STM JOURNALS
Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 5, Issue 2
www.stmjournals.com

Sentiment Analysis and Text Summarization Review

Mubashir Farooq**, *Afaq Alam Khan
Department of Information Technology, School of Engineering and Technology, Central University of Kashmir, Srinagar, Jammu and Kashmir, India

Abstract

In this global era, it has been a significant task to find out what other people think of the information extraction behavior of an online user. People often express or share their moods, opinions, attitudes, ideas, on the social media sites and blogs. The platform of social media sites and blogs became the research base for analyzing the sentiments or opinions of the users. These sites became so popular that it became new opportunities and challenges for observing the opinions of other people. The sentiments are expressed in the form of statuses, tweets, reviews etc. Sentiment analysis is an area of computation in which opinions, sentiments, expressions, and emotions are analyzed. In this paper, our focus is on the sentiments expressed in online reviews. After sentiment analysis, it is mandatory to do a summarization of those reviews so that it will be helpful for a user to buy an online product and for better decision making. This paper discusses the most recent sentiment analysis and text summarization techniques and approaches that are being used.

Keywords: *Sentiment analysis, opinion mining, machine learning, text summarization, social media*

****Author for Correspondence*** E-mail: mubashirnaqash@gmail.com

INTRODUCTION

User reviews are essential when it comes to shopping on online shopping websites. The shopping vendors have made a facility to post reviews about the product which the customer had purchased. After posting a review, it becomes helpful for other customers who are trying to buy the product. Also, the reviews become helpful for the vendors to give the best to their customers. Now, with the rapid growth of online reviews, a large number of users also express their opinions, sentiments, and the mood in these reviews; thus need of sentiment analysis arises to extract relevant information from them. In the research, online reviews find their applications in the field of e-commerce, healthcare, financial risks etc. Although online reviews express numerous functions, sentiment analysis of online reviews is a challenging task.

Online reviews that are posted about products, movies and other things carry sentiments in them that need to be adequately analyzed. The sentiments are of two categories, i.e., positive and negative. However, often these reviews are being classified haphazardly. So, after sentiment analysis, classification is an essential part. Also, whenever a customer is interested in buying a product, he goes for reading reviews posted by other customers who have already purchased that product. The user cannot check thousands of reviews as it often wastes time and also leads to poor decision making. Even if the reviews are classified as positive and negative, still the customer cannot decide whether to buy the product or not. There is a need to do a summarization of the reviews so that it will be easy for a customer to go for summarized reviews rather than detailed reviews.

Sentiment analysis aims to extract the mood of the author on a particular subject matter from the given text. Sentiment analysis is also known to be as opinion analysis and subjectivity detection. The use of natural language processing and statistical or linguistic patterns is more to extract attitude of a person. Sentiment analysis categorizes the sentiments into three categories, i.e., positive, negative and neutral. Sentiment analysis often deals with polarity and subjectivity detection.

Sentiment analysis involves bringing out and identifying personal details from the natural language which is often inconsistent or unstructured. It is the pulling out structured information from unstructured or semi-structured data. Various features are being taken into consideration to do the sentiment analysis like POS tags, emotions, etc. We perform sentiment analysis at various levels like document level, sentence level, and aspect level.

Text summarization is the area of study in which significant content is being identified to produce a concise and accessible summary that should convey the essential description of the input. Text summarization takes place at two levels, i.e., extractive summarization and abstract summarization. There has been much research in the extractive summarization. Extractive summarization aims to identify the most important phrases and sentences in the input data to make a shorter versed summary. Abstractive summarization is the process to give a summary in different words than in the original text.

Information extraction has found its applications in the field of text mining due to the availability and generation of the massive amount of text information online. Text mining is the field of study that is used to obtain various similarities from the unstructured text data. In the information retrieval techniques, text summarization is one among them which is used to process unstructured text data into useful information. It reduces uniformed input text data into precise version while keeping its content of information and comprehensive connotation.

Sentiment analysis and text summarization can be merged to generate a more refined and helpful interface that can help in both analyses of sentiments and summarization. Sentiment analysis will help in the classification of reviews, while summarization will help to do summaries of the already classified categories so that it will help a customer in better decision making about a product and it also saves time to go for summarized reviews instead of elaborated reviews that can be thousands in number. Both can be combined to get the job done to make it easier for both online customer and the sellers.

RELATED WORK

Khan *et al.* present a hybrid approach for twitter tweets classification. The proposed system includes different preprocessing techniques before transferring input data to the classifier. The proposed work introduced new algorithm in which, first, data collection takes place, then different preprocessing steps are done, and finally, a three-tier classification is done. The results showed that the proposed method achieved greater accuracy and overcame the past limitations when compared to the identical techniques. The proposed method achieved an average accuracy of 85.7% with 85.3% precision and 82.2% recall. The work is done on six different data sets which consist of tweets [1].

Duwairi *et al*. proposed Arabic review sentiment analysis using machine learning techniques. The proposed work used already developed a dataset of tweets. Three supervised learning classifiers were run on the dataset. The three classifiers are: SVM, Naïve Bayes, and K-nearest neighbor. The performance evaluation shows SVM achieved the highest precision while KNN achieved the highest recall [2].

Joshi *et al.* proposed a practical approach for twitter sentiment analysis. The proposed approach used Gujarati tweets and classified them into two categories: Positive and Negative. The approach used various preprocessing techniques to improve accuracy and focused on feature extraction. The data was taken from twitter. They used five different machine learning techniques like Naïve Bayes, KNN, decision trees, SVM and multilayer perceptron in which SVM performed better than others. The features used here are POS tagging and SVM for classification. The proposed approach achieved 92% accuracy [3].

Kolog *et al.* used a machine learning technique to analyze sentiment and social influence in text-based student’s life stories. The life stories first dealt with preprocessing techniques, and they were clustered using K-means clustering approach with the help of Euclidean distance. The main influences on student’s academic development are identified as a peer, domestic and school staff. For supervised classification, labels are divided into classes in which various influences are used. Three decision tree classifier variants are used in this work, i.e., SMO, MNB and J48, in which, SMO is found to be better than MNB and J48 performed better in classification [4].

Alomari *et al.* presented Arabian Jordanian corpus of twitter with manual annotation of tweets as positive and negative. The supervised techniques SVM and Naïve Bayes were used in this work in which SVM performed better than Naïve Bayes. The SVM used stemmer with term frequency, inverse document frequency weighting scheme with bigrams achieved higher performance. The accuracy and f-score using SVM are 88.72 and 88.27% respectively [5].

Aggarwal *et al.* performed sentiment analysis on online mobile reviews by using modified K means clustering algorithm with Naïve Bayes and KNN classification. The method starts with data extraction and then different preprocessing steps are done and then given as input to the classifier. The proposed system achieved an overall accuracy of 91% on the test data of about 500 online mobile reviews [6].

Kulkarni *et al.* presented feature based sentiment analysis on a mobile phone. Different reviews are taken into consideration about a mobile phone from different online websites. Data extraction is done through web scraping. Different preprocessing techniques are used to make data consistent. The features used here are bag of words, word cloud, and Tf-idf. The system achieved an excellent accuracy of 84.61% and determined the opinion orientation [7].

Panigrahi *et al.* presented an approach for extracting app features that are present in user reviews and the opinions that are associated with them. The proposed method starts with the extraction of user reviews from Google play store. Preprocessing techniques are used to make data structured. The reviews are classified into three categories, i.e., positive, negative and neutral. After that, feature extraction is done like tokenization, sentence splitting, parts of speech tagging, named entity recognition and stop word removal. The proposed work helps app analysts and developers about user opinions about the features of the app so that to identify user requirements and plan to release future updates [8].

Rout *et al.* used both supervised and unsupervised techniques on different datasets. For automatic identification of sentiment for tweets extracted from microblogging website, unsupervised twitter approach is used. The algorithms used here are Maximum entropy, support vector machines, and Multi nominal Naïve Bayes. The proposed work achieved an accuracy of 80.68% using the approach of unsupervised techniques. The supervised techniques are used for combining unigram, bigram and parts of speech (POS) as a feature that helps in finding sentiment and emotion of unstructured data. Using unigram feature with multinominal naïve Bayes classifier for short message services, achieved overall accuracy of 67% [9].

Chakraborty *et al.* used the deep learning approach for sentiment analysis on a set of movie reviews. The proposed approach here deals sentiment analysis on movie review sets. Here the WordtoVecGoogle algorithm is tested on a dataset of movie reviews for classification of text to preserve the semantic relationships between terms. The dataset used has been collected from Kaggle using the bag of words model. Also, the proposed system presented the comparative performance of clustering algorithms [10].

Mudasir *et al.* proposed the task of automatic text summarization by developing a tool that summarizes the input text data. Various NLP features and machine learning techniques were used in this work. They have also presented how WordNet is used for abstractive summarization. The proposed approach first fetches sentences from the input text data by using the ranking algorithm. Based on various features, ranking is done, and after extracting sentences, some of the words and phrases are taken into consideration which are then transformed into simple ones to make the final summary using a hybrid technique of summarization. The algorithm works in three steps, i.e., preprocessing, feature extraction and post-processing [11].

Bhat *et al*. proposed the hybrid approach of both extractive and abstractive summarization. The proposed approach uses statistical and semantic relationships to generate the final summary. Emotion in the text is used as a semantic feature. The final extractive summary is given as input to the novel language generator which consists of Lesk algorithm, WordNet, and parts of speech tagger which helps in the transformation of extractive summary into the abstractive summary. The proposed approach is evaluated on DUC 2007 dataset and achieved better results in comparison to the MS Word [12].

Naidu *et al.* proposed an algorithm that fetches keywords for summarization of text automatically in Telegu e-newspaper datasets. The proposed approach used a combination of statistical method and machine learning method. A POS tagger in Telegu language is used to identify related parts of speech information in Telegu text data. Stop words are removed from the data. The proposed algorithm first fetches keywords by the automatic keyword extraction algorithm. At last, summarization algorithm selects sentences accordingly to form the final summary. The algorithm works on the single article at a time [13].

Song *et al*. proposed an LSTM-CNN based ATS framework that generates new sentences by investigating more close-grained snippets than sentences like semantic phrases. The algorithm consists of two main stages in which the first stage fetches phrases from input data and the second stage generates text summaries using deep learning. The performance evaluation on the CNN and Daily mail datasets shows the proposed work performs better than the state of the art techniques regarding both syntactic and semantic structure [14].

Chen *et al*. proposed an embedding method for unsupervised paragraph known as Essence vector model which extracts the essential information from a paragraph, but it also excludes the background information to generate a more new vector representation for the paragraph. They proposed an extension of essence vector model known as de-noising essence vector model. The D-EV model not only inherits features of EV model but also gives a more strong view for an input spoken paragraph. Also, a new summarization framework is proposed which can take both applicability and redundancy information into consideration simultaneously. The three approaches are evaluated on two summarization corpora. The evaluation outcome shows the effectiveness and relevance of the proposed framework as compared to several state-of-the-art summarization approaches [15].

Vo *et al*. proposed a method to fetch and summarize aspects and connected opinions of products from a vast domain of reviews in a specific domain using NLP tools. The proposed approach increased the accuracy and importance of the review summaries by influencing knowledge about aspects extraction of products. The evaluation of the proposed system achieved F1 scores of 0.774 for laptop and 0.714 for camera reviews [16].

Wan *et al*. proposed a technique for extracting and ranking of multiple summaries. The proposed framework first extracts multiple individual summaries for improving the quality of the summaries. Also, they proposed a method known as ensemble ranking method which is used for ranking candidate summaries with the help of bilingual features. ROUGE 1.5.5 toolkit is used for performance evaluation [17].

Hu *et al.* proposed a multi-text data summarization method for recognizing the top-k most factual sentences of hotel reviews. K-medoids clustering algorithm is used to divide sentences into k groups for identification. For summarization results, the methods from the groups were then taken into consideration. Two sets of reviews are collected from TripAdvisor.com. The framework involves five steps: hotel review gathering, review preprocessing, sentence significance computation, sentence similarity computation, and top-*k* sentence mentions. The outcome indicated that the sentences summarized using the method can provide comprehensive information in positive and negative aspects [18].

Roy *et al*. proposed a method that is used for classification of linguistic sentences using the statistical technique. The method does the summarization from the sentences which are categorized, ranking keywords by fuzzy scores. The algorithm consists of three steps: Matching keywords by scanning the whole text, checking repetition of keywords, and ranking each category that matches the result [19].

Sehgal *et al.* focused on unsupervised Text Rank algorithm which is a graph-based approach used to overcome the problem of automatic text summarization and proposed a method to assign scores to the sentences during extraction. The method also calculated the normalized score of each sentence. For evaluation, four sample articles' results were taken and calculated using ROUGE 2.0 toolkit based on several parameters [20].

TOOLS AND TECHNIQUES

Table 1 given below summarizes some of the prior works done in the field of sentiment analysis:

Table 1: *Prior Work Done in Sentiment Analysis.*

Paper by	Year	Dataset Used	Approach Used	Performance/Evaluation
Khan *et al*. [1]	2013	Twitter	A hybrid scheme using an enhanced form of emoticon analysis	Accuracy: 85.7%, Precision: 85.3%, Recall: 82.2%
Duwairi *et al*. [2]	2016	Tweets/Comments	Supervised Learning Algorithms	Accuracy: 69.97%
Joshi *et al.* [3]	2017	Gujarati Tweets	Supervised Learning Algorithms	Accuracy: 92%
Panigrahi *et al.* [8]	2017	Google Play store Reviews	Extracting app features mentioned in user reviews and their associated sentiments	The result is for judging Android App, and the developers are also able to predict the problem and the improvement needed in the app for its popularity within less time.
Alomari *et al.* [5]	2017	Arabic Tweets	Supervised Learning Algorithms	Accuracy: 88.72%, F-score: 88.27%
Aggarwal *et al*. [6]	2017	Mobile Reviews	Supervised Learning Algorithms	Accuracy: 91%
Rout *et al.* [9]	2017	Movie Reviews/Tweets	Supervised Learning Algorithms	Precision: 72.53% Recall: 78.21% F1 score: 75.26%
Kolog *et al*. [4]	2018	Students Life Stories	Supervised Learning Algorithms	SMO (80%) J48 (72%) MNB (69%)
Kulkarni *et al.* [7]	2018	Product Reviews	Supervised Learning Algorithms	Accuracy: 84.61%
Chakraborty *et al.* [10]	2018	Movie Reviews	Deep Learning	No. of clusters:10 Time required for clustering: 477 s

Table 2 given below summarizes some of the prior works done in the field of text summarization.

Table 2: *Prior Work Done in Text Summarization.*

Paper by	Year	Data Used	Approach Used
Mudasir *et al*. [11]	2016	Documents data	NLP and Machine learning techniques
Hu *et al.* [18]	2016	Online Hotel Reviews	Novel text summarization technique to determine the top- *k* most informative sentences from online hotel reviews
Roy *et al*. [19]	2017	Linguistic Text	Classification of Linguistic Sentences using statistical approach and ranking keywords by different fuzzy value as scores
Song *et al.* [14]	2018	Daily mail and CNN	Deep Learning
Chen *et al.* [15]	2018	DUC 2001, 2002, 2004	Novel unsupervised paragraph embedding method
Vo *et al*. [16]	2018	Camera and Laptop Reviews	Aspect relationships using the DP, CR, and NER NLP tools
Wan *et al*. [17]	2018	DUC 2001, 2002	New framework for addressing the cross-language document Summarization task and top-*K* ensemble ranking method
Bhat *et al*. [12]	2018	Documents Data	Hybrid approach to a single-document summarization
Naidu *et al.* [13]	2018	e-newspaper articles	Interdependent algorithms in keyword extraction and text summarization
Sehgal *et al.* [20]	2018	Articles	Text Rank Algorithm

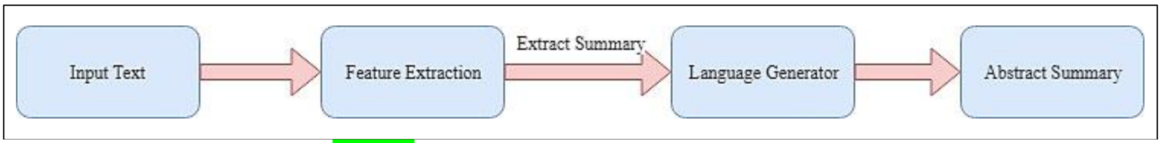

Fig. 1: *Text Summarization Process [12].*

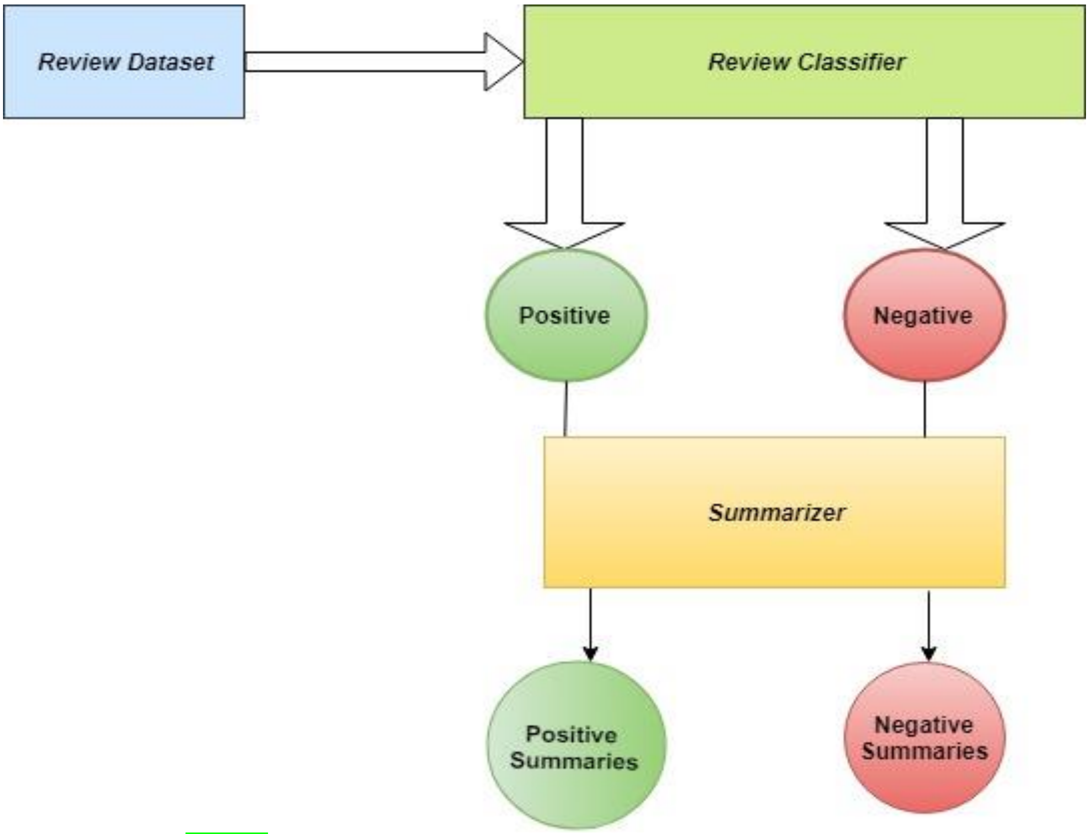

Fig. 2: *Sentiment Analysis and Text Summarization [1] [12].*

CONCLUSIONS

Upon reviewing the previous works done in the field of sentiment analysis and text summarization using different approaches and techniques, we conclude that much work has been done in both the fields.

So, if we combine both fields to develop the hybrid techniques; that will be useful for everyone; it will make the task more comfortable and will save time.

ACKNOWLEDGEMENTS

This research has been made possible by my guide, Mr. Afaq Alam Khan, who was very generous and supportive throughout the study.

REFERENCES

1. Farhan Hassan Khan, Saba Bashir, Usman Qamar. TOM: Twitter Opinion Mining Framework Using Hybrid Classification Scheme. *DSS (Decis Support Syst)*. 2014; 57: 245–257p.
2. Duwairi Rehab M, Islam Qarqaz. Arabic Sentiment Analysis Using Supervised Classification. *International Conference on the Future Internet of Things and Cloud*, Barcelona, Spain, IEEE. 27–29 Aug 2014; (8): 579–583p.
3. Joshi Vrunda C, Vekariya Vipul M. An Approach to Sentiment Analysis on Gujarati Tweets. *ACST*. 2017; 10(5): 1487–1493p.
4. Emmanuel Awuni Kolog, *et al.* Using Machine Learning for Sentiment and Social Influence Analysis in Text. *Proceedings of the International Conference on Information Technology & Systems*. 2017; 721: 453–463p.
5. Khaled Mohammad Alomari, *et al*. Arabic Tweets Sentimental Analysis Using Machine Learning. In: Benferhat S, Tabia

K, Ali M, editors. *IEA/AIE* 2017. *Lecture Notes in Computer Science (LNCS)*. Cham: Springer; 2017; 10350: 602–610p.
6. Ruchika Aggarwal, Latika Gupta. A Hybrid Approach for Sentiment Analysis Using Classification Algorithm. *IJCSMC*. 2017; 6(6):149–157p.
7. Sanket Kulkarni, Neha Prabhune, Vasanti Sathe. Sentiment Analysis of Product Reviews. *IJCSIT*. 2018; 9(1): 19–22p.
8. Sangeeta Panigrahi, *et al*. Sentiment Analysis of Application Reviews on Google Playstore. *IJSRCSEIT*. 2017; 2(2): 510–513p.
9. Jitendra Kumar Rout, *et al*. A Model for Sentiment and Emotion Analysis of Unstructured Social Media Text. *Electron Commer Res*. 2018; 18: 181–199p.
10. Koyel Chakraborty, *et al*. Comparative Sentiment Analysis on a Set of Movie Reviews Using Deep Learning Approach. In: Hassanien A, Tolba M, Elhoseny M, *et al*., editors. *The International Conference on Advanced Machine Learning Technologies and Applications, (AMLTA2018)*. Cham: Springer; 2018; 723: 311–318p.
11. Mudasir Mohd, *et al*. Sumdoc: A Unified Approach for Automatic Text Summarization. In: Pant M, Deep K, Bansal J, *et al*., editors. *Proceedings of Fifth International Conference on Soft Computing for Problem Solving*. Singapore: *AISC*, Springer; 2016; 436: 333–343p.
12. Bhat IK, Mudasir Mohd, Rana Hashmy. Sum It Up: A Hybrid Single-Document Text Summarizer. In: Pant M, Ray K, Sharma T, *et al*., editors. *Soft Computing: Theories and Applications*. Singapore: *AISC*, Springer; 2018; 583: 619–634p.
13. Reddy Naidu, *et al*. Text Summarization with Automatic Keyword Extraction in Telugu e-Newspapers. In: Satapathy SC, *et al*., editors. *Smart Computing and Informatics, Smart Innovation, Systems and Technologies*. 2017; 77: 555–564p.
14. Shengli Song, Haitao Huang, Tongxiao Ruan. Abstractive Text Summarization Using LSTM-CNN Based Deep Learning. *MTA*. 2018; 1–19p.
15. Kuan-Yu Chen. An Information Distillation Framework for Extractive Summarization. In *IEEE/ACM Trans Audio, Speech, Lang Process*. Jan 2018; 26(1): 161–170p.
16. Vo AD, Nguyen QP, Ock CY. Opinion-Aspect Relations in Cognizing Customer Feelings via Reviews. In *IEEE Access*. 2018; 6: 5415–5426p.
17. Xiaojun Wan, *et al*. Cross-Language Document Summarization via Extraction and Ranking of Multiple Summaries. *Knowl Inf Syst*. 2018; 1–19p.
18. Ya-Han Hu, Yen-Liang Chen, Hui-Ling Chou. Opinion Mining from Online Hotel Reviews: A Text Summarization Approach. *IPM*. 2017; 53(2): 436–449p.
19. Nirmalya Roy, *et al*. Extraction and Classification of Linguistic Text for Text Summarization. *IJEECS*. Oct 2017; 6(10): 314–318p.
20. Sehgal S, Kumar B, Maheshwar, *et al*. A Modification to Graph Based Approach for Extraction Based Automatic Text Summarization. In: Saeed K, Chaki N, Pati B, *et al*., editors. *Progress in Advanced Computing and Intelligent Engineering*. Singapore: AISC, Springer; 2018; 564: 373–378p.

Cite this Article
Mubashir Farooq, Afaq Alam Khan. Sentiment Analysis and Text Summarization Review. *Journal of Artificial Intelligence Research & Advances*. 2018; 5(2): 1–7p.