International Journal of Modern Electronics and Communication Engineering (IJMECE) ISSN: 2321-2152
Volume No.-7, Issue No.-2, March, 2019

# Review on Significance of Text Pre-processing Techniques in Sentiment Analysis and Opinion Mining

Kulpreet Kaur
Department of Computer Science
SGGSWU
Fatehgarh Sahib, India
E-mail: jhajjkulpreet@gmail.com

Shruti Aggarwal
Department of Computer Science
SGGSWU
Fatehgarh Sahib, India
E-mail: Shruti_cse@sggswu.org

**Abstract:** Data Mining is a non-trivial process of analyzing large databases or data warehouses to extract knowledge or useful patterns. Sentiment analysis is an area in data mining that uses 'Web' as a data source. It is also known as Opinion Mining and is a very important and widely studied topic in Web content mining and Natural language processing. It is a process of extracting sentiments or emotions from a database of reviews or comments. It is generally seen as a classification task that groups the reviews according to their polarity but it can actually accomplish tasks such as summarization, spam content detection, product recommendation, buying behaviour recognition and objectionable content detection on social media also. Mostly data collected from web is in unstructured and noisy form so pre-processing of web content is a crucial step and increases the accuracy of classifier in sentiment analysis but it is often ignored. Keeping in mind the importance of this step, major pre-processing techniques are reviewed and various combinations are analyzed for their advantages and disadvantages. This paper presents an overview of sentiment analysis and different pre-processing techniques used in it.

**Keywords:** Web content mining, Sentiment analysis, Opinion Mining, Text pre-processing, Lexicon based method, Machine learning methods.

## I. INTRODUCTION

Data Mining is a process of extracting knowledge from raw data. Raw data here can be relational databases, log files, transactional databases, Web, surveillance records, geographical data or other types of multimedia data. Fig. 1 shows various types of data in data mining.

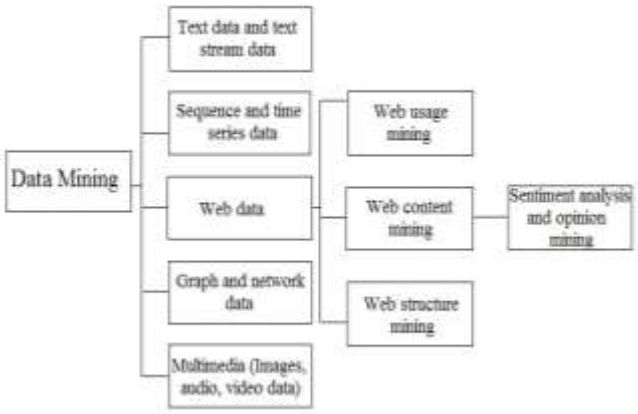


Figure 1 Sentiment analysis in Data Mining

The web is a huge repository of data of various kinds. The main components of web mining are web content mining, web structure mining and web usage mining while sentiment analysis comes under web content mining. Sentiment Analysis (SA), also called Opinion Mining (OM) is the field of study that analyses people's opinions, sentiments, evaluations, attitudes and emotions toward entities such as products, services, organizations, individuals, issues, events, topics or their attributes [1]. SA uses natural language processing, machine learning, text analysis, statistical and linguistics knowledge to analyze, identify and extract information from documents [2].

In the era of computerization, web 2.0 is growing significantly. Web 2.0 includes dynamic applications and websites like blogs, social networking sites (Facebook, Twitter, Instagram), media sharing platforms (YouTube, Flickr), information providing sites (Wikipedia) along with online shopping sites (Amazon, Flipkart).This extensive information explosion suggests that customers often check other users' suggestions in online shopping sites, blogs, and reviews on social networking sites before buying a product or getting a service. According to the Oxford dictionary, SA is the process of computationally identifying and categorizing opinions expressed in a piece of text to determine whether the writer's attitude toward a particular topic, product, and so on is generally positive, negative, or neutral [3].

Basically, there are three major categories to divide the polarity of the sentence or review but in some research works more than three categories are also introduced.

Pre-processing in SA is the task of refining the content collected from online sites or other online sources for specific applications. E.g. To calculate the polarity of a sentence or a document, stop words, factual sentences need to be removed and aspects or features need to be collected. This refinement is generally done in pre-processing step. Pre-processing decreases the overhead of the classifier by decreasing number of words or aspects to be classified.

In SA, various pre-processing techniques can be applied on data according to the nature or domain of data. Although it is believed that there is no proper subset of pre-processing techniques that increases accuracy in every dataset and every classification algorithm. So, different techniques or combinations of techniques are applied on different datasets.

### A. Applications of SA and OM

SA has various applications in computer science ranging from business development, marketing to online shopping. The data

acquired from online sources is used in automated or semi-automated manner to extract the sentiments from it. The various applications of SA and OM are as follows:

1. Classification of reviews according to their polarity where manual classification is difficult [4].
2. Analyzing users' buying behavior and product recommendations [4].
3. Analyzing hospitality in tourism sites [5].
4. Policy making and electoral campaigns [6].
5. Analyzing stock market trends and social trends from social networking sites [7].
6. Checking the trustworthiness of reviews or comments [8].

### B. Process of Sentiment Analysis

In general, SA can be considered as a classification process. SA can be done at three levels of classification i.e. document level, sentence level and aspect level. At document level, SA classifies the whole document as positive, neutral or negative. Considering document as a unit for sentiment classification is generally not very advantageous when the dataset is small and sentiment of each sentence is varying. At sentence level, opinion or sentiment of each sentence is considered as a unit. The classification is carried out in two steps. First step excludes factual sentences that do not express any sentiment from the dataset. At next step, subjective sentences are differentiated as positive, negative and neutral. At aspect level, classification is done on the basis of particular words present in the sentence. Various classifiers are used that determine the polarity of sentence by the presence of particular words. Some other fields in SA include Emotion Detection, Building Resources and Transfer learning [9]. Resources include lexical resources that are used as a dictionary in the task of SA. These resources have pre assigned polarity for many words along with synonyms and antonyms. Transfer learning is a typical machine domain that stores knowledge acquired while solving a problem to apply it on solving another related problem. The basic process of SA is given below that comprises of basic three steps:

- Pre-processing of the dataset
- Feature selection or feature acquisition
- Polarity classification

The final results give the sentiment polarity of each word or sentence that can be used for summarization or classification of reviews on a specified scale.

Pre-processing is the first step in text classification, after collection of data. Pre-processing is the procedure of cleansing and preparing texts that are going to be classified [10]. Right combination of pre-processing techniques can efficiently improve the accuracy of classification task. Tokenization is the first and common step in pre-processing techniques. Tokenization is a task of separating the full text string into a list of separate words [11]. The various Pre-processing techniques that can be applied on review data are given in fig. 2.Techniques like URL removal, Unicode removal, punctuation removal, removing numbers and stop word removal decrease the noise in review data. Lemmatizing and stemming are used to bring each word to its basic form so that classification can be done easily. Negation handling, POS tagging, slang word replacement increase the classification accuracy significantly. Spelling correction brings the misspelled words, internet shorthand words and slang words to their actual representation that can be efficiently matched with the dictionary terms.

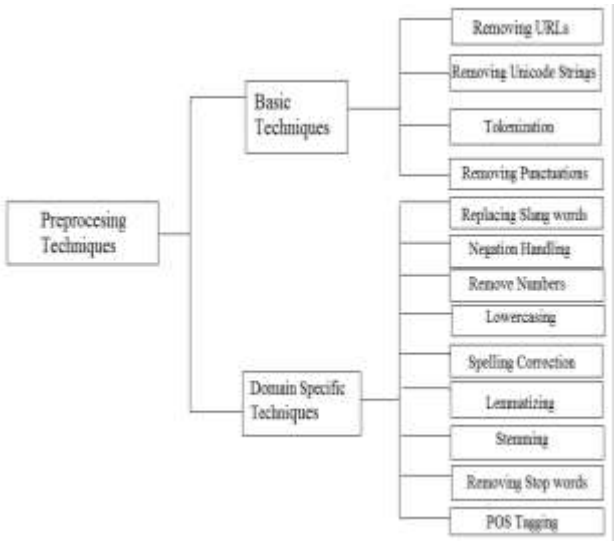


Figure 2 Pre-processing techniques in sentiment analysis

After pre-processing next step is classification of reviews according to their polarity. Sentiment classification can be divided into machine learning techniques, semi supervised techniques and lexicon based techniques. Each technique has various algorithms associated with it. Lexicon based techniques are unsupervised learning techniques and there is no need of training data in advance. This technique can be divided into two basic approaches i.e. manual approach and dictionary based approach. The manual approach is a time consuming approach as lexicon needs to be coded by hand. The other technique i.e. dictionary based approach generally deals with already available lexical resources like Wordnet, Sentiwordnet, Vader, Sentistrength etc. To increase the efficiency of these lexical resources, synonyms and antonyms are utilized for polarity assignment [12].

Machine learning is a supervised approach that provides a solution to classification problem in two steps i.e. learning the model from a corpus of training data then classifying the unseen data based on trained model [13]. Mostly used machine learning algorithms are Naïve Bayes classification, Support vector machine, artificial neural networks, random forest, genetic algorithms, decision trees, ensemble vote methods and decision trees [14].Some semi-supervised techniques are also proposed by various researchers that include the benefits of both machine learning and lexicon based approaches.

## II. RELATED WORKS

This chapter includes various research studies related to the Sentiment analysis and opinion mining using different methods. Review classification, summarization, spam detection are various research directions that attract the researchers in this field. The previous research related to pre-processing techniques and sentiment analysis is as follows:

Symeon Symeonidis, et al. [10] gathered common pre-processing techniques from previous studies, added some new ones and evaluated them on two datasets using four machine learning algorithms. An ablation study was performed in order to determine high-performance techniques and their interactions. A winning set of techniques was concluded that performed best among all.

V. V. Nhlabano, et al. [15] considered text pre-processing as a crucial step in challenging text classification tasks. Further, the effect of stemming, stop word removal and feature selection on social media data classification was discussed. The applied techniques improved predictive accuracy as well as decreased the dimensionality of the classification model.

Walla Mehdat et al. [9] gave comprehensive overview of recently proposed SA algorithms and applications of SA. Feature selection methods and classification are explained briefly. The related topics like transfer learning, emotion detection and building resources are discussed. It is concluded that Naïve Bayes and SVM as most used Machine learning algorithms and WordNet as most used Lexicon in SA.

Shreyas Wankhede, et al. [16] analyzed the preprocessing usage on social media data. The methodology introduced pre-processing techniques like URL and hash tag removal, spell checking and correction before applying the classification method. N-gram algorithm was used for spell checking that automatically generated candidates for spell correction. The study concluded that accuracy of classifier can be increased significantly by applying pre-processing.

Tajinder Singh, et al. [17] evaluated the effect of pre-processing on twitter data especially in terms of slang words. N-gram method was used to bind slang words with coexisting words and impact of these words on sentiment of the tweet. Experiments indicate the improvement in accuracy of classification using these pre-processing techniques.

Giulio Angiani, et al. [18] aimed to highlight the importance of pre-processing techniques and show how they can improve system accuracy. This method applies each of most known filters along with the basic cleaner. Some of the techniques removed useless noise in raw data, while others increased relevance of some concepts. This research was conducted on data originated from twitter.

Erik Cambria et al. [19] discussed natural language processing problems that need to be solved for human like performance of sentiment analysis system. The problems were organized into three layers i.e. semantic layer, syntactic layer and pragmatic layer. This study emphasized on major NLP problems including classification task and paved a path to ensemble approach that used both data driven algorithms and theory driven methods to understand and implement the way human decode and understand natural language.

Nilesh M. Shelke et al. [20] presented the categorization of features and feature selection techniques and gave a simple framework of text classification using lexicon based approach. Noun phrases and verb phrases were recognized in the sentences and then score was calculated using SentiWordNet as a lexical resource.

Emma Haddi, et al. [21] explored the role of text pre-processing in sentiment analysis and demonstrated that appropriate feature selection and representation improve accuracy in support vector machine classifier. This paper investigates the sentiment of online movie review dataset and used various combinations to reduce noise in data. The results showed that appropriate text pre-processing methods including data transformation and filtering can significantly increase the classifier's performance.

Atanu Dey et al. [22] considered that the lexicon based approaches outperform learning based ones when training data is inadequate but the existing lexicons use unigrams with assigned sentiment scores. N-grams can be formed when unigrams are combined with negations or intensifiers. So, the study presented a methodology to create n-gram lexicon using a rule based approach.

Akrivi Krouska et al. [23] analyzed the effect of pre-processing on twitter dataset. This research used unigrams, bigrams and 1-3 grams for data representation and applied TF-IDF, stemming, stop word removal and tokenization on each dataset. The effect of pre-processing on the quality of classification process was investigated and it was concluded that such techniques can help in personalization of e-learning systems for students.

D. S. Kulkarni et al. [24] presented the study on recent opinion mining methods and its comparative analysis with aim to identify the current research problems and their significance. It was concluded that there are three main components of opinion mining and analysis methods namely pre-processing, feature extraction and classification. The outcome of the study was research gaps in present solutions.

B. Dalila et al. [25] considered aspect extraction as most vital and extensively explored phase of opinion mining to carry out the classification of sentiments in precise manner. The study gave a comprehensive analysis for different aspect extraction techniques and also highlighted major recent works in Arabic language. Aspect extraction performance of supervised, unsupervised and semi-supervised in terms of precision, recall and F-measure was measured and compared and strengths of presented approaches were illustrated.

Shahid Shayaa et al. [14] presented systematic literature review to discuss both technical and non-technical aspects of opinion mining and sentiment analysis (OMSA) and highlighted challenges in OMSA techniques. Various techniques

of classification i.e. keyword based classification, lexicon based classification and machine learning based classification, different datasets on which these techniques were applied were discussed.

## III. COMPARATIVE ANALYSIS OF PRE-PROCESSING TECHNIQUES

Various techniques of pre-processing can be used alone or in combination with other techniques in sentiment analysis and opinion mining. Comparative analysis of various pre-processing techniques is given below:

| S No | Technique / Combination | Domain | Advantages | Disadvantages |
|---|---|---|---|---|
| **1** | Replacement of URLs and User mentions | Twitter Dataset | This method is beneficial in case of twitter dataset and increases the efficiency by approx. 0.2 % if applied alone. | It is domain specific pre-processing method and not useful in dataset other than social media. |
| **2** | Slang word removal and abbreviations expansion | Social media or online product reviews | Replacement of slang words using bigram, trigram causes better classification of sentiments and classifier accuracy is improved. | Certain classifiers like Convolutional Neural Networks (CNN) on certain datasets can reduce the accuracy when this technique is applied. |
| **3** | Removing numbers | Any | Removal of numbers is generally considered efficient as it is believed that numbers do not contain any sentiment. | In some slang words and internet shorthand such as "gr8", numbers are important and considering them may improve classifier efficiency. |
| **4** | Removing negation and replacement with antonyms | Any | Negation words generally affect the meaning of the sentence and ignoring negation can lead to misclassification. | Some researchers remove words less than 3 characters and hence removal of negation like no and not leads to misclassification. |
| **5** | Lowercasing | Any | It is observed that lowercasing increase the accuracy in most of domains. | Lowercasing is not necessary in some languages. |
| **6** | Stemming | Any | Stemming allows using and considering nouns, verbs and adverbs which have same radix in same way and improves performance of classifier. | Over stemming is a main problem with this technique that changes the meaning of the word. |
| **7** | Lemmatization | Any | Lemmatization improves the performance of the model and has better results than stemming. | In some classifiers like CNN, lemmatization does not work very well. Lemmatization can sometimes be a slow process as technique is time consuming. |
| **8** | Stop word removal | Any | Removing stop words reduces dimensionality of the term space and time complexity of classification is decreased. | The accuracy of output of stop word removal depends upon the proper training in machine learning. |
| **9** | Stemming and stop word removal | Document Classification | The precision of the classifier is highest when both | With the combination of stop word removal and not stem- |

| | | | | |
|---|---|---|---|---|
| | | | techniques are applied collectively. | ming applied, precision is lowest. |
| **10** | Tokenization , Stemming , lowercasing Stop word removal | English email and news dataset | Maximum F1 measure is obtained when using all methods except stop word removal. | Status of other tasks than lowercase conversion varies depending on nature. |
| **11** | Tokenization, Stemming and stop word removal | Any | This combination removes noisy data from text data and reduces overall size of dataset. | Sometimes over stemming may decrease the understandability of words. |
| **12** | Noise reduction, Feature extraction and stemming | Spam test dataset | Precision is improved and false positives were reduced significantly. | Precision in case of certain dataset is lightly reduced by using this combination. |
| **13** | Stemming, stop word removal and feature selection | Movie reviews | Removing stop words and reduced no. of features can improve up to 20% of accuracy | This technique can have different results on other datasets. |
| **14** | Stemming, punctuation removal and stop word removal | Trip advisor dataset | Precision of random forest classifier increases after removing punctuation marks. | Precision of RF decreases with applying all pre-processing steps together. |
| **15** | URL replacement, POS tagging, hash tag removal , spell correction and using emoticon | Any | Accuracy using spell correction, exploiting emoticons can be maximized. | Accuracy using spell correction, dictionary look up approach can be minimized among the combinations. |
| **16** | URL replacement, punctuation removal, stop word removal and slang replacement | Twitter dataset | Use of bigram and trigram method increase accuracy and also decrease the size of data. | This combination works well with social media datasets but effect on other datasets is unknown. |
| **17** | URL replacement, Hash tag removal, stemming, stop word removal, negation handling, dictionary lookup | SemEval dataset | Basic cleaning i.e. stemming, stop word removal, tokenization strongly increases accuracy. | Using dictionary look up did not enhance the performance of test but increase the elaboration time for cleaning raw data. |
| **18** | Slang word replacement, spell checker and stemming | Log files | The process of stemming produces great result in terms of accuracy. | The process of stemming can increase false positives in result. |

Table 1 Pre-processing Techniques in sentiment analysis

The given table analyses major pre-processing techniques and combinations along with their advantages and disadvantages.

## IV. CONCLUSION

Sentiment analysis is a widely studied topic in various fields like computer science, business, marketing and many more. As a result there are more than 1 million research papers available that contain the term "sentiment analysis"[4]. When we discuss in context of computer science sentiment analysis is still a growing field. As sentiment analysis is an NLP problem, therefore, there is always a need of perfection in this study to achieve human like imperceptibility by machines and to perform classification and other tasks. As online platforms have data in crude form, text pre-processing is a crucial step in this process as it reduces the overhead and time complexity of sentiment analysis process. Various techniques have their own advantages and disadvantages but their usage varies. So there cannot be a particular subset of techniques applicable on every domain. So, combinations of various techniques can be analyzed for best results.

## ACKNOWLEDGMENT

I (Kulpreet Kaur) would like to express my gratitude towards my guide Ms. Shruti Aggarwal, Assistant Professor of depart-

ment of Computer Science for her constant encouragement and her valuable suggestions.

## REFERENCES

[1] U. Students, T. Selvind, and A. Professor, "Automatic Sentiment Analysis of User Reviews", IEEE international Conference on Technological innovations in ICT for Agricultural and rural development, 2016.

[2] P. Gupta, R. Tiwari, and N. Robert, "Sentiment Analysis and Text Summarization of Online Reviews: A Survey", International conference on Communication and signal processing, 2016.

[3] Valdivia, M. V. Luzón, and F. Herrera, "Sentiment Analysis in TripAdvisor," IEEE Intell. Syst.,vol 32, issue 4, pp. 72-77, 2017.

[4] I.K.C.U. Parera,"Aspect based opinion mining on Restaurant reviews",IEEE International Conference on Computational intelligence and Applications,2017

[5] N. Akhtar, N. Zubair, A. Kumar, and T. Ahmad, "Aspect based Sentiment Oriented Summarization of Hotel Reviews," in Procedia Computer Science, vol 115, pp. 563-571,2017.

[6] R. Jose and V. S. Chooralil, "Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble Approach," in Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016.

[7] K. Rudra, A. Sharma, N. Ganguly, and S. Ghosh, "Characterizing and Countering Communal Microblogs during Disaster Events," IEEE Trans. Comput. Soc. Syst., vol 5, issue 2, 2018.

[8] S. Bajaj, N. Garg, and S. K. Singh, "A Novel User-based Spam Review Detection," in Procedia Computer Science, vol 122, pp. 1009-1015, 2017.

[9] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, issue 4, p.p. 1093-1113,2014.

[10] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," Expert Syst. Appl., vol 110, pp. 298-310, 2018.

[11] Balazs, J. A., & Velásquez, J. D "Opinion mining and information fusion: A survey", Information Fusion, 27, 95–110, .2016.

[12] N. N. Yusof, A. Mohamed, and S. Abdul-Rahman, "Reviewing classification approaches in sentiment analysis," in Communications in Computer and Information Science, vol 545, pp. 43-53, 2015.

[13] J. Khairnar and M. Kinikar, "Machine Learning Algorithms for Opinion Mining and Sentiment Classification," Int. J. Sci. Res. Publ., vol 3, issue 6, 2013.

[14] S. Shayaa et al., "Sentiment analysis of big data: Methods, applications, and open challenges," IEEE Access, 2018.

[15] V Nhlabano and P. E. N. Lutu, "Impact of Text Pre-processing on the Performance of Sentiment Analysis Models for Social Media Data", International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD),2018 .

[16] S. Wankhede, Ranjit Patil ,Sagar ,ProfAshwini, "Data Pre-processing for Efficient Sentimental Analysis",Proceedings of 2nd International conference on Inventive communication and computer technologies,2018.

[17] T. Singh and M. Kumari, "Role of Text Pre-processing in Twitter Sentiment Analysis," in Procedia Computer Science, Vol. 89, P.p. 549-554,2016.

[18] G. Angiani, L. Ferrari, T. Fontanini, P. Fornacciari, E. Iotti, F. Magliani, and S. Manicardi, "A comparison between Pre-processing techniques for sentiment analysis in Twitter," in CEUR Workshop Proceedings, 2016.

[19] E. Cambria, S. Poria, A. Gelbukh, I. P. Nacional, and M. Thelwall, "Sentiment Analysis Is a Big Suitcase", IEEE Intelligent Systems ,vol 32 , issue 6, 2017.

[20] N. M. Shelke, S. Deshpande, and V. Thakare, "Statistical feature based approach for aspect oriented sentiment analysis," in Proceedings of the International Conference on Inventive Communication and Computational Technologies(ICICCT) , 2017.

[21] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," Procedia Comput. Sci., vol 17, pp. 26–32, 2013.

[22] Dey, M. Jenamani, and J. J. Thakkar, "Senti-N-Gram: An n-gram lexicon for sentiment analysis," Expert Syst. Appl., vol 103, pp. 92-105,2018.

[23] Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," in IISA 2016 - 7th International Conference on Information, Intelligence, Systems and Applications, 2016.

[24] D. S. Kulkarni and S. F. Rodd, "Extensive Study of Text Based Methods for Opinion Mining",2nd International conference on Inventive Systems and control,2018 .

[25] A. Dalila, A. Mohamed, and H. Bendjanna, "A Review of Recent Aspect Extraction Techniques for Opinion Mining Systems", 2nd International Conference on Natural Language and Speech Processing (ICNLSP), 2018.