International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

Exploring Preprocessing Techniques for Natural LanguageText: A Comprehensive Study Using Python Code

[1]MR. ADEPU RAJESH, [2]DR. TRYAMBAK HIWARKAR
[1]Research Scholar, [2]Professor and Dean
[1,2] Department of Computer Science and Engineering, School of Engineering and Technology, Sadar Patel University , Balaghat (MP) , India
Corresponding Author Orcid ID: https://orcid.org/0009-0000-4072-9267

Abstract
The paper highlights the significance of efficient text preprocessing strategies in Natural Language Processing (NLP), a field focused on enabling machines to understand and interpret human language. Text preprocessing is a crucial step in converting unstructured text into a machine-understandable format. It plays a vital role in various text classification tasks, including web search, document classification, chatbots, and virtual assistants. Techniques such as tokenization, stop word removal, and lemmatization are carefully studied and applied in a specific order to ensure accurate and efficient information retrieval. The paper emphasizes the importance of selecting and ordering preprocessing techniques wisely to achieve high-quality results. Effective text preprocessing involves cleaning and filtering textual data to eliminate noise and enhance efficiency. Furthermore, it provides insights into the impact of different techniques, such as raw text, tokenization, stop word removal, and stemming, using a Python implementation.
Keywords: Text preprocessing, Natural Language Processing, Web search, Chatbots, Tokenization, Stemming, Python

1. Introduction

Text processing in natural language processing (NLP) involves employing a range of techniques and operations to extract meaningful information and conduct analysis on raw text data. Its primary objective is to convert unstructured text into a structured format that can be comprehended and processed by machine learning algorithms. text representation is crucial to enable the application of these algorithms to text-based data. Text representation in NLP refers to the process of converting text data into a numerical or machine-readable format that can be utilized by algorithms for various NLP tasks. Before conversion suitable data preprocessing techniques should undergo. Data pre-processing techniques are essential for removing unnecessary words, characters, and punctuations that are not useful for machines to interpret. Natural Language Processing (NLP) offers a range of techniques that can be selected based on the specific use case. The order in which these techniques are applied is crucial in certain scenarios. The primary goal of data or text pre-processing is to transform raw data into a structured format, with a focus on extracting keywords that highlight the context of sentences or paragraphs. Natural Language Processing is found various real life applications such as:

Chatbots	Sentiment Analysis	Speech Recognition
Language Translation	Text Summarizers	Autocorrect

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

2. Literature Review

[1] This study focuses on optimizing and enhancing a text classification algorithm. The experiments conducted employ two datasets, namely TREC2007 and Enron-spam, while employing support vector machine, naive Bayes classifier, and k-nearest neighbor classifier for the classification process. The results of the experiments indicate that the proposed method demonstrates superior performance compared to other methods when evaluating F1 value, particularly for the TREC2007 and Enron-spam datasets. Importantly, this improved performance is achieved while minimizing the burden of sample annotation

[2] This paper presents an analysis of the influence of text preprocessing choices on the performance of a standard word-based neural text classifier. The study emphasizes the importance of making careful decisions regarding data preprocessing and maintaining consistency when comparing different systems. Generally, simple tokenization is as effective as, if not superior to, more complex preprocessing techniques such as lemmatization or multiword grouping. However, it should be noted that domain-specific datasets, like the medical dataset used in our experiments, exhibit poor performance with sole tokenization. Surprisingly, word embeddings trained on multiword-grouped corpora demonstrate remarkable effectiveness when applied to straightforward tokenized datasets. This particular attribute has often been overlooked and, to the best of our knowledge, we are the first to test this hypothesis. In fact, this finding may partially elucidate the enduring success of pre-trained Word2vec embeddings, which explicitly learn multiword embeddings as part of their pipeline

Furthermore,the analysis reveals significant variance in results depending on the choice of preprocessing method (on average ±2.4% for the best performing model), especially when the training data is insufficient for generalization. To comprehensively comprehend the significance of these findings, further analysis and experimentation are necessary. Nevertheless, this work serves as a starting point for investigating the impact of text preprocessing in deep learning models. We hope that our discoveries will inspire future researchers to meticulously select and report their preprocessing decisions when evaluating or comparing different models

[3] This paper provided a comparison between the chi-square and TF-IDF with cosine similarity score methods for text classification based on their content. The proposed system incorporates text preprocessing and feature extraction using a thresholding parameter to categorize texts into one or more categories. Through evaluation metrics, it was determined that TF-IDF with cosine similarity score outperformed chi-square in classifying ten general categories. The results indicate that text preprocessing enhances the accuracy of text recognition and improves the overall system performance for text classification

[4] This paper examined a comparison between the chi-square and TF-IDF with cosine similarity score methods for the purpose of classifying texts based on their content into one or more categories. The proposed system incorporates text preprocessing and feature extraction using a thresholding parameter to facilitate this categorization process. According to the evaluation metrics, TF-IDF with cosine similarity score demonstrated superior performance in classifying the ten general categories. The results emphasize the positive impact of text preprocessing on enhancing the accuracy of text recognition and improving the overall system performance for text classification

[5] This paper focuses on investigating the sentiment of online movie reviews. To address the inherent noise in the text, we employed a combination of preprocessing techniques. Additionally, we utilized the chi-squared method to eliminate irrelevant features that have no impact on sentiment orientation. Our study encompasses comprehensive experimental results, demonstrating that the accuracy achieved through appropriate text preprocessing on the two datasets is comparable to the accuracy achieved in topic categorization, which is considered a less challenging problem

[6] This paper projected on the function of text pre-processing in sentiment analysis of the two online data sets (Dat-1400 and Dat-2000) of movie reviews. It combined a number of pre-processing techniques, including the removal of HTML tags, non-alphabetic signs, white space, abbreviation expansion, stemming, stop-word removal, and negation handling, to reduce the amount

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

of background noise in the text. They also used the chi-squared method to eliminate unnecessary features. Authors claimed that SVM, a large number of word unigrams, suitable feature selection techniques, and pre-processing techniques may considerably enhance sentiment analysis. Their accuracy outcomes are on par with those that can be obtained in topic categorization, a considerably simpler task. Using 9,058 word unigrams, they achieved an accuracy of 93.5% on the Dat-2000 dataset.

[7] This Paper explained the effects of text pre-processing techniques on sentiment classification performance in two different classification tasks, and summarised the classification results of six pre-processing techniques (replacing negative mentions (e.g., "won't" into "will not"), removing URL links, changing words that have repeated letters to their original English form, removing numbers, removing stop words, and expanding acronyms to their long forms) using two feature models (the word n-g and the word s-g). removal of stop words, reversing words, removal of URLs, and according to the experimental findings, was replacing negative mentions in the N-grams model. On all five datasets, this technique significantly improves the accuracy and F1-measure of practically all classifiers (the greatest accuracy improvement is 8.23% using SVM and the maximum F1-measure improvement is 10.21% using RF on the SemEval2014 dataset). In the N-grams model, acronym expansion only significantly improved NB (accuracy and F1-measure got increases of 6.85% and 6.08%, respectively).

[8] This Paper discussed the effects of preprocessing on TC utilising tokenization, stopword removal, lowercase conversion, and stemming as preprocessing techniques. They examined four datasets—Turkish emails, English emails, Turkish emails, and English emails—using every combination of the preprocessing techniques. With feature sizes of 10, 20, 50, 100, 200, 500, 1,000, and 2,000 words, they solely used the SVM ML technique. Their key finding was that although improper combinations can actually reduce classification accuracy, right combinations, depending on the domain and language, can significantly enhance classification accuracy. Their research shows that regardless of the language or domain, lowercase conversion increases classification success in terms of precision and dimension reduction.

[9] This article focuses on identifying characteristics and evaluating the performance of stop word removal, stemming, and text lemmatization algorithms used in the well-known NLTK and Spacy libraries. Work is based on the analysis of the results of numerous experiments on the application of the considered text processing algorithms for such typical topics of web publications.

3. **Text Preprocessing**

Text preprocessing techniques play a crucial role in transforming unstructured textual data into a structured format, facilitating easier analysis and prediction for various tasks, such as sentiment analysis. Within the field of natural language processing, there exists a diverse range of text preprocessing techniques that are commonly utilized. Here, I highlight some of the most frequently employed techniques along with python implementation

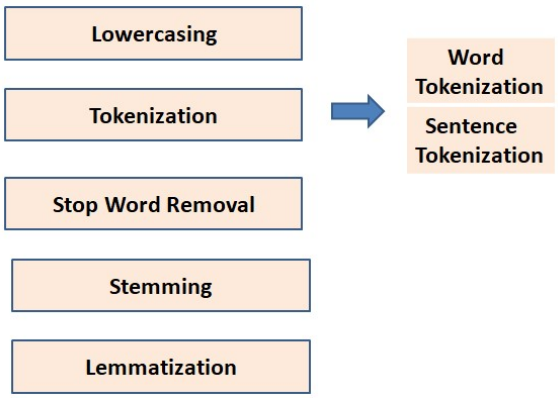


Figure 3.1 : Preprocessing Techniques

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

Raw Data:

The following unstructured comments are considered for preprocessing. These comments are collected from public opinion over internet on Titan submarine incident

@dogstar5927 3 days ago The best, honest and realist interview and discussion about this entire sordid saga since it first started !
@jocelyncarley5124 2 days ago The CEO was wreckless and irresponsible.The sad part was that-- it was preventable, that man should've never taken those people down with him, sad
@maryearll3359 3 days ago ALWAYS overthink a situation
@msv-yu9ui 2 days ago I don't think it would've made a difference, there were so many people who warned Stockton Rush and told him that he shouldn't do it, he brushed it all aside and did it anyways, it didn't seem like it was possible to talk to him. He would've ignored Camerons advice too.
@martinwhite418 2 days ago The goofball saying "go get them"? There was no "them" to get. How do people not understand this?

3.B1-Raw data: Titan submarine incident

a.Lowercasing: In text preprocessing, lowercasing text is the process of making all the letters in a given text lowercase. By treating all words as lowercase, regardless of their original case, lowercasing aims to standardize the text data and simplify analysis. Lowercasing is done as part of text preparation for the following reasons: Normalization, Consistency ,Vocabulary reduction and matching comparisons too.

```
# LOWERCASING

comments='''@dogstar5927 3 days ago The best, honest and realist interview and discussion about this entire sordid saga since it first started !
@jocelyncarley5124 2 days ago The CEO was wreckless and irresponsible.The sad part was that-- it was preventable, that man should've never taken
those people down with him, sad
@maryearll3359 3 days ago ALWAYS overthink a situation
@msv-yu9ui 2 days ago I don't think it would've made a difference, there were so many people who warned Stockton Rush and told him that he
shouldn't do it, he brushed it all aside and did it anyways, it didn't seem like it was possible to talk to him. He would've ignored Camerons
advice too.
@martinwhite418 2 days ago The goofball saying "go get them"? There was no "them" to get. How do people not understand this?'''

comments=comments.lower()
comments
```

```
@dogstar5927 3 days ago the best, honest and realist interview and discussion about this entire sordid saga since it first started !
@jocelyncarley5124 2 days ago the ceo was wreck less and irresponsible.the sad part was that-- it was preventable, that man should've never takenthose people down with him, sad
@maryearll3359 3 days ago always overthink a situation
@msv-yu9ui 2 days ago i don't think it would've made a difference, there were so many people who warned stockton rush and told him that heshouldn't do it, he brushed it all aside and did it anyways, it didn't seem like it was possible to talk to him. he would've ignored camerons advice too.
@martinwhite418 2 days ago the goofball saying "go get them"? there was no "them" to
```

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

```
get. how do people not understand this?
```

3.B2- Lowercasing: Python implementation

b.Tokenization: Tokenization is the process of dividing a text or document into smaller parts known as tokens in text preprocessing. Depending on the amount of tokenization selected, a token in the context of natural language processing (NLP) can range in length from a single letter to a complete word, phrase, or sentence. Tokenization's primary objective is to divide the text into meaningful units that can be utilized as input for various NLP tasks, including sentiment analysis, machine translation, named entity recognition, part-of-speech tagging, and more. There are various levels of tokenization as

i)Word Tokenization- The text is divided up into individual words in this type of tokenization, which is the most prevalent. The phrase "Hello, how are you?", for instance, might be tokenized as ["Hello", ",", "how", "are", "you", "?"].

```
# WORD TOKENIZATION

from nltk.tokenize import word_tokenize
word_tokens_comments=word_tokenize(comments)
for word in word_tokens_comments:
  print(word,end=" ")
```

```
@ dogstar5927 3 days ago the best , honest and realist interview and discussion about this entire sordid saga since it first started ! @ jocelyncarley5124 2 days ago the ceo was wreckless and irresponsible.the sad part was that -- it was preventable , that man should 've never taken those people down with him , sad □ @ maryearll3359 3 days ago always overthink a situation ♥ @ msv-yu9ui 2 days ago i do n't think it would 've made a difference , there were so many people who warned stockton rush and told him that he should n't do it , he brushed it all aside and did it anyways , it did n't seem like it was possible to talk to him . he would 've ignored camerons advice too . @ martinwhite418 2 days ago the goofball saying `` go get them '' ? there was no `` them '' to get . how do people not understand this ?
```

3.B3-Word Tokenization: Python implementation

ii) Sentence Tokenization- The process of dividing a text or document into individual sentences is known as sentence tokenization in text preprocessing. Sentence tokenization's objective is to divide the text into digestible sentence units that can later be independently examined or processed.Sentence tokenization is a crucial stage in natural language processing (NLP) jobs because many NLP algorithms and models work at the sentence level. The input data is organised into discrete units by breaking the text down into sentences, which makes it simpler to apply different NLP approaches and algorithms that operate on sentence-level data.

```
# SENTENCE TOKENIZATION

from nltk.tokenize import sent_tokenize
sent_tokens_comments=sent_tokenize(comments)
sent_tokens_comments
```

```
['@dogstar5927 3 days ago the best, honest and realist interview and discussion about this entire sordid saga since it first started !', "@jocelyncarley5124 2 days ago the ceo was wreckless and irresponsible.the sad part was that-- it was preventable, that man should've never taken\nthose people down with him, sad □\n@maryearll3359 3 days ago always overthink a situation ♥\n@msv-yu9ui 2 days ago i don't think it would've made a difference, there were so many people who warned stockton rush and told him that he\nshouldn't do it, he brushed it all aside and did it anyways, it didn't seem like it was possible to talk to him.", "he would've ignored camerons\nadvice too.", '@martinwhite418
```

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

```
2 days ago the goofball saying "go get them"?', 'there was no "them" to get.', 'how do people not understand this?']
```

3.B4-Sentence Tokenization: Python implementation

c. **Stop word Removal:** Stop word removal is a text preprocessing method used in natural language processing (NLP) to get rid of words that are commonly used but are deemed unnecessary or have little value for a certain job. Stop words include those phrases. "The," "is," "and," "of," "to," "a," "in," and similar words are examples of stop words in English.Prior to further analysis or modelling, stop words are found in the text and removed using a procedure called stop word removal.

To decrease the dimensionality of the text input and increase the effectiveness and efficiency of NLP operations, stop words are removed.

```
# STOP WORD REMOVAL

from nltk.corpus import stopwords as spw_words
spw_words_comments=spw_words.words('english')
filt_spw_comments= [w for w in word_tokens_comments if w not in spw_words_comments]
print(filt_spw_comments)
```

```
['@', 'dogstar5927', '3', 'days', 'ago', 'best', ',', 'honest', 'realist', 'interview', 'discussion', 'entire', 'sordid', 'saga', 'since', 'first', 'started', '!', '@', 'jocelyncarley5124', '2', 'days', 'ago', 'ceo', 'wreckless', 'irresponsible.the', 'sad', 'part', '--', 'preventable', ',', 'man', "'ve", 'never', 'taken', 'people', ',', 'sad', '□', '@', 'maryearll3359', '3', 'days', 'ago', 'always', 'overthink', 'situation', '🖤', '@', 'msv-yu9ui', '2', 'days', 'ago', "n't", 'think', 'would', "'ve", 'made', 'difference', ',', 'many', 'people', 'warned', 'stockton', 'rush', 'told', "n't", ',', 'brushed', 'aside', 'anyways', ',', "n't", 'seem', 'like', 'possible', 'talk', '.', 'would', "'ve", 'ignored', 'camerons', 'advice', '.', '@', 'martinwhite418', '2', 'days', 'ago', 'goofball', 'saying', '``', 'go', 'get', "''", '?', '``', "''", 'get', '.', 'people', 'understand', '?']
```

3.B5-Stop word removal: Python implementation

d. Stemming: Stemming is a text preparation method that helps natural language processing (NLP) break down words into their "stem," or root, form. Words with various inflected forms but the same fundamental meaning can be normalized through stemming. This aids in lowering the vocabulary's dimensionality and enhancing the effectiveness and efficiency of various NLP tasks.

Stemming, for instance, would change the words "running," "runs," and "ran" into their standard stem "run." The words "jumps," "jumped," and "jumping" are all originated from "jump." NLP algorithms use different stemming algorithms with some of the popular ones being- Porter Stemming, Snowball Stemming, Lancaster Stemmking, Lovins Stemming

```
# STEMMING - PORTER STEMMER

import nltk
from nltk.stem import PorterStemmer

# Initialize Python porter stemmer
ps_comments = PorterStemmer()
for word in filt_spw_comments:
  print(ps_comments.stem(word),end=" ")
```

```
@ dogstar5927 3 day ago best , honest realist interview discuss entir sordid saga sinc first start ! @ jocelyncarley5124 2 day ago ceo wreckless irresponsible.th sad part -- prevent , man 've never taken peopl , sad □ @ maryearll3359 3 day ago alway overthink situat 🖤 @ msv-yu9ui 2 day ago n't think would 've made differ , mani peopl warn stockton rush told n't , brush asid anyway , n't seem like possibl talk . would 've ignor cameron advic . @
```

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

```
martinwhite418 2 day ago goofbal say `` go get '' ? `` '' get . peopl understand ?
```

3.B6-Stemming: Python implementation

e. Lemmatization: Lemmatization is a text preparation method used in natural language processing (NLP) to condense words into their "lemma," which is their "base" or dictionary form. By taking into account their grammatical characteristics and identifying the base or root word that best captures their fundamental meaning, lemmatization seeks to normalize words. Lemmatization guarantees that the resulting term is a valid word in the language and keeps its semantic meaning, in contrast to stemming, which only removes prefixes or suffixes. Lemmatization is crucial for maintaining language integrity and meaningful word forms in text classification, sentiment analysis, machine translation, and information retrieval applications.

```
# LAMMATIZATION

from nltk.stem import WordNetLemmatizer
lem_comments = WordNetLemmatizer()
for word in filt_spw_comments:
  print(ps_comments.stem(word),end=" ")
```

```
@ dogstar5927 3 day ago best , honest realist interview discuss entir sordid saga sinc first start ! @ jocelyncarley5124 2 day ago ceo wreckless irresponsible.th sad part -- prevent , man 've never taken peopl , sad  @ maryearll3359 3 day ago alway overthink situat ❤ @ msv-yu9ui 2 day ago n't think would 've made differ , mani peopl warn stockton rush told n't , brush asid anyway , n't seem like possibl talk . would 've ignor cameron advic . @ martinwhite418 2 day ago goofbal say `` go get '' ? `` '' get . peopl understand
```

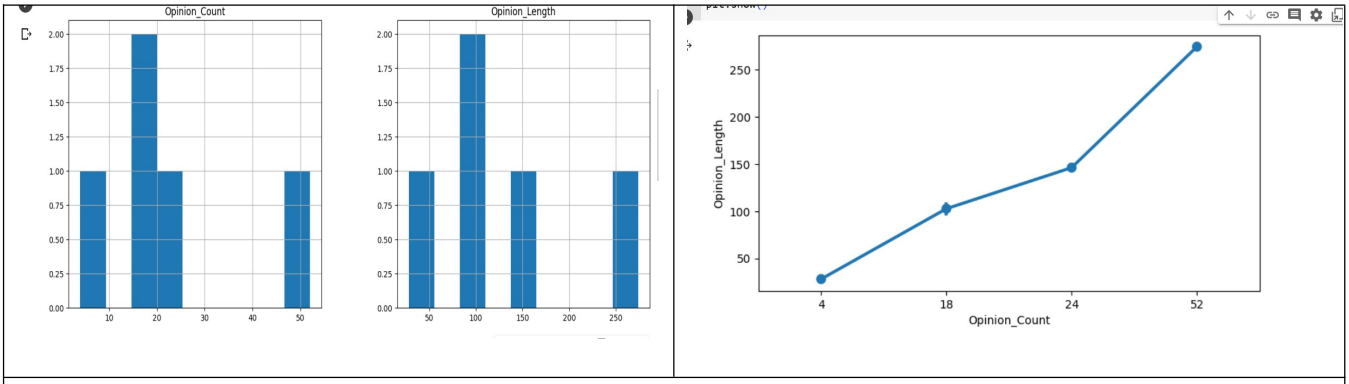
3. B7- Lemmatization: Python implementation

4. Experimental Results:

4.1 Before Text Pre-processing

	Name	Opinion
1	@dogstar5927	The best, honest and realist interview and discussion about this entire sordid saga since it first started !
2	@jocelyncarley51242	The CEO was wreckless and irresponsible.The sad part was that-- it was preventable, that man should've never taken those people down with him, sad
3	@maryearll3359	ALWAYS overthink a situation
4	@msv-yu9ui	ago I don't think it would've made a difference, there were so many people who warned Stockton Rush and told him that he shouldn't do it, he brushed it all aside and did it anyways, it didn't seem like it was possible to talk to him. He would've ignored Camerons advice too.
5	@martinwhite418	The goofball saying "go get them"? There was no "them" to get. How do people not understand this?

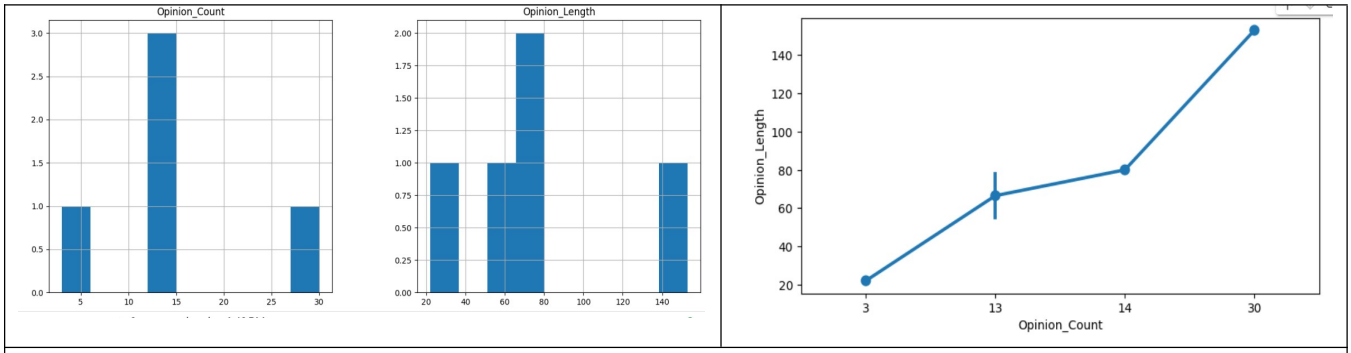
International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 5 Volume No.7 September - October – 2023
DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621



4.1.a Visualization

4.2 After Text Pre-processing

	Name	Opinion
1	@dogstar5927	best , honest realist interview discuss entire sordid saga since first start !
2	@jocelyncarley5124	wreckless irresponsible.th sad part -- prevent , man 've never taken peopl , sad
3	@maryearll3359	alway overthink situat
4	@msv-yu9ui	n't think would 've made differ , mani peopl warn stockton rush told n't , brush asid anyway , n't seem like possibl talk . would 've ignor cameron advic
5	@martinwhite418	goofbal say `` go get '' ? `` '' get . peopl understand



4.2 a. Visualization

5. Text Preprocessing Applications

Text preprocessing finds application in various NLP tasks, including:

a. **Sentiment Analysis:** The purpose of sentiment analysis is to identify the emotional tone of a given text, whether it is positive, negative, or neutral.

b. **Text Classification:** Text preprocessing is used to clean and prepare the text data for categorising documents into predetermined categories in tasks like document classification or spam detection.

c. **Information Retrieval** : Building search engines and information retrieval systems requires text preprocessing since text data needs to be indexed, tokenized, and normalised for effective searching and matching.

d. **Text Clustering**: For text clustering jobs, where comparable papers are grouped together based on their content, preprocessing text data is essential.

e. **Speech Recognition**: Preprocessing is the process of handling speech-specific problems, such as noise removal and turning audio signals into text data, in speech-to-text operations

DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

f. T**ext Summarization**: By omitting superfluous information and maintaining crucial content, text preprocessing aids in the conversion of extensive text documents into succinct summaries in text summarization.

6. Conclusion

The paper focused on the overview of text processing. We have reported experimental results of the text preprocessing stages on the tweets of titan submarine tragedy. In the study of natural language processing tokenization, stop words removal and lemmatization plays an import ant role in the preprocessing approach. Python implementation results are derived for each preprocessing technique. The results of text processing are also projected based on the opinion length and opinion count of the comments. From the results it is observed that after preprocessing the raw data, insignificant words punctuation aspects are eliminated thus makes the model with most optimum. Further the paper discussed on application areas of text preprocessing.

7. References

1. H,Research Article Text Classification Based on Machine Learning and Natural Language Processing Algorithms ,Hindawi Wireless Communications and Mobile Computing Volume 2022, Article ID 3915491, 12 pages https://doi.org/10.1155/2022/3915491
2. On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis Jose Camacho-Collados School of Computer Science and Informatics Cardiff University camachocolladosj@cardiff.ac.uk Mohammad Taher Pilehvar School of Computer Engineering 2018
3. An Evaluation of Preprocessing Techniques for Text Classification, International Journal of Computer Science and Information Security, 16(6):22-32 Ammar Kadhim June 2018
4. International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 6, June 2018 . An Evaluation of Preprocessing Techniques for Text Classification Ammar Ismael Kadhim
5. Department of Computer Science Information Technology and Quantitative Management (ITQM2013) The Role of Text Pre-processing in Sentiment Analysis Emma Haddia , Xiaohui Liua , Yong Shib
6. Haddi E., Liu X., & Shi Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17, 26–32.
7. Pre-processing methods on twitter sentiment analysis. *IEEE Access*, 5, 2870–2879 Hianqiang Z., & Xiaolin G. (2017). Comparison research on text preprocessing
8. Ayedh A., Tan G., Alwesabi K., & Rajeh H. (2016). The effect of preprocessing on arabic document categorization. *Algorithms*, 9(2), 27.
9. Kirill Smelyakov; Danil Karachevtsev; Denis Kulemza; Yehor Samoilenko; Oleh Patlan, Effectiveness of Preprocessing Algorithms for Natural Language Processing Applications, Published in**:** 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T) Date of Conference: 06-09 October 2020
10. Krouska A, Troussas C, Virvou M. The effect of preprocessing techniques on twitter sentiment analysis. In: 7th International Conference on Information, Intelligence, Systems & Applications; Chalkidiki, Greece; 2016. pp. 740- 752
11. Ghag KV, Shah K. Comparative analysis of effect of stopwords removal on sentiment classification. In: IEEE International Conference on Computer, Communication and Control; Indore, India; 2015. pp. 1-6.
12. Sharma P, Agrawal A, Alai L, Garg A. Challenges and techniques in preprocessing for twitter data. International Journal of Engineering Science and Computing 2017; 7 (4): 6611-6613.
13. Safeek I, Kalideen MR. Preprocessing on facebook data for sentiment analysis. In: Proceedings of 7th International Symposium on Multidisciplinary Research for Sustainable Development; Oluvil, Sri Lanka; 2015. pp. 69-78.

DOI:10.46647/ijetms.2023.v07i05.047 ISSN: 2581-4621

14. Ghag K, Shah K. Optimising sentiment classification using preprocessing techniques. International Journal of IT & Knowledge Management 2015; 8 (2) : 61-70