IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

A Brief Study on Sentiment Analysis & Opinion Mining

Jasneet Kaur

Department of Information Technology, Dr. Akhilesh Das Gupta Institute of Technology & Management, New Delhi, India

DOI: https://doi.org/10.26438/ijcse/v7i5.10511056 | Available online at: www.ijcseonline.org

Accepted: 18/May/2019, Published: 31/May/2019

Abstract- Due to Access of Internet and social media now a days, everyone, irrespective of age and area of concern is able to express his opinions regarding any entity, which can be a product, an article, a blog or just a simple tweet. These reviews thus plays a major role for marketers, customers or product analysts in creating opinions regarding a particular entity. This has led to creation of 2.5 quintillion bytes of data every day. Sentiment analysis or opinion mining is a branch of data mining which deals with the study of opinions and sentiments of the peoples which are expressed over internet. This paper presents a detailed study of approaches made so far for opinion mining, the comparison of data mining techniques and algorithm and their accuracy on various data sets. The paper also include various challenges that may be faced during the analysis of opinionated data.

Keywords: Data mining, Knowledge discovery, Opinion Mining, Polarity check, Sentiment analysis.

I. Introduction

Data on web is increasing day by day. Most of this data is in the form of text. This data is required to be organised efficiently. Information over web can be categorised into two parts: one can be fact and other can be opinions [1]. Facts can be the particular information that is available on web whereas opinion are the ideas of any individual which are expressed on various social media or ecommerce websites. Initially the aim was just to organise the facts but with the emerging trend of social media and websites where opinions were evolving in huge numbers there was a need to manage this data.

Sentiment analysis, also known as opinion mining, is a way of analysing the opinionated text on the web and extracting the sentiments or emotions of the writer. Sentiment analysis therefore is a way of developing a system which collects the data containing people's opinions, applies data mining and Natural language processing techniques to categorise the opinions according to the likes and dislikes of the users. A considerable amount of opinions otherwise will be very difficult to analyse and summarize.

The summarized opinion will be useful for the customers as it is not convenient for them to read all the reviews and decide which product will be good for them. Similarly, it will be useful for the company to improve their future marketing strategy in order to progress their sales and to keep their customers satisfied [2].

An opinion or sentiment of a person can be expressed as a five tuple structure [3]:

(e,a,s,h,t)

Where,

e- is the entity on which the opinion has been expressed,

a- is the aspect or feature of entity e,

s- is the sentiment of opinion holder h associated to the aspect a of entity e. It can be positive, negative or neutral,

h- is opinion holder,

t- is the time at which s is expressed

The paper is organized as follows: Section I contains the introduction of sentiment analysis , Section II explains the various levels at which sentiment analysis is performed, Section III defines the steps and technique for sentiment classification, Section IV includes the comparison of various techniques of sentiment classification and their respective accuracies, Section V mentions the challenges which are faced during the process of sentiment analysis, Section VI contains the conclusion and Section VII includes references

II. Levels of Opinion mining

Sentiments of a writer can be expressed at various levels of a text [4]:

Document level: In this level, the entire document is analysed to determine its polarity as positive, negative or neutral. A sole polarity is assigned to whole document.

Sentence level: Each sentence is analysed for its polarity separately. Here first, the sentence is test for its subjectivity

or objectivity. A subjective sentence is a sentence which has some sentiment in it, for eg: *"This is a great phone under 20k"*. This sentence has a positive sentiment i.e. "great" in it. An objective sentence does not have any opinion in it , for eg: "*My friend bought this phone recently*". A subjective sentence is therefore only judged for the polarity and each sentence is classified as negative, positive or neutral.

Neutral statements must not be confused by objective sentence. Example of a neutral polarity sentence can be "*I don't feel like studying*." This sentence have a sentiment but it cannot be classified as a negative or positive sentence therefore it is a neutral sentence.

Aspect level: Aspect level sentiment analysis deals with the opinions expressed on features of any product. For example "*The battery life of this phone is awesome but the picture quality is not so good*". Here the opinions are expressed on two features, one positive and one negative, of a particular entity.

III. Methodology for sentiment classification:

1. **Extraction of a Database [5,6]:** Dataset containing opinions of the users are to be extracted first from web. The dataset can be of any e commerce website, twitter, restaurant/hotel review website or any news or social media blog.
2. **Pre-processing of the data [5,6]:** Pre-processing means refining the data to make it suitable for application of classification techniques. Pre-processing include following steps:
   - **Tokenization:** It involves dividing the document into tokens or separate set of characters so that the document can be represented as a bag of words. Tokenization can be done by n-gram model where n is the consecutive n words. i.e. unigram, bigram, trigram. Each token is then considered as a separated feature.
   - **Removal of stop words:** Stop words are commonly used words such as 'a','an','the','is','for','to' , any html tags, url's, punctuation marks etc .These words doesn't have any specific meaning and they do not contribute much in the process of classification. Therefore to make the dataset more simplified these words are removed in pre-processing.
   - **Stemming:** Stemming is a way of converting a particular word into its base form for example: 'played' & 'playing' can be converted into 'play'. Also converting words like "soooooooo good" into "so good".
3. **Sentiment Analysis Techniques [7,8]:** After pre-processing phase, Sentiment classification of the text can be done by any following approach:
   - **Dictionary Based Approach:** This is the most easiest and quick approach to sentiment classification. A predefined set of positive and negative word list or certain rules are maintained and then number of positive and negative words are counted in the text. If appearance of positive keywords are more than negative keywords then then positive sentiment is returned otherwise a negative sentiment is returned.
     One drawback of this approach is that the dictionary must be kept updated with every possible occurrence of sentiment words in both positive and negative lists.
   - **Machine learning approach:** Machine learning algorithms classifies a given text into negative, positive or neutral without creating any pre-defined rules or any set of dictionary. Before applying any of the machine learning algorithm on the text firstly feature extraction of the text if performed. In feature extraction, the text is classified as bag of words. Where each word is assigned a frequency i.e. number of times that word has appeared in the text.
     Machine learning algorithm is then fed with the extracted features and a training dataset which has predefined labelled set of positive, negative and neutral wordlist. The output from machine learning algorithm is then fed into the classifier model which classifies the extracted features into one of the categories of the opinions.
     Machine learning algorithm can be further divided into two types:
     - Supervised learning: In this type of machine learning, the algorithm is fed with some labelled training set according to which classification is done. Some of the widely used machine learning algorithms are Naïve Bayes classifier, Support Vector machine (SVM), Maximum Entropy, K Nearest Neighbour, linear & logistic regression & decision trees.
     - Unsupervised learning: Every time it is not possible that the labelled dataset is available for sentiment classification. Unsupervised learning algorithm works without a training dataset by forming clusters or groups of data based on the commonalities among each different portion of data. Some of the common unsupervised learning algorithms are K-Means , DBSCAN & Neural

Networks(self-organizing map (SOM) and adaptive resonance theory (ART))

4. **Evaluation & Decision Making [5]:** The accuracy of the classifier is calculated on the basis of TP(True positive), TN(True negative), FP(False positive) & False negative. These values are the comparison of the sentiment label that a document has been assigned by the classifier with the sentiment class that item actually belongs to.

TP: are the items that are correctly classified as positive by the classifier

FP: are the items that should have been classified as positive by the classifier but they are not classified as such.

TN: are the items that are correctly classified as negative by the classifier

FN: are the items that should have been classified as negative by the classifier but they are not classified as such.

Table 1. Confusion matrix

		Correct labels	
		Positive	Negative
Classified labels	Positive	TP(True positive)	FP(False positive)
	Negative	FN(False negative)	TN(True negative)

Other parameters used for evaluation are:

- ✓ Accuracy : It is evaluated on the basis of correctly classified items with respect to total number of items

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- ✓ Precision & Recall [8] : Precision defines the exactness of the classification i.e how many items were classified correctly of a given category out of all the items(incorrect or correct) that were labelled for that category whereas recall defines how many items were classified correctly of a given category out of all the items(incorrect or correct) that should have been labelled for that category.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

- ✓ F - Measure: F Measure combines the score of precision and recall.

$$F = \frac{2*precision*recall}{precision+recall}$$

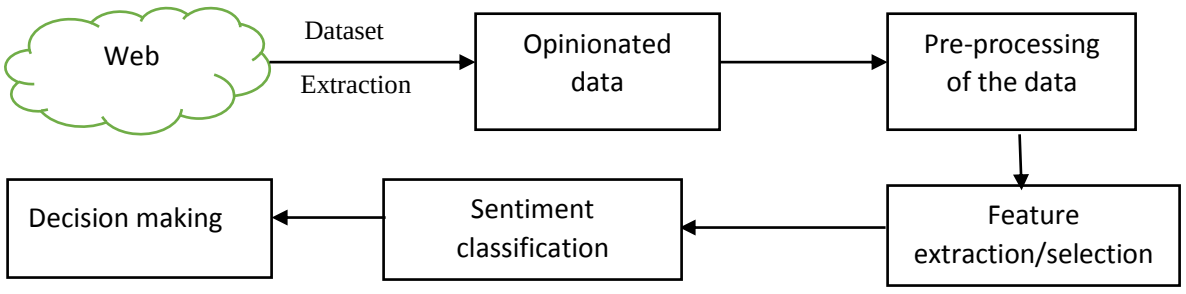

Fig 1. Steps involved in Sentiment classification

IV. Comparison Table

Table 2. Accuracy Comparison

no.	Ref.	Authors	Title of the paper	Classification Techniques used	Accuracy (%)
1.	[9]	Gautami Tripathi and Naganna S.	Feature selection and classification approach for sentiment analysis	SVM + TF - IDF	85%
2.	[10]	Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li	Dual Sentiment Analysis considering two sides of one review	Dual Sentiment Analysis (DSA) model.	73%
3.	[11]	Onam Bharti Mrs. Monika Malhotra	Sentiment analysis on twitter data	Naïve Bayes KNN	79.66 83.59
4.	[12]	Akshay Amolik, Niketan Jivane	Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques	Naïve Bayes SVM	86 90
5.	[13]	Jaspreet Singh, Gurvinder Singh	Optimization of sentiment analysis using machine learning classifiers	Naïve Bayes J-48 BF TREE ONE R	85 87 84 87
6.	[14]	Raheesa Safrin, K.R.Sharmila	Sentiment analysis on online product review	K means clustering	90.47
7.	[15]	Naw Naw	Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers	SVM KNN	72.4 65.2
8.	[16]	H. M. Keerthi Kumar , B. S. Harish ,	Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method	SVM Naïve Bayes Max Entropy KNN (Using Information gain feature selection)	75.4 54.7 78.3 72.2
9.	[17]	Bharat Gaind, Varun Syal, Sneha Padgalwar	Emotion Detection and Analysis on Social Media	SMO J48	91.7 85.4
10.	[18]	Ali Hasan , Sana Moin	Machine Learning-Based Sentiment Analysis for Twitter Accounts	Naïve Bayes *TextBlob *Sentiword Net *W-WSD SVM *TextBlob *Sentiword Net *W-WSD	 76 54.75 79 62.67 53.33 62.33

Above table compares the techniques used by different authors for sentiment analysis and their corresponding accuracy:

V. Challenges in Sentiment Analysis [19,8]

Classification of text based on sentiments can be quiet challenging sometimes:

- **Using satirical words**: Satirical words are those positive words which are used to taunt or insult any entity. These words are used to mock any person/product. These comments can be misleading in sentiment classification as they contains positive words but infers a negative emotion.
- **Language Barriers**: Many of the times people use a different language other than English to express their sentiment or a language other than the one in which the classifier is trained. Also a person may use abbreviation such as F9, B4, OMG, TTYL etc. , which is unknown to the classifier.
- **False & Misleading Comments:** False reviews are often used in ecommerce websites as an immoral strategy to defame any product of rival establishment. These comments can lead to incorrect sentiment classification and wrong decision making.
- **Comparative Sentences:** Detecting opinion in a comparative sentence is again a difficult task.
  For e.g.**:** This product is better than the one launched last week.
  Analysis of such comments is difficult as we don't know whether the product launched last week was better than this product or not.
- **Feature extraction**: Feature extraction or selection is the most important step in sentiment classification. Selected feature of one product may not be appropriate for another product. For eg if we take a bigram feature "very long" in case of reviews which were for a laptop. If a comment includes "very long start-up time", it points toward a negative sentiment. Now if we extract reviews for a mobile phone and perform classification using the same extracted features and a comment includes "very long battery backup", it must points to a positive opinion but it will not do so because this feature i.e. "very long " is supposed to be a negative set feature .
- **Implicit Comments**: Implicit comments are the comments which are not mentioned directly, or indirect comments are called implicit comments. For e.g. "What a capacity!" comment on a mobile phone must be regarding the RAM of the device. Such comments face difficulty in classification because it doesn't have any keyword related to the RAM.
- **Handling of Neutral Comments:** Classification of neutral comments can also be difficult as they might be confused with objective sentences I.e. the sentences having facts & not any opinion

VI. Conclusion

Sentiment Analysis / Opinion mining is the one of the leading research areas in data science. OM is a way of retrieving opinions from large amount of data extracted from ecommerce & social media sites. This paper provides a reader with a brief review of sentiment analysis and techniques used for it. It also compares the accuracy of these techniques in some of the latest research papers.

Future work that can be done is the sentiment extraction from images & audio files which are very popular in social media

References

[1]Neha Raghuvanshi, J.M. Patil , "A Brief Review on Sentiment Analysis", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), India,2016,IEEE

[2] Jawad khan, Byeong soo jeong, "*Summarizing customer review based on product feature and opinion*" Proceedings of the 2016 International Conference on Machine Learning and Cybernetics, Jeju, South Korea, 10-13 July, 2016, IEEE

[3] Premnarayan Arya, Dr.Amit Bhagat, "*Deep Survey on Sentiment Analysis and Opinion Mining on Social Networking Sites and E-Commerce Website*" , International Journal of Engineering Science and Computing, March 2017 Volume 7 Issue No.3

[4] Harpreet Kaur, Veenu Mangat, *"A Survey of Sentiment Analysis techniques"* International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017)

[5] Ms.K.Mouthami ,Ms.K.Nirmala Devi, Dr.V.Murali Bhaskaran, "*Sentiment Analysis and Classification Based On Textual Reviews*" International Conference on Information Communication and Embedded Systems (ICICES) ,India, Pages 1-622, 2013

[6] Jawad khan, byeong soo jeong "*Summarizing customer review based on product feature and opinion*" Proceedings of the 2016 International Conference on Machine Learning and Cybernetics, Jeju, South Korea, 10-13 July, 2016

[7] Ms.A.M.Abirami , Ms.V.Gayathri "*A survey on sentiment analysis methods and approach*" 2016 IEEE Eighth International Conference on Advanced Computing (ICoAC)

[9] Gautami Tripathi, Naganna.S, "*Feature Selection and Classification approach for Sentiment Analysis*", An International Journal, vol.2, pp.1-16, (2015).

[10] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li, "*Dual Sentiment Analysis: Considering Two Sides of One Review*", IEEE Transactions on Knowledge and Data Engineering, MANUSCRIPT ID, 1041-4347 (c) 2015 IEEE.

[11] Onam Bharti,z Monika Malhotra ,"Sentiment analysis on twitter data", IJCSMC, Vol. 5, Issue. 6, June 2016, pg.601 – 609

[12] Akshay Amolik, Niketan Jivane, Mahavir Bhandari, Dr.M.Venkatesan," Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques.", International Journal of Engineering and Technology (IJET) Vol 7 No 6 Dec 2015-Jan 2016. Pg 2038-2044

[13] Jaspreet Singh , Gurvinder Singh and Rajinder Singh," Optimization of sentiment analysis using machine learning classifiers" *Human-centric Computing and Information Sciences* 2017, Springer. https://doi.org/10.1186/s13673-017-0116-3
[14] Raheesa Safrin, K.R.Sharmila, T.S.Shri Subangi, E.A.Vimal ," SENTIMENT ANALYSIS ON ONLINE PRODUCT REVIEW" Volume: 04 Issue: 04 | Apr -2017 IRJET Pg 2381-2388
[15] Naw Naw (2018); Twitter Sentiment Analysis Using Support Vector Machine and K-NN Classifiers; International Journal of Scientific and Research Publications (IJSRP) 8(10) (ISSN: 2250-3153), DOI: http://dx.doi.org/10.29322/IJSRP.8.10.2018.p8252
[16] H. M. Keerthi Kumar, B. S. Harish, H. K. Darshan. Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method, International Journal of Interactive Multimedia and Artificial Intelligence, (2018), http://dx.doi.org/10.9781/ijimai.2018.12.005
[17] Bharat Gaind, Varun Syal, Sneha Padgalwar "Emotion Detection and Analysis on Social Media",proceedings of International Conference on Recent Trends In Computational Engineering and Technologies (ICTRCET'18), May 17-18, 2018, Bengaluru, India.
[18] Ali Hasan , Sana Moin , Ahmad Karim and Shahaboddin Shamshirband," Machine Learning-Based Sentiment Analysis for Twitter Accounts "Mathematical Computer Application. 2018, 23, 11; doi:10.3390/mca23010011
[19] Siddu P. Algur, Jyoti G. Biradar, "Opinion Mining and Review Spam Detection: Issues and Challenges" IJARCSSE Volume 7, Issue 1, January 2017 DOI:10.23956/ijarcsse/V7I1/0170
[20] Ketan Sarvakar, Urvashi K Kuchara, "*Sentiment Analysis of movie reviews: A new feature-based sentiment classification*", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.3, pp.8-12, 2018
[21] Amit Palve, Rohini D.Sonawane, Amol D. Potgantwar, "*Sentiment Analysis of Twitter Streaming Data for Recommendation using, Apache Spark*", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.3, pp.99-103, 2017

Author's profile

Jasneet Kaur completed her Bachelor of technology from Guru Gobind Singh Indraprastha University New Delhi in 2014 and Masters of Technology from YMCA University Faridabad in 2018. She is currently working as an Assistant Professor in Department of Information Technology.