IJMS 2022 9(2): 69 - 81

International Journal of Multidisciplinary Studies (IJMS)
Volume 9, Issue II, 2022

Learning to Use Normalization Techniques for Preprocessing and Classification of Text Documents

Karunarathna K.M.G.S. and Rupasingha R.A.H.M.*

Department of Economics and Statistics, Faculty of Social Sciences and Languages,
Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka

ABSTRACT

Text classification is the most substantial area in natural language processing. In this task, the text document is divided into various types according to the researcher's purpose. In the text classification process, the basic phase is text preprocessing. In text preprocessing, cleaning, and preparing text data are significant tasks. To accomplish these tasks under the text preprocessing, normalization techniques play a major role. Different kinds of normalization techniques are available. In this research, we mainly focus on different normalization techniques and the way of applying them to text preprocessing. Normalization techniques reduce the words of the text files and change the word form to another form. It helps to analyze the unstructured texts and predefine the text into standard form. This causes to improve the efficiency and performance of the text classification process. For text classification, it is important to extract the most reliable and relevant words of the text files, because feature extraction causes successful classification. This study includes the lowercasing, tokenization, stop word removal, and lemmatization as normalization techniques. 200 text documents from two different domains, namely, formal news articles and informal letters obtained from the Internet in the English language were evaluated using these normalization techniques. The experimental results show the effectiveness of the use of normalization techniques for the preprocessing and classification of text documents and for comparison between before and after using normalization techniques to the text files. Based on the comparison, we identified that these normalization techniques help to clean and prepare text data for effective and accurate text document classification.

KEYWORDS: *Preprocessing, Normalization, Techniques, Cleaning documents, Text classification*

Corresponding author: Rupasingha R.A.H.M., Email: hmrupasingha@gmail.com

1 INTRODUCTION

This research paper deals with the normalization techniques used for data preprocessing. Preprocessing is the first step of text classification. In our research, we mainly target the normalization techniques phase.

1.1 Text Classification

The development of internet and digital technology has given rise to diverse research, and text classification is one of them. This is due to the fact that researchers are increasingly interested in natural language processing and text mining. Text classification is playing a huge role when searching, classifying, and organizing information in various types. Using a search engine like Yahoo, Google, Ask-researchers, one can find a lot of documents according to their purpose. Therefore, text classification has become one of the most valuable fields today. There are two types of text classification; they are manual classification and automatic classification. Manual classification can be inconsistent because researchers must have knowledge about the categories that may cause to represent the negative impact of the classification. Automatic classification based on the machine learning algorithm can be complex, but it verifies the consistency of the classification (Malik & Bhardwaj 2011).

Text classification is the act of labeling or tagging documents using categories depending on their content (Pascual & F 2021). Text classification can be used successfully for different domains such as topic detection, spam e-mail filtering, SMS spam filtering, sentiment analysis, author identification, and web page classification (Basarkar 2017) (Adetunji et al. 2018). As shown in Fig. 1, there are several processes to complete in the text classification such as preprocessing text file, feature extraction, feature selection, and classification stage (Uysal & Gunal 2014)

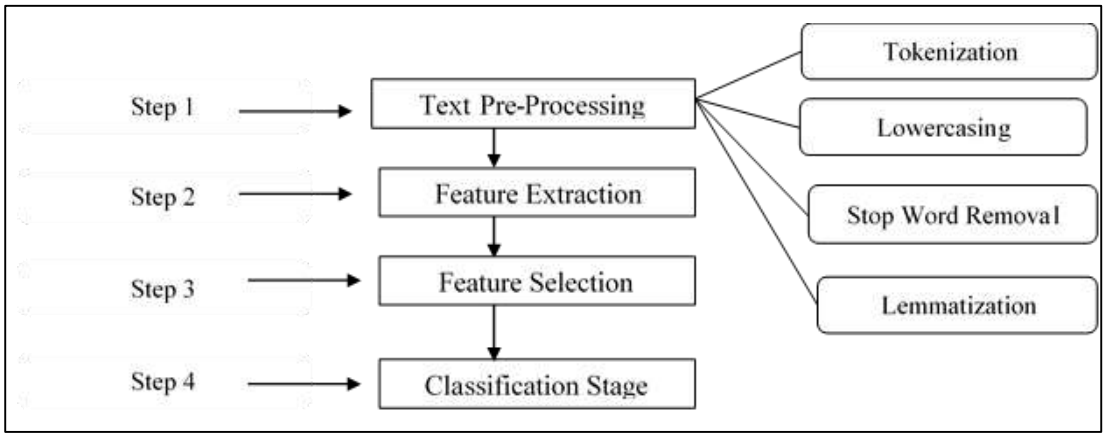


Figure 1: Text classification process

1.2 Text Preprocessing

In this research, we dealt with the first step of the text classification process, "Text preprocessing". "Text preprocessing is to bring the text into a form that is predictable and analyzable for task" (Ganesan 2019). The unstructured text usually contains a lot of useless information such as noisy, unnecessary, repetitive words, numbers, punctuations, HTML tags, URLs, scripts, advertisements, stop words, abbreviations, emoticons, slangs, misspellings, shortcuts, and specific terminology. Because each word counts as a dimension feature set, having unnecessary words causes the model time waste (Işik & Dağ 2020). Text preprocessing is often the first step of Natural Language Preprocessing (NLP) and has the potential impact on its final performance (Camacho-Collados & Pilehvar 2018). Under the text preprocessing, we remove these unnecessary data and clean the data set for further process. The purpose of the pre-processing is to make the input less

complex in a way that does not adversely affect interpretability or conclusion of the model (Denny & Spirling 2018). This is the most important sub-task of the text classification, and the accuracy of the further steps depends on its output. The importance of preprocessing is to make the text classification more effective using preprocessing output. The most successful results in the text classification can be achieved by correctly preprocessing the texts. Because of that, it is important to identify the preprocessing techniques. Normalization techniques play a major role here. As mentioned before, there are various noises in the documents, but in our case, we mainly focus on using four normalization techniques to clean the data.

1.3 Normalization Techniques

Normalization is the search for patterns that reduce the various forms of words presented in the textual dataset and maintains the essential meaning at the same time. Normalization helps to display the semantic behavior similar to nouns. Normalization techniques are best suited for text classification (Da Silva Conrado, Laguna Gutiérrez & Rezende 2012) (Esuli & Sebastiani 2009)(Kadhim 2018). This paper discusses the various types of normalization techniques that can be used for preprocessing when we are doing text classification. We applied different normalization techniques such as lowercasing, tokenization, stop-word removal, and lemmatization for the text files extracted from the Internet. Based on the results, we could understand the importance of applying normalization techniques for data preprocessing.

This research mainly deals with the important and main normalization techniques that can be helpful for text classification. The research also studies the normalization techniques one by one and shows how these normalization techniques can apply to clean the text files. Finally, this research compares the text files before and after applying normalization techniques.

Although a lot of researchers have used normalization techniques, there are instances where some researchers have focused on only one or two normalization techniques (Kannan et al. 2015). In the (Toman et al. 2006) research, the authors have got only stemming and lemmatization as the normalization techniques. As well as the research (Korenius et al. 2004) mainly focuses on stemming and lemmatization as normalization techniques.

Some previous researchers have done similar research under the field of Preprocessing Techniques. However, they have missed some parts of preprocessing techniques. In the research by Uysal & Gunal (2014), tokenization, stopping the word removal, lowercase conversion, and stemming have been considered as the preprocessing techniques. The research by Kadhim (2018), has got only tokenization, stop-word removal and stemming as preprocessing techniques. It has become the main issue when identifying the most relevant and the best normalization for text preprocessing. Although so many researchers have conducted research about cleaning text files for the text classification, most of them have employed only stemming and lemmatization as normalization techniques. And, some of the researchers have used a few input data for the evaluation process. When we do preprocessing, we cannot be limited only for a minimum number of normalization techniques if we need to get a better output. Instead, we have to employ each and every normalization technique since each technique does a specific cleaning task. If we employ a limited number of techniques, the dataset will be cleaned at its minimum level. Alternatively, with the application of a maximum number of techniques, we can clean the data set at its thoroughgoing level and hence, more accuracy can be obtained. Therefore, our research investigates the essential normalization techniques for the text classification using 200 text documents.

.

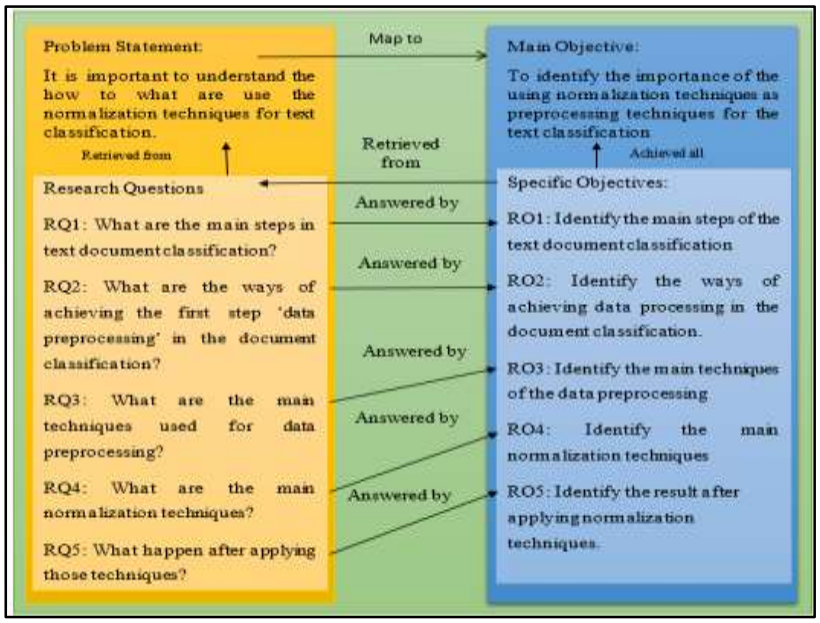


Figure 2: Mapping of research questions to research objectives

We have identified the research questions and objectives as in Fig.2

2 RESEARCH METHODOLOGY / MATERIALS AND METHODS

2.1 Scope of the Study

The main aim of this study is to identify the importance of the normalization techniques for text preprocessing which are beneficial to text classification. This study will investigate the four normalization techniques: tokenization, stop-word removal, lowercasing, and lemmatization. In the process of research, these techniques will be employed for the text preprocessing. Then, the texts without the application of the above techniques and their new condition after applying them will be compared.

2.2 Research Philosophy

The key purpose of the research is spreading the knowledge in normalization techniques for text classification and it primarily deals with one variable: normalization techniques. The present research is based on constructivism philosophy because it generates new knowledge in the normalization techniques based on the human intelligence with the experience of the real world. The reality of this research is subjective because it is constructed being centered to the human mind. This research is innovative with the employment of single method and validates the accuracy of the results.

Moreover, this research falls under qualitative research because it deals with words of the text documents which belong to language category. Besides, the study does not employ numerical data.

2.3 Data Collection

In this experiment, 200 text files are used to apply normalization techniques. These are secondary data which have been downloaded from the internet. The text files are based on both formal and informal styles. They are divided into two groups, namely formal news articles and informal letters. News data set is acquired from the documents produced by the 'Washington Post' news website and the data set of letters is acquired from the documents presented in several websites: 'Answershark', 'Writing Help-central', and 'Letters Free' website. The website where we gathered news articles (Washington Post) is one of the most popular websites in the internet. People all over the world regularly use this website to read

news. In addition, other three websites are also used by various people for their education requirements and to obtain knowledge. These four websites possess decent reviews and they consist with significant amount of data to gather. Hence, we have selected these four websites to collect data.

These text files are in English language and each document belongs to a specific category. Fig. 3 explains the process of extracting datasets from internet and Fig. 4 explains an example for extracting the text files from internet. The Uniform Resource Locator of the internet link is collected first. Then the final data set is downloaded using those URLs. Those HTML files are converted into text files by removing the HTML tags and other tags using the Java programming language. Similar to the other dataset, this output dataset also contains TXT files. Each sentence in the text file has been separated with the full stop. Therefore, those can be loaded easily. The dataset is included in two folders named as 'formal' and 'informal' and each folder contains 100 text files

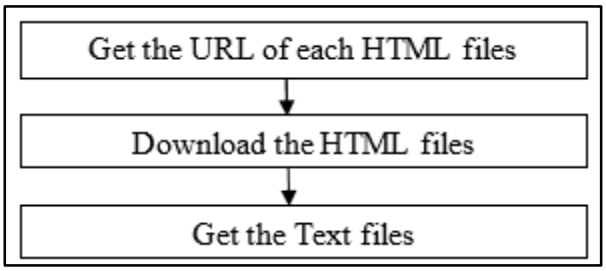


Figure 3. Process of extracting datasets

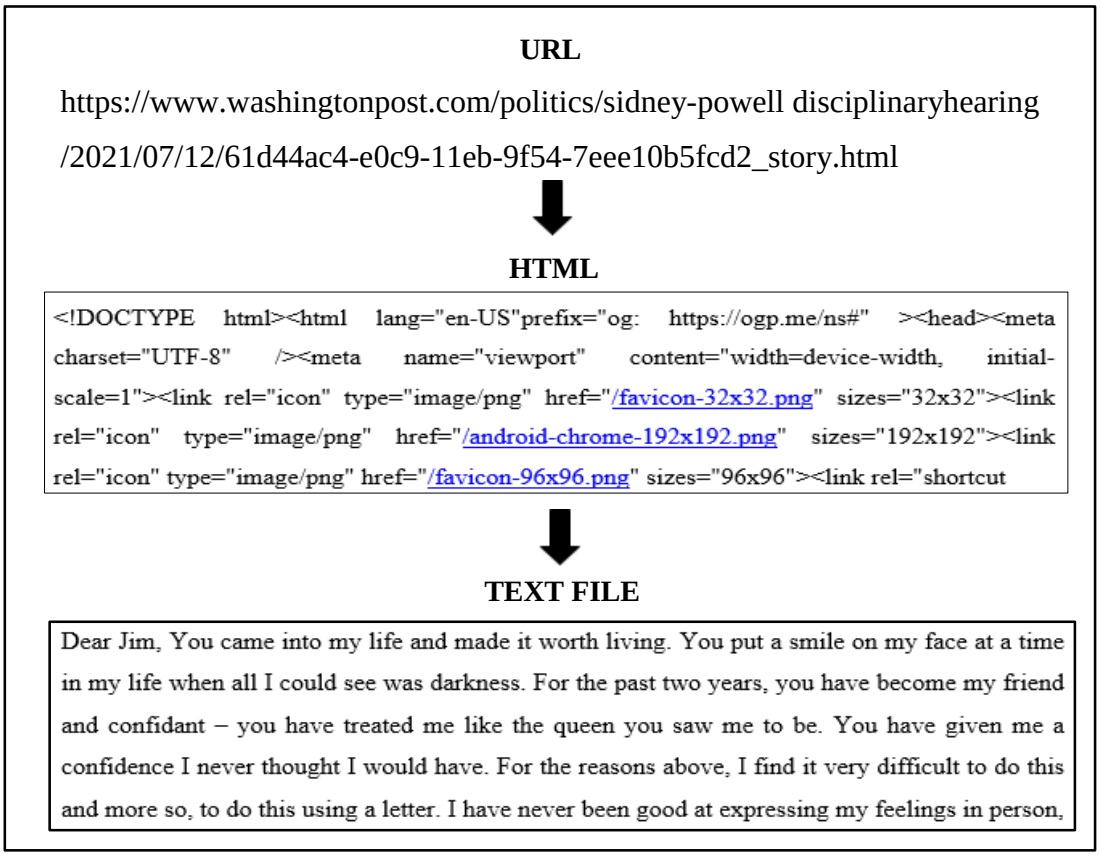


Figure 4: Example for extracting text files from the Internet

2.4 Text Normalization Steps

The main aim of normalizing the text file is to split the text into individual words and to remove the unwanted words. As the scope of the study mentions, the four normalization techniques employed for the research are tokenization, stop-word removal, lowercasing, and lemmatization. The foremost step of the normalization is reading all the text documents that are analyzed in the research. From this stage, all the words in the text document are fragmented into structures (words, phrases, or meaningful parts), which is called tokenization. The next step is removing all the unwanted words in the text document. It is termed as stop-word removal. This step varies from research to research. In the current research, all the prepositions, conjunctions, and articles are removed. Subsequently to this stage, remaining text goes to the next step, lowercasing. Lowercasing is used to represent the text in a simple way. As the final step of normalization, all the words in the text files are lemmatized. Fig. 5 explains each step followed under the normalization techniques.

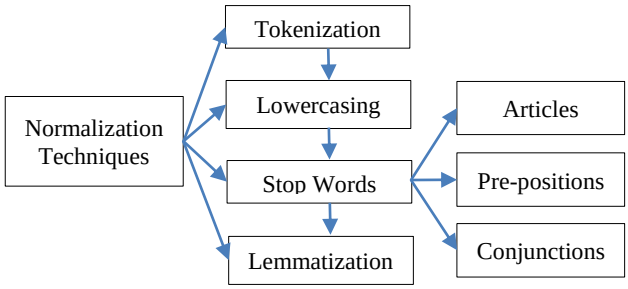


Figure 5: Conceptual Framework

2.4.1 Experiment A: Tokenization

Tokenization can generally be understood as a preset of any kind of natural language normalization technique. It is a default way of breaking the flow of text into words, phrases, symbols or other meaningful elements. Those fragments are named as tokens. Example of Tokenization is shown in Table 1. Simply, this method is used to tokenize content into single words. This step, which can also be called as fragmentation, is usually done with the usage of numbers and letters. Nevertheless, non-numerical characters (punctuations) are not employed. Tokenization is not only separating the words into the basic preprocessing unit but also it interprets and divides the symbols to create higher-level tokens. The main purpose of this step is to examine the words in a sentence.

There are three stages in tokenization.

1. Converting the HTML document into a text document (to get the input text files)
2. Removing empty sequence (Whitespace)
3. Listing down all the words (Tokens)

Table 1: Example for tokenization

	Multiple terms of words	**Phrase**	**Symbol**	**Meaningful elements**
Ex: -	He returned home for the first time in ten years.	The four men were talking because they wanted to spend time.	Dollar	Life is a matter of choices, and every choice make you happy
Tokenized Sentence	{“He”, “returned”, “home”, “for”, “the”, “first”, “time”, “in”, “ten”, “years”}	- The four men were talking. - Because they wanted to spend time.	$	“Life is a matter of choices,” “and every choice make you happy”

2.4.2 Experiment B: Lowercase Conversion

Although lowercasing is overlooked by the researchers, it is the simplest and most efficient way to preprocess a text due to its assistances to condense the text. Capitalization is used to mark the beginning of the sentence in any kind of document. It seems to be that there is no difference between uppercase and lowercase words; nonetheless, in a document with many sentences, capitalization can be a considerable problem when the text is categorized. The most common approach to deal with capitalization is changing all the letters into lowercase. Before classification, all uppercase letters are generally converted to lowercase letters. This technology projects all the words in the text file to be consisted of an identical feature. Lowercasing is applied mostly in text mining and NLP issues and is especially cooperative with the consistency of expected output. Table 2

explains the example for Lowercase Conversion.

Table 2: Example for lowercase conversion

Raw Word	Lowercasing Word
America	**america**
AmericA	
AMERICA	

2.4.3 Experiment C: Stop-Word Removal

Stop-words are the words that are irrespective to the topic of the text. Assuming that stop-words are not relevant to the text, they are removed before the classification. Stop-words are specific to the language being studied as same as in the lemmatization. On the other hand, a list of stop-words can be considered as a list of recurring features that appear in every text file. Conjunctions, pronouns, prepositions, articles and so forth can be considered as types of stop-words. Stop-words should be removed as they have no effect and a value in the classification process. However, removing the stop-words depends on the aim of the particular research. According to the aim of the current study, articles, prepositions, and conjunctions have been removed.

Articles: An article is a word used to modify a noun which is a person, place, object or idea. The definite article is 'the', and it refers directly to a specific noun or groups of nouns. Indefinite articles are the words 'a' and 'an'. Each of these articles is used to refer to a noun which is not specific. (Lyndsay Knowles 2021)

Prepositions: A prepositions is a word or group of words used before a noun, pronoun, or noun phrase to show direction, time, place, location, spatial relationships, or to introduce an object. Some examples of prepositions are words like 'against', 'on', 'at', 'in', and 'of'. (micguides.waldenu.edu)

Conjunctions: A conjunction is a part of speech that connects words, phrases, clauses, or sentences. Some examples of conjunctions are 'for', 'and', 'nor', 'but', and 'or'.

Example for removing stop-words:

Ex: "The bus is full. You'll have to wait for the next one"

For this sentence, stop-words are:

'The', 'to', 'for'

After removing those stop-words, the sentence:

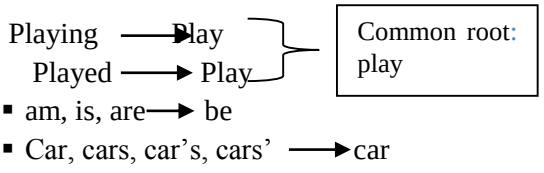
"bus is full. You'll have wait next one"

2.4.4 Experiment D: Lemmatization

Lemmatization is an NLP process that removes the suffix of the word to get the basic form of the word. Lemmatization is a challenging task, especially within wide-ranging languages. Although lemmatization is a more difficult method to normalize the words, sometimes it can be beneficial. The benefit of the lemmatization is that the searcher can match a specific search to the exact index key.

Aimed at the grammatical accuracy, the documents are prepared with the usage of various forms of words such as 'go', 'gone', 'went', 'going', and 'goes'. Besides, there are families of words with related meanings like 'democracy', 'democratic', and 'democratization'. In most cases, searching for the root word seems to be convenient in returning documents that contain another word in the same family.

The common connotation of Lemmatization is undertaking things with a proper usage of vocabulary and morphological analysis aiming only at removing distinct endings and obtaining the root form of a word, which is termed as a lemma. The following examples show the lemmatization process.



Using the above mapping, a sentence could be normalized as follows:

The boy's cars are different colors

↓

The boy car be differ color

The overall process of text normalization is described in Fig. 6.

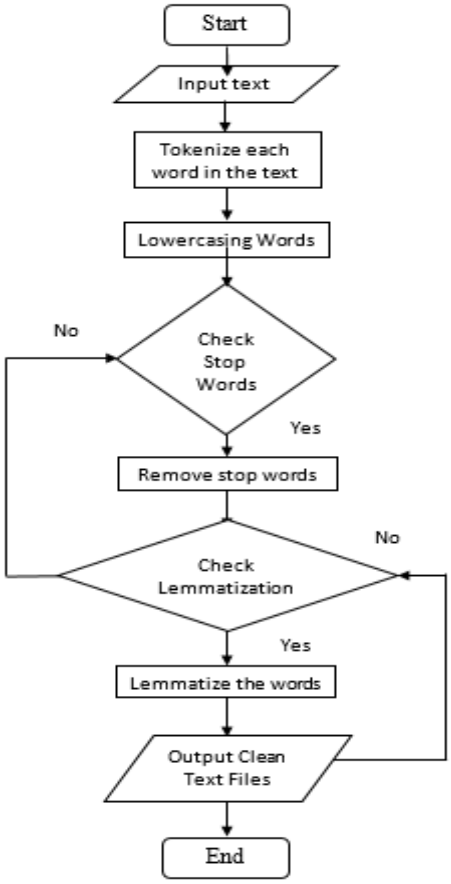


Figure 6: Flow chart for text normalization

3 RESULTS & DISCUSSION

This section discusses the results of the analysis based on four normalization techniques, tokenization, lowercasing, stop-word removal, and lemmatization respectively.

3.1 Tokenization

As stated above, tokenization is used to split the sentence, phrases, or the entire document into small meaningful units. This experiment has been employed in StringTokenizer of 'Java' programming language. It has facilitated splitting the text documents into tokens. By way of the first experiment, tokens in each text document are identified. As mentioned earlier, there are 200 data sets and they are divided into two categories as 'formal' and 'informal'. Following table 3 shows the number of tokens in 10 files among 100 text files from the 'informal' folder.

Table 3: Tokenization – Informal Folder

Name of the Text File	Quantity of the Tokens
Output0.txt	482
Output1.txt	546
Output2.txt	282
Output3.txt	401
Output4.txt	590
Output5.txt	548
Output6.txt	353
Output7.txt	421
Output8.txt	339
Output9.txt	531

Following table 4 shows the number of token in 10 files among 100 text files from the 'formal' folder.

Table 4: Tokenization – Formal Folder

Name of the Text File	Quantity of the Tokens
Output0.txt	2758
Output1.txt	1596
Output2.txt	1865
Output3.txt	2939
Output4.txt	1885
Output5.txt	1580
Output6.txt	1864
Output7.txt	1014
Output8.txt	1544
Output9.txt	1323

Above table 3 and table 4 present the tokens that are included in the text documents before cleaning the text documents. These tokens help to lemmatize the words easily and to scrutinize the comparison between the previous text documents and the cleaned text documents. The following Fig. 9 shows the sample document and the split tokens.

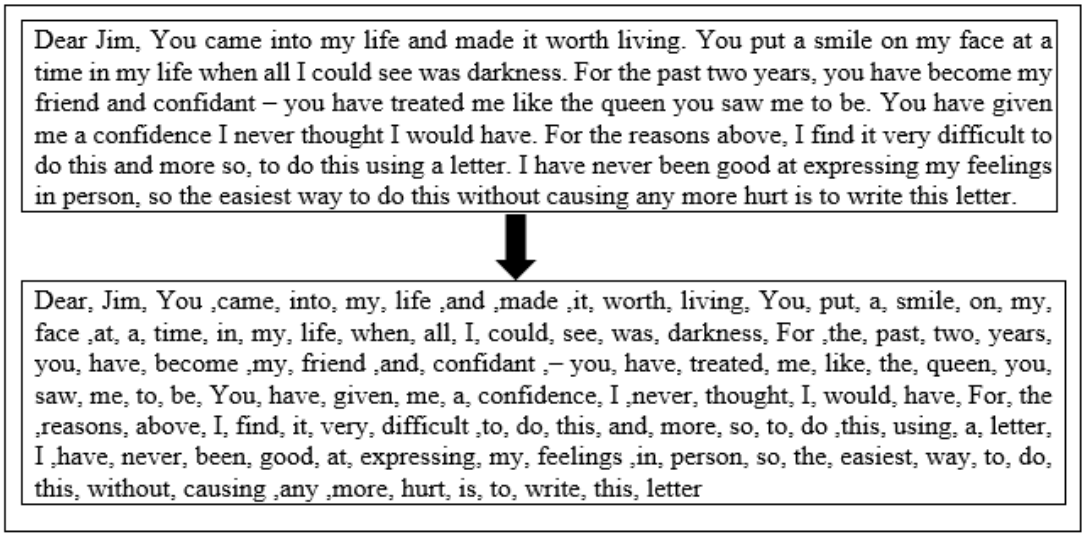


Figure 9: Tokenization- Tokenize File

3.2 Lowercasing

After completing the tokenization, the next step is lowercasing. Lowercasing is the most efficient way of text classification process. Therefore, as experiment 02, all the words included in the text document are lowercased. The following Fig. 10 shows a sample which demonstrates how to lowercase a text document.

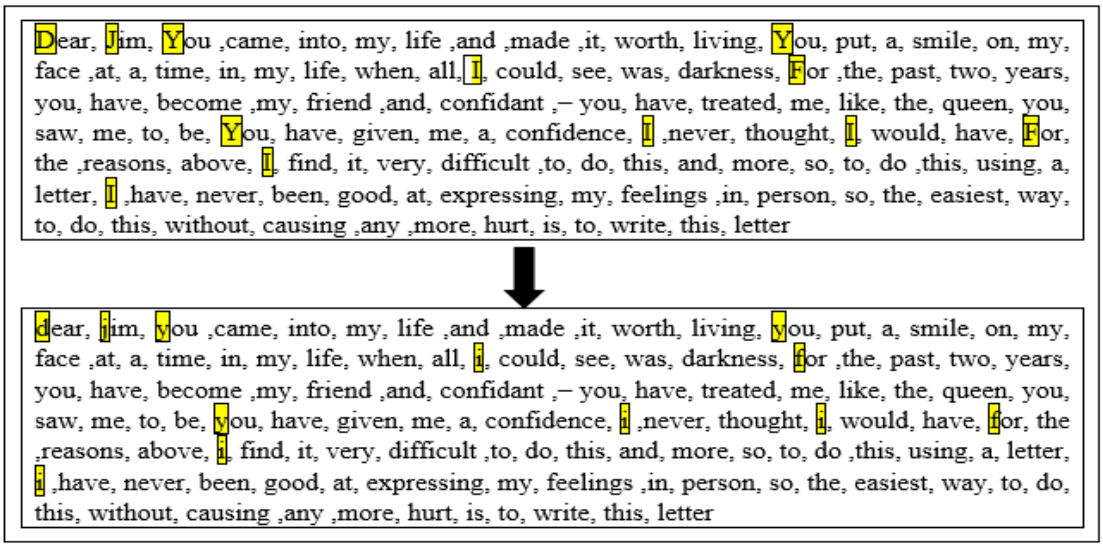


Figure 10: Lowercase Conversion- Lowercase File

In this Fig. 10, first part shows the original text file and it contains capital letters. The second part shows the output after converting the entire document into lowercase.

3.3 Stop-Word Removal

Next to the completion of tokenization and lemmatization, the following step is removing all the stop-words. Here, we have removed the stop-words in each text document. There are various types of stop-words as mentioned formerly. However, for the purpose of this particular research, prepositions, conjunctions, and articles have been considered as the stop-words. Fig. 11 shows the example for removal of stop-words in a text document.

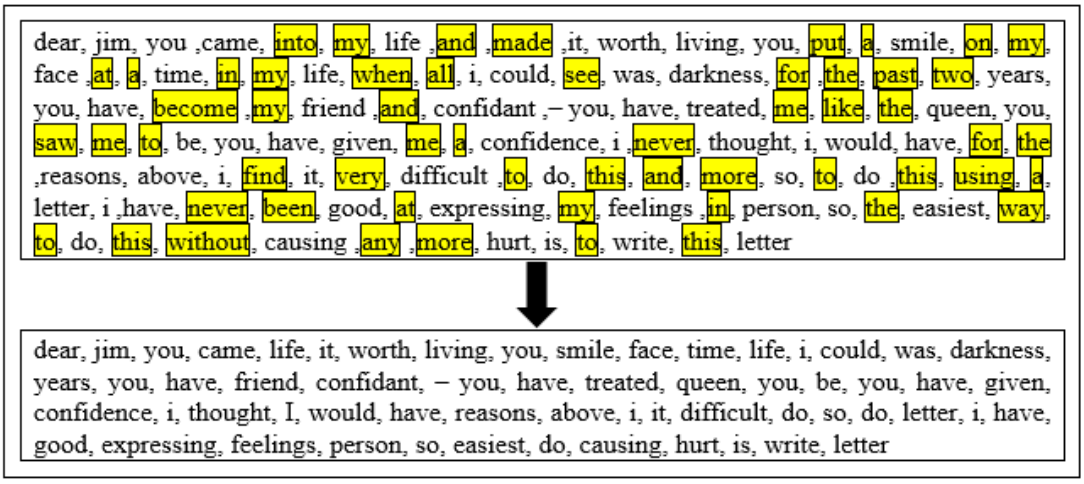


Figure 11: Stop-Word Removal

3.4 Lemmatization

After completing the first three normalization techniques, lemmatization is used to convert a word to its basic form (word root). Following Fig. 12 shows how to lemmatize the words.

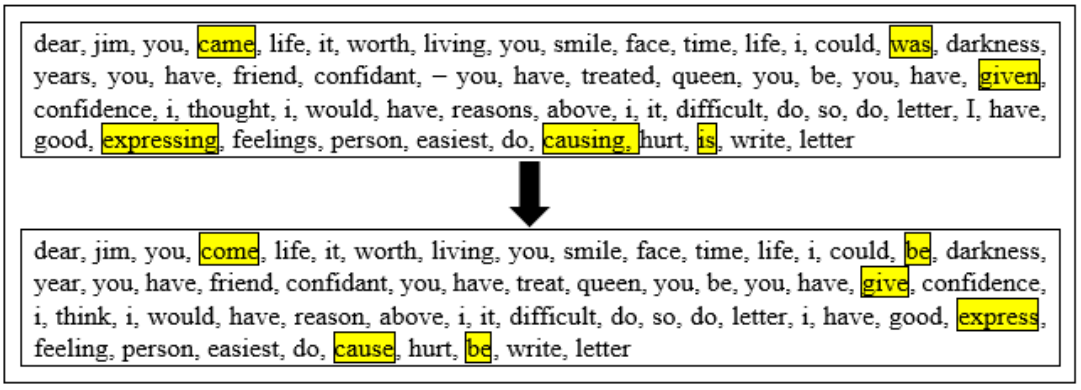


Figure 12: Lemmatization

In Fig. 12, the first part highlights the words before lemmatizing the words included in the returning documents that contain another word in the same family.

According to the final result, we could clean documents using normalization techniques as presented in the following verification.

Table 5: Comparison between the Quantities of Tokens - Formal Documents Folder

Name of the Text File	Quantity of the Tokens	After Adding Normalization Techniques
Output0.txt	2758	2070
Output1.txt	1596	1259
Output2.txt	1865	1413
Output3.txt	2939	2251
Output4.txt	1885	1455
Output5.txt	1580	1215
Output6.txt	1864	1422
Output7.txt	1014	797
Output8.txt	1544	1167
Output9.txt	1323	1009

Table 5 shows the difference between the quantity of tokens included before cleaning the text files and that quantity after applying normalization techniques for the 10 text files selected from the formal folder.

Table 6: Comparison between the Quantities of Tokens - Informal Documents Folder

Name of the Text File	Quantity of the Tokens	After Adding Normalization Techniques
Output0.txt	482	377
Output1.txt	546	436
Output2.txt	282	224
Output3.txt	401	307
Output4.txt	590	472
Output5.txt	548	391
Output6.txt	353	259
Output7.txt	421	314
Output8.txt	339	255
Output9.txt	531	370

Table 6 shows the difference between the quantities of the tokens included before cleaning the text files and after applying normalization techniques for the 10 text files selected from the informal folder.

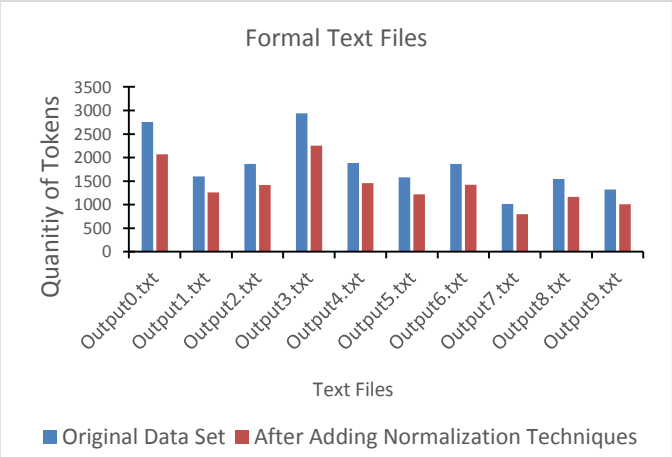


Figure 11. Comparison between the Quantities of Tokens - Sample of Formal Documents

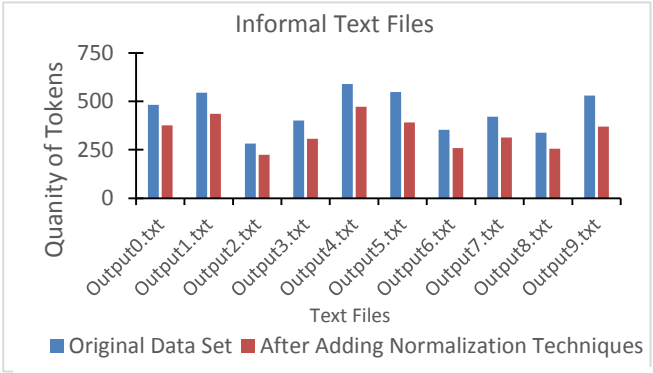


Figure 12. Comparison between the Quantities of Tokens - Sample of Informal Documents

Fig. 11 and Fig. 12 have verified that normalization techniques are an appropriate way to clean the text files. In the first step of preprocessing, which is known as tokenization, words in the document are tokenized. Moreover, tokenization seems to be ensuring the greatest role of the text cleaning process as it is evident through the comparisons between the results after applying the normalization techniques. The next, lowercasing method, also helps to condense the text data. The stop-word removal method has been used as the third step. At that stage, all the stop-words are removed. Then, the lemmatization method has played a significant role as a normalization technique by converting all the words into their basic forms. Accordingly, all the documents have been cleaned following these four steps. A summary of the comparison between the 100 formal documents and the 100 informal documents before and after the normalization is shown in the following Fig.13.

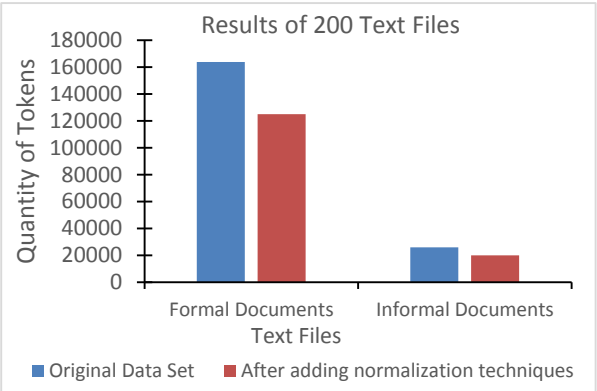


Figure 13. Comparison between the number of tokens – All 200 text files

In conclusion, this research has defined all four normalization techniques and compared the state of all the selected text files with and without the application of the normalization techniques. For the purpose of tokenization, lowercasing, stop-word removal, and lemmatization, Java programming language version 8 and Net Beans Integrated Development Environment (IDE) 8.2 have been used.

There are various kinds of normalization techniques that can be employed to preprocess a particular document according to its needs. In

our case, we have employed four normalization techniques according to the purposes of our further document classification. By these four normalization techniques we have been able to acquire the result which satisfies our requirements. It is challenging to compare the results of this normalization with other normalization techniques since each normalization technique stands for a different cleaning process.

4 CONCLUSION & RECOMMENDATIONS

There are various normalization techniques. Among them, this research has discussed four techniques: tokenization, lowercase conversion, stop-word removal, and lemmatization which are employed in the text preprocessing. Preprocessing can be employed for a successful text classification. These normalization techniques have been applied for 200 text documents and results of each have been analyzed.

Moreover, this paper presents the comparison between the text files with and without the application of normalization techniques. The proposed methods of this study can be used for text preprocessing and for classification of the text documents into various categories. Applying normalization techniques is a vital step which should be done in advance to classify a text. Apart from the various types of normalization techniques, this research paper has described how to apply them to text files for the purpose of reducing words, cleaning the text file, and to open a fruitful path for text classification. Accordingly, we have reached our aim of identifying the importance of the usage of normalization techniques as a preprocessing method for text classification. Based on the evaluation, we have recognized that normalization helps to preprocess data and to improve the accuracy of a text file for further processes.

As a forthcoming effort, we are planning to apply the remaining techniques by changing the input text documents and then identify the most suitable normalization techniques for diverse datasets. Also, these text files are being planned to be employed for text classification based on the formality and informality of them.

REFERENCES

Adetunji, AB, Oguntoye, JP, Fenwa, OD & Akande, NO 2018, 'Web Document Classification Using Naïve Bayes', *Journal of Advances in Mathematics and Computer Science*, pp. 1–11. doi:10.9734/jamcs/2018/34128.

Basarkar, A 2016, "DOCUMENT CLASSIFICATION USING MACHINE LEARNING," San Jose State University [Preprint]. Available at: https://doi.org/10.31979/etd.6jmu-9xdt..

Camacho-Collados, J & Pilehvar, MT 2018, 'On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis', *Proceedings of the 2018 EMNLP Workshop* BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP [Preprint]. Available at: https://doi.org/10.18653/v1/w18-5406.

Denny, MJ & Spirling, A 2018, 'Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It', Political Analysis, pp. 168–189. Available at: https://doi.org/10.1017/pan.2017.44.

Esuli, A, & Sebastiani, F 2009, 'Training Data Cleaning for Text Classification', Lecture Notes in Computer Science, pp. 29–41. Available at: https://doi.org/10.1007/978-3-642-04417-5_4.

Ganesan, K 2018, 'All you need to know about text preprocessing for NLP and Machine Learning', Medium. Towards Data Science.

Available at: https://towardsdatascience.com/all-you-need-to-know-about-text-preprocessing-for-nlp-and-machine-learning-bc1c5765ff67 (Accessed: December 14, 2022).

IŞIK, M, & DAĞ, H 2020, 'The impact of text preprocessing on the prediction of review ratings', TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES, pp. 1405–1421. doi:10.3906/elk-1907-46.

Kadhim, AI 2018, 'An Evaluation of Preprocessing Techniques for Text Classification', International Journal of Computer Science and Information Security, pp. 22–32. Available at: https://sites.google.com/site/ijcsis/.

Kannan, S, Gurusamy, V, Vijayarani, S, Ilamathi, J & Nithya, M 2015, 'Preprocessing Techniques for Text Mining Preprocessing Techniques for Text Mining', *International Journal of Computer Science & Communication Networks*, 5(October 2014), pp. 7–16.

Korenius, T, Laurikkala, J, Jarvelin, K & Juhola, M 2004, 'Stemming and lemmatization in the clustering of finnish text documents', Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management - CIKM '04 [Preprint]. Available at: https://doi.org/10.1145/1031171.1031285.

Malik, HH & Bhardwaj, VS 2011, 'Automatic Training Data Cleaning for Text Classification', 2011 IEEE *11th International Conference on Data Mining Workshops* [Preprint]. Available at: https://doi.org/10.1109/icdmw.2011.36.

Pascual & F (2021) Document Classification.

Silva Conrado, MD, Laguna Gutiérrez, VA & Rezende, SO 2012, June, 'Evaluation of normalization techniques in text classification for portuguese', In *International Conference on Computational Science and Its Applications*, pp. 618-630, Springer, Berlin, Heidelberg.

Toman, M, Tesar, R & Jezek, K 2006, 'Influence of word normalization on text classification', Proceedings of InSciT, pp. 354–358. Available at: http://www.kiv.zcu.cz/research/groups/text/publications/inscit20060710.pdf.

Uysal, AK & Gunal, S 2014, 'The impact of preprocessing on text classification', Information Processing & Management, [online] pp. 104–112. doi:10.1016/j.ipm.2013.08.006