BioGecko Vol 12 Issue 01 2023

ISSN NO: 2230-5807

COMPARITIVE ANALYSIS OF PRE-PROCESSING TECHNIQUES FOR TEXT DATA

1st M.Priya Alagu Dharshini 2nd Dr.S.Antelin Vijila
1.Rearch scholar,ManonmaniamSundaranar University Tirunelveli, India.
2. Assistant professor, ManonmaniamSundaranar University Tirunelveli, India.
**preethi.murugiah10@gmil.com*

Abstract
Due to the advent of the internet, there has been a tremendous expansion in digital assessments made by e-commerce and social media websites such as Amazon, Flipkart, Facebook, Twitter and others. Users share their views, feelings in an exceedingly convenient approach in these applications. Sentiment analysis is an intellectualistic technique of extracting user feelings and emotions through the Natural Language Process. The sentiment analysis encompasses the classification of input text into three parts "Positive," "Negative" or "Neutral." The motivation of sentiment analysis is to distinguish the data polarity within the text and classify them. In most cases, sentiment analysis is a text-based process, but there are challenges in finding the correct polarity of the sentence. Thus, the text cleansing method is necessary for converting raw text to cleansed text. Pre-processing is the initial phase in text classification and choosing effective pre-processing procedures can improve classification accuracy. This work utilizes five kinds of pre-processing methodology and their different combinations to gauge the effectiveness based on time complexity and text cleansing quality. The pre-processing strategies used in this work are Spelling Correction, Part-of-Speech Tagging, Lemmatization, Emoji Removal and Replace Emoticons. For sentiment analysis, 4680 records data set have been collected from Amazon and Twitter which contains 3816 from Amazon 864 from Twitter. In order to measure the efficiency deep learning-based Convolutional Neural Network classification algorithm is applied to weigh the score of cleaned text to classify the sentiment analysis. Among all the pre-processing strategies the Replace Emoticons techniques achieve a classification accuracy score 0.84 on the Amazon dataset and 0.81 on the Twitter Dataset. This result emphasizes that replace emoticons can be used in sentiment classifiers to achieve better performance.

Keywords— pre-processing, deep learning, accuracy, CNN

1. INTRODUCTION

Nowadays, the advanced digitalization of social media and e-commerce websites is expanded for digital marketing. According to digital marketing, clients purchase their products through e-commerce websites [1]. It also gives the insight to share their reviews and comments about the quality of the products. Likewise for data-gathering, these reviews will become a source of information for the new clients, product manufacturers or sellers. This will lead to know about the product quality, which will assist them to make the proper choice to buy, produce or market the products. Reviews are identified through comments, but this constitutes a huge amount of raw text data and henceforth it needs the proper formation to acquire the significant data. Pre-processing is the main source in providing cleansed text data for better sentiment classification results [2], so this work briefly discusses pre-processing techniques.

Selecting the most appropriate techniques for doing pre-processing can improve the efficiency in classifying text.Reviews from the internet have lots of irrelevant data, for example Instagram, Facebook comments and Snap deal reviews include large quantities of irrelevant data. During the cleaning and preparing texts for sentiment classification, proper pre-processing methods should be used. Comparing the pre-processing techniques based on processing time could be inadequate in determining which pre-

processing technique is the best. Confusion matrix is one of the best metrics in estimating the efficiency of algorithms. Therefore, this analysis makes use of accuracy score and efficiency along with processing time. The Big data concept is utilized to manage vast amounts of data effectively. Following sections give the theoretical underpinnings of these concept and its potential with real-time scenarios.

2. Related works

In Sentiment Analysis, especially on microblogging texts, the role of preprocessing techniques is significant as a part of text classification. Many research efforts have been made to demonstrate the difference between these techniques and their contribution to the final result of classification.

The implementation of the pre-processing on twitter knowledge for sentiment polarity categorization is investigated by Singh and Kumari et al. [3] utilize the users' viewpoints on Twitter for specific themes like products, books, politics, movies, etc. The authors used the N-gram text normalization technique to predict the bindings and the significance of slang words identified using conditional random fields. These user reviews had punctuation, spelling shortcuts, new words, misspellings, URLs, slang, abbreviations and genre-specific terminology. Parts Of Speech tagging is used for removing stop words. Support Vector Machine (SVM) is a method that evaluates the data and determines the orders applied for the classification process. When three models are compared, POS tagging with N-gram normalization had the highest accuracy followed by N-gram normalization without normalization and this study reviews classification that are grouped into three classifications: positive, negative and neutral. Additionally, the proposed scheme is robust to different sizes of data and more accurate at classifying sentiment.

Haddi, Liu and Shi et al. [4] analysed the ways for doing sentiment analysis. This describes different pre-processing (white space removal, stop words removal, expanding abbreviation, stemming, online text cleaning) employed to filter the given piece of natural language text. Three distinct feature matrices are produced using various feature weighting approaches like Term Frequency Inverse Document Frequency (TF-IDF), Feature frequency (FF) and Feature presence (FP). Filtering is the next step done to choose the most feature by calculating chi-squared statistics for each feature in the document based on the feature weighting approaches. The sentiment polarity categorization of tweets is used in machine learning experiments for evaluating the product's attributes. SVMclassification of tweets determines whether a common perception is positive or negative. This presented approach evaluates the classifier performances based on the classification of each feature matrix (FF, TF-IDF and FP) achieving accuracy of 90.5%, 93.5% and 93% respectively.

Pre-processing methods were investigated by Uysal and Günal et al. [5] for the languages Turkish and English on E-Mail and news datasets. They intend to widely inspect the effect of pre-processing on text classification for the fluctuated perspectives like text language, classification accuracy, dimension reduction and text space. In this research, pre-processing methods' combinations were evaluated under lowercase conversion, stop-word removal, stemming and tokenization. Chi-square (CHI2) is applied in this work to choose informative features. SVM classifier is utilised for classification algorithm. Also, the success measure is considered on the basis of Micro-F1 score. These research achieved the highest Micro-F1 score of 0.9713 on the Turkish email dataset.

Martin Boldt et al. [6] presented a sentiment analysis of e-mail by integrating VADER lexical knowledge with the SVM classification of text. It uses a combined structure in which one can make use of the background lexical information. The word-class relationships wereutilized and it extracts the information for particular domains by employing convenient training examples. An SVM model is trained on the labelled data after labelling the e-mails with a lexicon-based sentiment tool (VADER). TF-IDF method separates e-mails into individual terms and weighs their value in a document in relation to the entire data set. E-mails were graded according to their sentiment, from positive to negative. Experimental results show that the sentiment is extracted using Linear SVM model with a mean AUC of 0.896 and a mean F1-score of 0.834.

The study proposed by Desheng Run Wu et al. [7] focuses on an examination of online assessment posts accessible to a business sector by utilizing a totally distinctive assumption strategy. A subset of postings can be manually polarized in terms of sentiment. Word segment, linguistics processing and

part-of-speech tagging techniques were used for pre-processing purposes and N-gram was chosen for feature selection to select efficient features. Automatic prediction of the sentiment polarity was done by identifying the features using sentiment analysis and labelled posts from written stock forum text of other posts. The performance comparison of different methods indicates that the statistical machine learning approach has 81.82% accuracy, whereas that of semantic approach is 75.58%.

Mukherjee et al. [8] tried to implement the streaming data classification in HBase using the Naive Bayes Algorithm. Classifying the streaming data is the most challenging task, even though HDFS (Hadoop Distributed File System) does not show an efficient output when small data blocks are handled in real-time. Authors used HBase database as it is a good option to handle huge amounts of data because of its properties like distributed, scalable and storing data as key-value pairs. The tobacco-affected student record (sample size ranges from 3000 to 40,000 records) from HBase database was handled and classified using the Naive Bayes algorithm, which offers more accuracy than a decision tree. Execution time and the classification error rate are compared among the naive bayes and decision tree.

Tyagi et al. [9] evaluated the word score for sentiment analysis using a Logistic Regression (LR) algorithm. The effective word score approach is utilized to filter the training dataset after pre-processing approach. In this work polarity score of each word is maintained in the range of -5 to 5. The dictionary has approximately 2500 words, which is sufficient to predict sentiment. They compared the navie bayes, SVM and LR methods and found that the LR method was the most efficient.

Ramadhan et al. [10] used Twitter API to collect the data about the candidates' names who participated in the Jakarta Governor Election. Initially, they have done the pre-processing step like deletion of punctuation, URLs, stop words and stemming. After that the dataset is split into two where 90 percent is used for training and 10 percent is used for testing. Authors performed a feature extraction process using TF-IDF. The process of extracting features includes grouping of words with the bag of words method and Multinomial Logistic Regression (MLR) is used for classification. The authors used cross-validation with k-fold (where k=10) for testing the results. From the test results, the greater value of the fold affects the acquisition accuracy of each method. MLR can achieve the highest accuracy up to 74%.

3. Methodology

This section describes the pre-processing techniques and deep learning algorithm (CNN), here the algorithm is used to weigh the pre-processing techniques.Overall analysis of this work is depicted in Fig 1.1

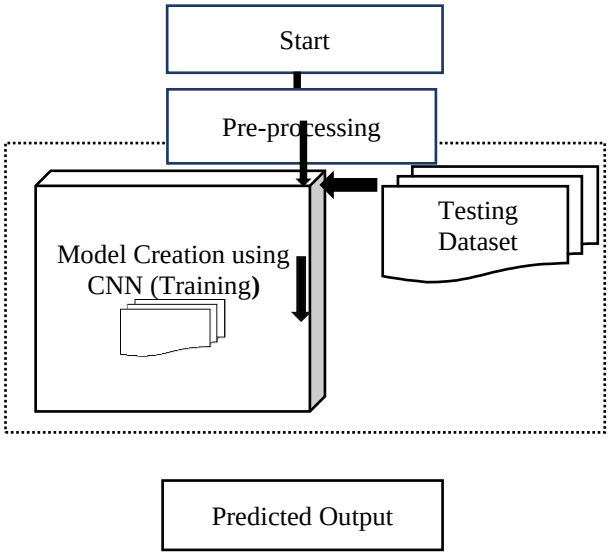

Figure 1.1 Flow diagram of CNN Sentiment Analysis

3.1 Data Pre-Processing

Pre-processing of text data makes it easier to extract useful data from the raw text,which will be given to deep learning algorithms. This is considered as one of the benefits of mining raw text. Data driven from any e-commerce sites are generally incomplete, inconsistent and are prone to errors, so it requires pre-processing which will cleans and convert information into a noticeable format.

In this paperseventeen pre-processing techniques are analyzed in order to find which can be used as the standard pre-processing technique based on accuracy score. The techniques are converting to lower characters, removing URLs, removing hash (#) tags, removing @ symbol,remove all non-English words, remove stop words, expand acronyms,remove special characters, remove recurring character, removenon-alphabetic letter start with a word, stemming, tokenization, spelling correction, POS Tagging, lemmatization, remove emoticons andreplace emoticons.

3.1.1 Basic Pre-Processing Techniques

From the seventeen pre-processing techniques, twelve methods are the most fundamental approaches and these are known as the basic pre-processing techniques.In these twelve methods, converting the text to its lower-case format is relatively easy and faster to process the text and removing URL's, #tags, the @ symbol, special characters, recurring characters, non-English words, stop words and words with non-alphabetic letters helps in removing non-relevant data.The non-relevant data are treated as noise as they do not provide any useful information during text classifications. Expanding acronyms is (like awe will be converted to awesome) helpful in finding the exact word. Stemming process finds the base form or stem of the word and tokenization techniquebreaks large text strings into tokens. It is possible to tokenize large text sections into sentences and sentences into words.

3.1.2Basicwith Spelling Correction (BSC)

In BSC, spelling correction is used along with twelve basic pre-processing techniques. spelling correction is a special pre-processing technique that detects and corrects spelling errors.Spell-Checker library is used in correcting the text. The spelling correction approach reduces the background noise.

3.1.3 BasicwithPOS TaggingandLemmatization (BPTL)

BPTL uses POS Tagging and lemmatization technique along with twelve basic pre-processing techniques.Initially basic pre-processing isapplied, followed by POS Tagging and lemmatization. A POS Tagging is a Natural Language Processing (NLP) task where a special label is assigned to each token (word) in a text. Eight parts of speech are there in the English language: conjunction, noun, verb, pronoun, adverb, interjection, adjective and preposition. These denote word functioning grammatically within the sentence.

Lemmatizationconverts a wordto its base form but it differ fromstemming in the context. The last few characters are removed in stemming leads to spelling errors and incorrect meanings. For example, Stemming will cut ‘ing’ part from caring and convert it to car whereas lemmatization identifies correctly as 'care'. Thus, lemmatization gives better meaning whereas stemming gives meaningless words. Multiple lemmas may also present for the same word where tag of 'part-of-speech' (POS) in that specific context for the word will be identified and the suitable lemma will be extracted.

3.1.4 Basic withEmoji Removal (BER)

Emoji Removal is a pre-processing technique that checks the corpus for the availability of emojis. When emoji is present in the sentence, it will be removed and white space is also removed using this technique.

3.1.5 Basicwith Replace Emoticons (BRE)

In social media people express their emotions using emojis. In the BRE technique, twelve basic pre-processing techniques are applied initially and then emojis are replaced with words using the emoji library. For example, a positive emoticon can be replaced by a word like "happy" to represent positive emotion. Likewise, negative emoticons can be replaced by a word like "sad" to represent negative emotion.

3.2Sentiment Analysis

In a lot of NLP tasks, the sentiment prediction of text is significant. In recent years, deep learning algorithms have significantly improved sentiment prediction [11-12]. CNN is a powerful tool which can be used toanalyse sentiment in prediction work[13-14]. In sentiment classification, the data set has divided into testing and training data.

3.2.1 Word Embedding

Anunsupervisedmodel like CNN creates a vocabulary of words from a large corpus of words and generates dense embeddings for each word. For deep learning algorithms to operate effectively on numbers instead of words, the words are turned into vectors. This transformation is referred to as word embedding.

3.2.2 Overviewof CNN

In neural network the algorithm cannot process the raw text data, the pre-processed text used Word2Vec model for converting the text to vector format and it is sequentially presented in embedding layer. The given sentence is converted to numerical sequence which is fed as the input to the embedding layer. The length of this input sequence is a 600-dimensional vector. The embedding layer is usually the first layer in a convolutional neural network which processes text input for classification tasks.An embedding layer maps numeric word indexes to corresponding 300-dimensional word vectors which creates an embedding matrix. The embedding layer produces embedded matrix. The data from the embedding layer will be given to the two convolution layers.Due to the nature of text data, a 1D convolution layer is used in the model. Each convolution layer uses two different filter sizes which capture contextual information unique to that layer.Each convolutional layer is mentioned by the letters C1 and C2 and its kernel size of the filters are 3, 5, which K1 and K2 of respectively denotes. Both C1 and C2 have default stride value. The convolutional layer gets a new matrix with convolved features by performing a convolution operation.After convolution, each feature should have a nonlinearity applied to it. Selecting Rectified Linear activation (ReLU) is simple which is quite effective and widely used because of its nonlinearity. The result of the activation function should then be passed to the global max pooling layer for processing. This layer is accountable for reducing the dimension and abstraction of the features by combining the feature maps. Thus, the computation speed is increased. In the second convolutional layer, it is used to extract features. A pooling function is applied to each feature map for producing a fixed-length vector. The output of the pooling function is concatenated into a single output known as C3. The resulting feature matrix is passed into the dropout layer. Over-fitting is not experienced during the CNN training process. But if it occurs, a dropout layer can resolve it. For converting to the dimensional feature map, flattening processes with dense layer. In this layer, each neuron receives input from all the neurons of the previous layer, making it a densely connected neural network. This feature matrix is fed into another dropout layer. No values will be dropped during the inference when the dropout layer and training are enabled simultaneously. Finally, the feature matrix is passed to the output layer. The last layer in a neural network is usually the classical dense or fully connected layer. The number of units in this layer equals the number of classes in the dataset. Some loss functions, including categorical cross-entropy and error prediction are included in the output layer. The fully connected layer automatically stores the weight of different text categorized under positive, negative or neutral classes.The sigmoid activation function is applied to all the units for getting final output class probabilities. Output is derived from the fully connected layer that can either be positive, negative or neutral.

3.3Data Analysis

The pre-processing results are evaluated using two datasets containing 3816 textual reviews and 864 tweets which are given to the deep learning-based sentiment prediction systems. The pre-processing experiments are performed on a Dell Inspiron 5770 with an Intel Core i7 processor having 1.8 GHz. The experiments are done in Windows 10 operating system having 16 GB of RAM.

4Results and Discussion

The performance analysis is done by implementing the methods in two datasets, Amazon and Twitter. The analysis was done among the methods Basic Techniques, Basic with spelling correction (BSC),

Basic with POS tagging and lemmatization (BPTL), Basic with emoji removal(BER), Basic with replace emoticons (BRE), BER-PTL, BRE-PTL, BER-PTL-SC, and BRE-PTL-SC. Table 4.1 shows performance analysis in terms of accuracy, specificity, sensitivity, precision, F1-score and recall for all pre-processing techniques in the Amazon dataset.

Table 4.1 Performance analysis on Amazon dataset

Method	Processing Time (Seconds)	Accuracy	Specificity	Sensitivity	Precision	F1 score
Basic Techniques	0.65	0.82	0.89	0.77	0.82	0.79
BSC-Basic with Spelling Correction	140.67	0.76	0.86	0.69	0.76	0.71
BPTL-Basic with POS Tagging and Lemmatization	13.94	0.80	0.88	0.73	0.78	0.75
BER-Basic with Emoji Removal	1.84	0.83	0.90	0.80	0.82	0.81
BRE-Basic with Replace Emoticons	1.33	0.84	0.90	0.80	0.86	0.82
BER-PTL	12.30	0.81	0.89	0.80	0.84	0.81
BRE-PTL	11.25	0.83	0.90	0.81	0.81	0.81
BER-PTL-SC	280.63	0.77	0.88	0.78	0.77	0.77
BRE-PTL-SC	121.25	0.79	0.89	0.78	0.80	0.79

Table 4.2 Performance analysis on Twitter dataset

Method	Processing Time (Seconds)	Accuracy	Specificity	Sensitivity	Precision	F1 score
Basic Techniques	0.81	0.79	0.88	0.73	0.79	0.74
BSC-Basic with Spelling Correction	560.66	0.75	0.86	0.73	0.79	0.74
BPTL-Basic with POS Tagging and Lemmatization	28.46	0.76	0.88	0.76	0.77	0.76
BER-Basic with Emoji Removal	1.52	0.78	0.88	0.77	0.79	0.78
BRE-Basic with Replace Emoticons	0.92	0.81	0.90	0.80	0.84	0.81
BER-PTL	10.31	0.76	0.88	0.76	0.77	0.76
BRE-PTL	11.16	0.78	0.87	0.73	0.77	0.75

BER-PTL-SC	623.79	0.75	0.87	0.75	0.77	0.75
BRE-PTL-SC	145.10	0.77	0.88	0.78	0.77	0.77

The above tables 4.1 and 4.2 show the performance metrics and time analysis for amazon and twitter dataset for comparing the nine different combinations of pre-processing techniques. This helps to find which technique gives better result for text classification. Effective text pre-processing is required to generate accurate sentiment analysis. Generally, people provide the reviews in short length, abbreviations form and use emoji in their reviews which makes it necessary to pre-process the text efficiently. In this study spelling correction and pos-tagging and lemmatization metrics values are very low and also it is time consuming while processing the text data. Basic with replace emoji performance better in cleansing process and it shows a better result in performancemetrics and also it takes very less time to pre-process the text when compared with spelling correction, pos-tagging and lemmatization techniques. Hence in this study it proved that spelling correction, pos-tagging and lemmatization is not providing a prominent result. Based on the analyzation BREtechniques performance well for text classification.

5. Conclusion

The pre-processing technique is evaluated with Amazon and Twitter datasets. This result emphasizes the significance of pre-processing techniques in sentiment analysis which is shown using a convolutional neural network classifier. The various combination of pre-processing techniques is used along with basic pre-processing techniques to improve the accuracy of three-way sentiment prediction.Results are evident that classification of sentiment polarity using basic with emoji replace pre-processing technique (BRE) achieves a higher classification of accuracy nearly 0.84% on Amazon dataset and 0.81% on the Twitter datasets. Future approach is to test these techniques on datasets from different domains such as news articlesand movie reviews.

REFERENCES

1. José M. Ponzoa&Anett Erdmann (2021): E-Commerce Customer Attraction: Digital Marketing Techniques, Evolution and Dynamics across Firms, Journal of Promotion Management.
2. Liu, Bing, 2012, ‘Sentiment Analysis and Opinion Mining. Synthesis Lectures onHuman Language Technologies’, 5(1), 1–167.
3. Singh, T., & Kumari, M., 2016, ‘Role of Text Pre-processing in Twitter Sentiment Analysis’, Procedia Computer Science, 89, 549–554.
4. Haddi, E., Liu, X., & Shi, Y., 2013, ‘The Role of Text Pre-processing in Sentiment Analysis’. Procedia Computer Science, 17, 26–32.
5. Uysal, A. K., &Gunal, S., 2014, ‘The impact of pre-processing on text classification’, Information Processing & Management, 50(1), 104–112.
6. Borg, A., & Boldt, M., 2020, ‘Using VADER Sentiment and SVM for Predicting Customer Response Sentiment’, Expert Systems with Applications, 113746. (Journal)
7. Wu, D. D., Zheng, L., & Olson, D. L. (2014). A Decision Support Approach for Online Stock Forum Sentiment Analysis. IEEE Transactions on Systems, Man and Cybernetics: Systems, 44(8), 1077–1087. doi:10.1109/tsmc.2013.2295353. (Transaction)
8. Mukherjee A., Mondal S., Chaki N., Khatua S., 2019, ‘Naive Bayes and Decision Tree Classifier for Streaming Data Using HBase’. Advanced Computing and Systems for Security. Advances in Intelligent Systems and Computing, vol 897. Springer, Singapore.
9. Tyagi, A., & Sharma, N., 2018, ‘Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic’. International Journal of Engineering & Technology (Journal)
10. Ramadhan, W. P., Novianty, S. T. M. T. A., &Setianingsih, S. T. M. T. C., 2017, ‘Sentiment analysis using multinomial logistic regression’, 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC). (Conference)

11. Lei Zhang, Shuai Wang, Bing Liu, 2018 'Deep learning for sentiment analysis: A survey', Published by Wiley data mining and knowledge discovery.
12. Qurat,Mubashi, Amna R, Amna N, Muhammad K, Babar H and A.Rehman,',2017 Sentiment analysis using deep learning techniques: A review' International Journal of Advanced Computer Science and Applications.
13. Hannah Kim, Young-SeobJeong, 2019' Sentiment classification using convolution neural networks',Department of Future Convergence Technology, Soonchunhyang University, Asan-si 31538, Korea.
14. Haitao Wang, Keke Tian1, Zhengjiang Wu1, Lei Wang. 2020, 'A Short Text Classification Method Based on Convolutional Neural Network and Semantic Extension', International Journal of Computational Intelligence Systems, Vol. 14(1), 2021, *pp.* 367–375.