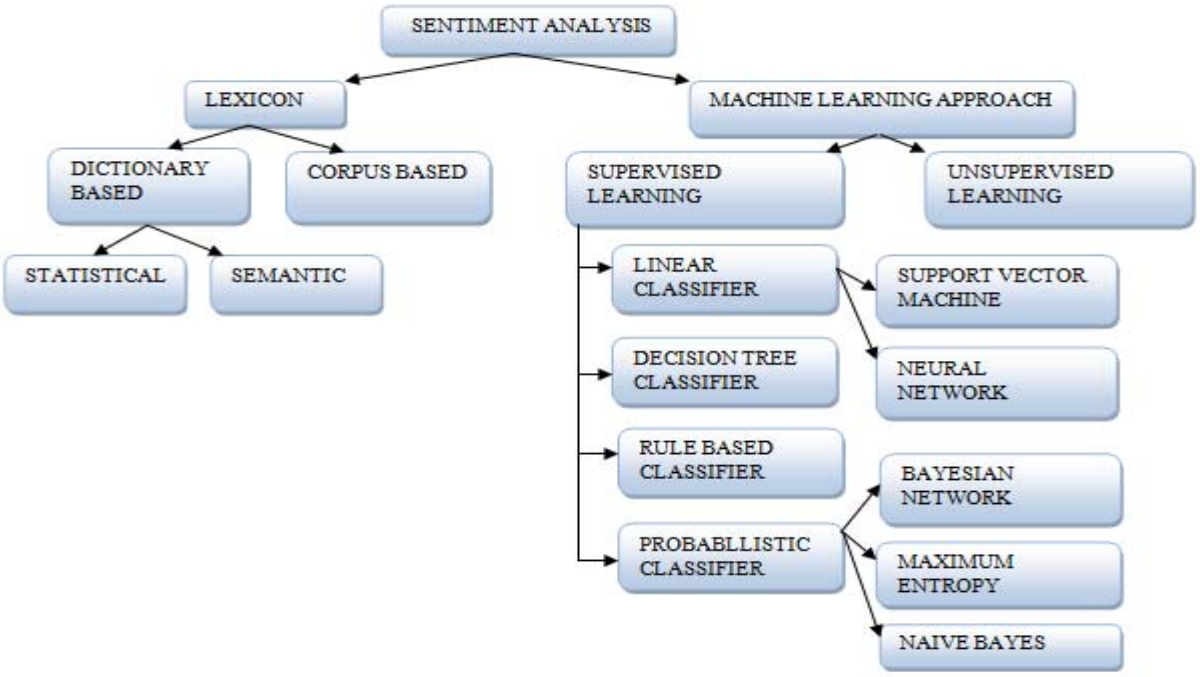

Figure 2: Sentiment Classification Techniques

2) Machine learning: This is an automatic classification technique. Classification is performed using text features. Features are extracted from text. It is of two types- supervised and unsupervised

a) Supervised- System is trained using labeled training examples. Each class represents different features and has a label associated with it. When a word arrives, its features are compared and labeled with a class with maximum matching.

1.Probabilistic classifier: This classifier is able to foresee a probability function over a set of classes for a given input data. It does not give only the most likely classes but a probability function over all classes. For example an ordinary classifier function assign a label y to input x as

$$y=f(x)$$

In case of probabilistic classifier this function is replaced with conditional distributors Pr(Y/X) i.e. for given x X , probability is assigned to all y Y as

$$y= \arg\max Pr(Y=y/X)$$

- Naïve bayes*:* This classifier uses bayes theorem to predict the probability that a given set of features is a part of particular label. It uses bag of words (BOW) model for feature extraction. This model assumes that all the features are independent.[12]

$$P(label/features)=P(label)* P(features/label)/P(features)$$

Where P(label)= prior probability of label
P(features/label)=prior probability that feature set is classified as label
P(features)= prior probability that feature set will occur

- Bayesian network: This model assumes that there is a strong dependence between features. It is a directed acyclic graph in which nodes represents random variables and edges represent dependencies. It is very expensive model therefore is not frequently used.
- Maximum entropy: By using encoding, this classifier converts the labeled feature sets to vectors.[11]This vector is used to calculate weights of features which can then be used to predict label for each feature set. Encoding maps each pair i.e. P(featureset, label) into vectors. Probability of label can be computed as

P(fs/label) = dotprod(weights, encode(fs, label)/
sum(dotprod(weights, encode(fs,l))forlinlabels)

2. Linear classifier-It performs classification based on the linear combinations value of the characteristics. Let W={w1, w2, w3….} is word frequency , vector C={c1, c2, c3…..} is linear coefficient vector and S is a scalar then output of linear predictor will be

$$LP=W.C+S.$$

This predictor is called hyperplane which separate two classes.

- SVM: Support vector machine is a supervised learning model which is used for classification. Its main aim to determine best linear separators for classification. It is a non probabilistic classifier[17]. For a given set of training data, each is labeled for belonging to one of the classes, SVM training algorithm create a model which assign new data into one or two classes. Hyperplane is used to separate

two classes. In the diagram below, for example, to classify triangle and circle shapes we compute three hyperplanes A, B and C. C is best separator as items on both sides are at maximum distance from Hyperplane and B is worst separator.

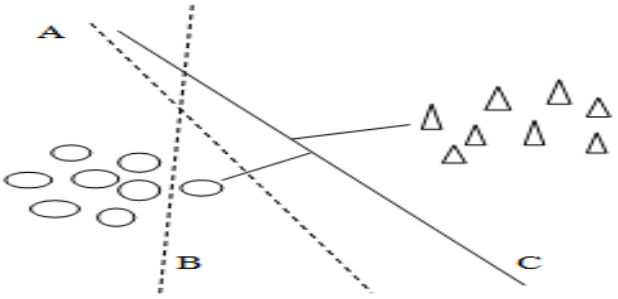

Figure 3: Classification using Support Vector Machine

- Neural network: Neural network electronic networks of neurons similar to neural structure of brain. Neuron is the basic element of this network. Neurons are placed in three layers- input, hidden and output. A neuron consists of set of inputs (vi) and associated weights (wi). It also has a function (fi) that adds the weights and maps these results to the final result i.e. output(y).[16]

3. Decision tree classifier- In this classification, a condition is used to divide the data. Data which satisfy the condition is placed in one class and rest data in other class. It is a recursive procedure. There are a number of splits : single attribute split which searches for particular word presence or absence to perform classification, similarity based multi attribute split which matches the given document words with predefined words to perform classification and Discriminate based multi-attribute split use discriminates to perform split.

4. Rule based classifier- This classifier makes use of certain rules as IF , THEN. It can be written as

IF condition THEN decision

The rules can be generated during training phase depending on our requirements. [2]

References	Year	Task	Data set	Algorithm
7	2011	Sentiment analysis	Digital camera reviews	Multi class SVM
8	2011	Sentiment analysis	Training data in Chinese	Semantic
9	2011	Sentiment classification	Movie reviews	Lexicon based, semantic
10	2011	Sentiment analysis	Product reviews	Statistical(machine learning), semantic
11	2012	Feature selection	Movie reviews	Statistical, maximum entropy
12	2012	Sentiment classification	Restaurant reviews	Naïve bayes, svm
13	2012	Sentiment analysis	News	Lexicon based
14	2012	Emotion detection	Blogs data	Corpus based
15	2012	Emotion detection	Emotions corpus	Lexicon based, SVM
16	2013	Sentiment classification	Movie, camera, book, GPS reviews	Artificial neural network, SVM
17	2013	Sentiment classification	Tweets and movie reviews	SVM, Naïve bayes
18	2014	Sentiment analysis	Facebook data	Lexicon based, machine learning
19	2015	Sentiment analysis	Tweets	Hybrid(lexicon+ learning algorithm)
20	2015	Sentiment analysis	Movie, book, product reviews	SVM
21	2016	Sentiment analysis	Tweets	Lexicon based
22	2016	Sentiment analysis	Starbucks twitter dataset	Dynamic architectural artificial neural networks

Table 1: Articles summary

CONCLUSION & FUTURE SCOPE

This paper presents a survey of sentiment analysis and classification algorithms. This survey concludes that sentiment classification is still an open field for research. There is a lot of scope for algorithms in it. SVM and naïve bayes are most popular algorithms for sentiment classification. Sentiment analysis of tweets is very popular. Datasets from sites like Amazon, IMDB, flipkart are widely used for sentiment analysis. Deeper analysis is required in case of social networking sites. In many cases, context consideration is very important. Therefore more research is required in this field.

REFERENCES

1. Pang B, Lee L. , "Opinion mining and sentiment analysis" FoundTrends Inform Retriev:1–135, 2008.

2. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey" Ain Shams Engineering Journal 5.4 :1093-1113, 2014.

3. Arora, Piyush. "Sentiment Analysis for Hindi Language." Diss. International Institute of Information Technology Hyderabad, 2013.

4. Kaur, Amandeep, and Vishal Gupta. "A survey on sentiment analysis and opinion mining techniques." Journal of Emerging Technologies in Web Intelligence 5.4: 367-371, 2013.

5. Esuli, Andrea, and Fabrizio Sebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." *Proceedings of LREC*. Vol. 6. 2006.

6. Hemnaath, R., and Low, B.W. "Sentiment Analysis Using Maximum Entropy and Support Vector Machine." Semantic Technology and Knowledge Engineering,2010.

7. Chin Chen Chien, Tseng You-De. "Quality evaluation of product reviews using an information quality framework". Decis Support Syst;50:755–68, 2011.

8. Zhou L, Li B, Gao W, Wei Z, Wong K. "Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities",conference on Empirical Methods in Natural Language Processing (EMNLP'11),2011.

9. Heerschop B, Goossen F, Hogenboom A, Frasincar F, Kaymak U, de Jong F., "Polarity Analysis of Texts using Discourse Structure", ACM Conference on Information and Knowledge Management (CIKM'11),2011.

10. Hu Nan, Bose Indranil, Koh Noi Sian, Liu Ling. ''Manipulation of online reviews: an analysis of ratings, readability, and sentiments''.Decis Support Syst ,52:674–84, 2012.

11. Duric Adnan, Song Fei., "Feature selection for sentiment analysis based on content and syntax models", Decis Support Syst,53:704–11, 2012

12. Kang Hanhoon, Yoo Seong Joon, Han Dongil., "Senti-lexicon and improved Naı¨ve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Syst Appl ,39:6000–10, 2012

13. Moreo A, Romero M, Castro JL, Zurita JM. "Lexicon-based comments-oriented news sentiment analyzer system" Expert Syst Appl ,39:9166–80, 2012

14. Keshtkar Fazel, Inkpen Diana., "A bootstraping method for extracting paraphrases of emotion expressions from texts" Comput Intell;vol. 0, 2012

15. Balahur Alexandra, Hermida Jesu´ s M, Montoyo Andre´ s. "Detecting implicit expressions of emotion in text: a comparative analysi" , Decis Support Syst ,53:742–53, 2012

16. Moraes Rodrigo, Valiati Joa˜o Francisco, Gavia˜o Neto Wilson P., "Document-level sentiment classification: an empirical comparison between SVM and ANN", Expert Syst Appl ,40:621–33, 2013.

17. Rui Huaxia, Liu Yizao, Whinston Andrew., "Whose and what chatter matters? The effect of tweets on movie sales", Decis Support Syst 2013.

18. Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." Computers in Human Behavior ,31 : 527-541, 2014

19. Khan, Aamera Z H; Atique, Mohammad; Thakare, V M. International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE), suppl. National Conference on "Advanced Technologies in..:89-91, 2015.

20. Agarwal, Basant, et al. "Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach." Cognitive Computation ,7.4 :487-499, 2015

21. Zimbra, David, M. Ghiassi, and Sean Lee. "Brand-Related Twitter Sentiment Analysis Using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks." 49th Hawaii International Conference on System Sciences (HICSS). IEEE, 2016.

22. Saif, Hassan, et al. "Contextual semantics for sentiment analysis of Twitter." Information Processing & Management 52: 5-19,2016.