International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

Pre-Processing Concepts and Techniques for Sentiment Analysis

M. Edison*[1], Dr. A. Aloysius[2]
[1]Research Scholar, Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, TamilNadu, India
[2]Assistant Professor, Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, TamilNadu, India

ABSTRACT
Sentiment Analysis is consider as a big task to analyse people's opinion, appraisal, and attitudes in the worldly communications. Many of the people can express their emotions with the text, symbols, and variety of ambiguous data through social media networks. Mainly, Twitter permits a 140-character limit to post one's comments. Therefore, users are posting their comments like ambiguous data. In that case, pre-processing techniques are very helpful to remove the unwanted data from the data set and solve the various research problems in sentiment analysis for supporting the same. This paper mainly deals with the importance of pre-processing concepts and techniques. Especially, pre-processing techniques are given an idea that cautious to select the suitable feature to analyse the sentiments, which gives better result to classify the sentiment words.
Keywords: Pre-Processing, Pre-Processing Tasks, Techniques, Sentiment Analysis, Feature Selection.

I. INTRODUCTION

In this decade, the fast growth of information are huge through internet. Nowadays, the internet users can share their opinions through Social Media Networks (SMNs) with different topics. The bearing of SMNs like Twitter, Facebook, WhatsApp etc., have increased data for daily usage. Especially, 500 million tweets per day and around 200 billion tweets per year tweets tweeted on Twitter, 2.5 billion pieces of content and 500+ terabytes of data each day, which contains unstructured data. The unstructured data contains Text, URLs, user name, numbers, symbols, special characters and so on. Therefore, the unwanted data to be cleaned from the data set. Especially, pre-processing is a vital task in text mining and it reduces complexity of the data. Various steps are covered in pre-processing such as: data cleaning, data reduction, data transformation, data integration and data discretization. Some of these techniques assisted to pre-process the data based on the research issues, then it produces the pre-processed data. This paper structure as follows. In section 2 is presented as literature review and the pre-processing tasks are presented in section 3. In section 4 is represented as the major pre-processing techniques and the conclusion is represented in section 5.

II. LITERATURE REVIEW

Vijayarani et. al [1] have given a quite information about the text mining pre-processing techniques. Mainly, the technique has helped to extract the data from the large dataset and it uses to remove the stop words and handling the stemming. Muskan et. al [2] have proposed pre-processing methods for bindings of slang words as well as coexisting words. It sees the effort of pre-processing on the Twitter data for sentiment classification and relies the sentiment translation of the slang words. Akrivi et. al [3] have selected the appropriate feature for pre-processing techniques without affect the quality of classification in sentiment analysis. Regarding the feature selection policy focused on the choice to select the algorithm to measure the features like attribute. Akila et. al [4] have proposed well-organized method for pre-processing of tweets, which is important pace in classification. It performes in three different task by the pre-processing technic such as: remove the URLs, remove the stop word and finally tokenize the sentences then the pre-processed data have been given an input to the machine learning algorithm for splitting the tweets.

III. PRE-PROCESSING TASKS

"Pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method for resolving such issues and prepares raw data for further processing [5]." The data pre-processing task is shown in figure. 1.

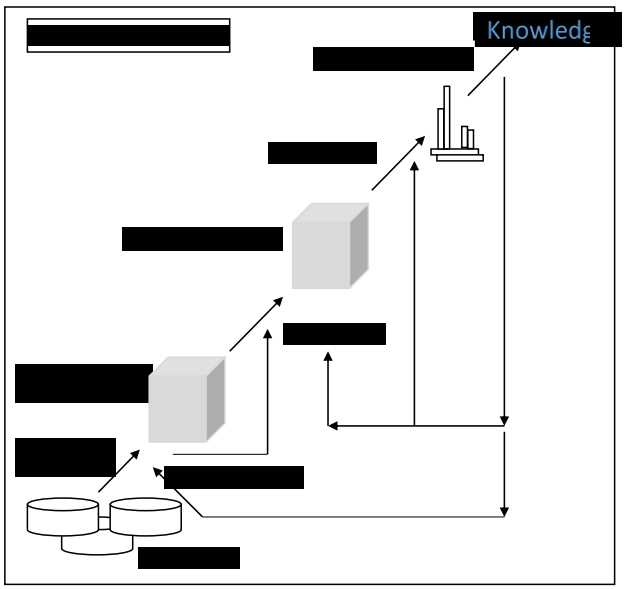

Figure. 1. Data Pre-Processing Tasks

A. Data Cleaning

Data cleaning is the process of removing the erroneous word from the data set. It fills in missing values, and removes the noisy data. In the case of noisy data occurrence in the data set, scattered errors. For example, "string is a numeric value", actually string is sequence of characters but not related to numeric value. So, the text "numeric value" is considered as noisy data in the data set [6, 7].

a. Stop Word Removal

Stop words commonly occur in the languages like English, Hindi, and Sanskrit etc. In Natural Language Processing (NLP), stop words often take place that are not considered as important in the data / data set. Practically, pre-processing techniques and applications are important to remove the stop words in text mining [8]. With different categories, the stop words can be spilt into determiners, coordinating conjunctions and prepositions [9].

i. Determiners

Determiner incline to mark nouns (Examples: the, a, an, another).

ii. Coordinating Conjunction

Coordinating conjunctions connect words, phrases, and clauses (examples: for, an, nor, but, or, yet, so).

iii. Prepositions

Prepositions express temporal or spatial relations (Examples: in, under, towards, before).

B. Data Transformation

Data transformation is one of the tasks involved in pre-processing which applies for data analysis purpose. Data transformation represents some operations such as normalization, aggregation, generalization and attribute construction which are additional pre-processing procedures that would contribute towards the success of the mining process.

a. Normalization

Normalization is a measurement of the data mining, which has to be analyzed with a specific range like [0.0, 1.0] for providing better results in data analysis. Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers): V'=(V-Mean) / StDev

Where,

V = vector

Mean = Means

StDev = Standard Deviation

b. Aggregation

Aggregation would be useful for data analysis to obtain aggregate information that gathered and expressed in summary form, and it obtains information about the specific variable or group.

Moving up in the concept hierarchy on numeric attributes.

c. Generalization

Generalization is the process of creating successive layers of summary data in the evolution database.

Moving up in the concept hierarchy on nominal attributes.

d. Attribute Construction

Replacing or adding new attributes are inferred by existing attributes.

C. Data Reduction

- Data reduction is a process for transformation of the numerical or alphanumeric information, which derives empirically into a corrected and simplified form.
- Reduce the amounts of data.
- Reducing the number of attributes.
- Removing irrelevant attributes

D. Data Discretization

Data discretization techniques can be used to reduce the number of values for a given continuous attributes by dividing the range of the attributes into intervals. Interval labels can be used to replace the actual data values. Replacing numerous values of a continuous attribute by a small number of interval labels reduces and simplifies the original data [10].

IV. PRE-PROCESSING TECHNIQUES

Pre-processing is a necessary data preparation step for sentiment classification. The pre-processing filter allows the following configurations:

A. TF-IDF weighting scheme

It is a standard approach for the feature vector construction. TF-IDF stands for the "Term Frequency-Inverse Document Frequency", which is numerical statistic that reflects how important a word to a document in a corpus.

B. Stemming

The stem word plays a vital role in sentiment analysis because the users can express their opinion in an improper way. The comments have been written in a noisy form like "happyyyyyyy", here more than one repeated character does not focus on any kind of information as well as a word in structured format. So, the word "happyyyyy" is converted into "happy" and the repeated characters have been removed. Hence, these words are handled in a proper manner.

The pre-processing algorithm has removed unwanted data from the data set using unigram feature by sentence level. The unigram feature checks for a word sequentially, then, it gives the original data. The original data have been taken to measure the polarity of different classes [11].

C. Tokenization

This setting splits the documents into words/terms, constructing a word vector, known as bag-of-words. The N-Gram Tokenizer is used to tokenize the words then compare word with unigram, bigram and 1-to-3-gram.

D. Feature selection

Feature Selection (FS) is called as variable selection or attribute selection. FS is the method for selecting and constructing the features of particular field in the data set. Concurrently, the FS is consists of many features like unigram, bigram, trigram N-gram etc., these features are supportive to sentiment analysis for selecting a feature of a particular sentiment.

V. CONCLUSION

Pre-processing is one of the major tasks in data analysis, which is an important and critical stepping process in sentiment analysis. Mainly, the pre-processing concepts like tasks are given many idea for preparing the data. Data pre-processing is a process to reduce the complexity of the data as well as it removes the unwanted data from the data set. Especially, data cleaning, data reduction and pre-processing techniques are very helpful to clean the data and select the particular feature and what actually very suitable techniques for the particular issue as given a better suggestion for the sentiment analysis.

VI. REFERENCES

[1] Vijayarani, Ilamathi and Nithya "Preprocessing Techniques for Text Mining – An Overview", International Journal of Computer Science & Communication Networks (IJCSCN), 2015, pp: 7-16.

[2] Muskan and Dr. Knawal Garg "An Efficient Algorithm for Data Cleaning of Web Logs with spider Navigation Removal", International Journal of Computer Application (IJCA), 2016, pp: 6-12.

[3] Akrivi Krouska, Christos Troussas and Maria Virvou "The effect of preprocessing techniques on Twitter Sentiment Analysis", Information Intelligence Systems & Applications (IISA), 7th International Conference on. IEEE, 2016, pp: 1-5.

[4] R. Akila, R. Praveena and PriyaDarsini "Twitter Data Preprocessing Using Natural Language Processing",

South Asian Journal of Engineering and Technology (SAJET), 2017, pp: 46-49.
[5] https://www.techopedia.com/definition/14650/data-pre-processing.
[6] http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-3.html.
[7] L. Sunitha, M. Bal Raju and B.Sunil Srinivas "A Comparative Study between Noisy Data and Outlier Data in Data Mining", International Journal of Current Engineering and Technology (IJCET), 2013, pp: 575-577.
[8] Jaideepsinh K. Raulji, Jatinderkumar R. Saini and Dr. Babasaheb Ambedkar "Stop-Word Removal Algorithm and its Implementation for Sanskrit Language", International Journal of Computer Applications, 2016, pp: 15-17.
[9] http://text-analytics101.rxnlp.com/2014/10/all-about-stop-words-for-text-mining.html.
[10] S.S. Baskar , Dr. L. Arockiam and S.Charles "A Systematic Approach on Data Pre-processing In Data Mining", COMPUSOFT, An international journal of advanced computer technology, 2013, pp: 335-339.
[11] M. Edison and A. Aloysius 'Lexicon based Acronyms and Emoticons Classification of Sentiment Analysis (SA) on Big Data", International Journal of Database Theory and Application (IJDTA), 2017, pp: 41-54.

BIBLIOGRAPHY

M. Edison is a research scholar in the department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He is doing his Doctor of Philosophy in the area of Big Data. He has published many research articles in the International Journals/Conferences in his research area, he has attended many workshops, conferences, and he has acted as a resource person for national and international workshops.

Dr. A. Aloysius is working as an Assistant Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 17 years of experience in teaching and research. He has published many research articles in the National / International conferences and journals. He has acted as a chairperson for many national and international conferences. Currently, eight candidates are pursuing Doctor of Philosophy Programmed under his guidance.