# Text Pre-Processing Techniques in Natural Language Processing: A Review

**Aachal Jakhotiya[1], Harshada Jain[2], Bhavik Jain[3], Ms. Charmi Chaniyara[4]**

*[1]Aachal Jakhotiya, BE Student, ACE, Mumbai*
*[2]Harshada Jain, BE Student, ACE, Mumbai*
*[3]Bhavik Jain, BE Student, ACE, Mumbai*
*[4]Charmi Chaniyara, Assistant Professor, ACE, Mumbai*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract –** *Artificial Intelligence (AI) is the driving force of many upcoming technologies in recent times. It is the intelligence that needs to be induced in machines to make them as intelligent as humans are. For machines to act logically and rationally, they need to be able to understand and interpret human language. This is possible with the help of Natural Language Processing (NLP) which basically is the subset of AI that deals with processing of text or sentences into machine understandable format. With NLP being very important for the computer systems to understand texts, it is very much critical to preprocess the text data to remove noise and structure the data in the correct format for machines to accept as input. Text preprocessing is quite useful in structuring the data so as to design highly efficient models as per requirements.*

***Key Words*: NLP, Natural Language Processing, AI, Text Pre-processing, Tokenization, Stemming, Lemmatization, Stopwords, POS Tagging.**

## 1. INTRODUCTION

Text pre-processing focusses on converting the raw data into a well-defined structure where the words that do not contribute into the contextual meaning of the sentence are discarded. Being an important part of natural language processing, text pre-processing can be done in various ways as there are various techniques available for the same. The choice of technique can be as per the demand of the problem but there are a few techniques that must be used with every problem to enhance the performance of the models. The techniques need to be applied in a specific order to provide the best results possible.

In the world where social media is ruling the minds of people, slangs and short forms are very commonly and extensively used. Humans are aware of all the recent trends and can figure out meanings of any new found words on their own. The same thing cannot be done by machines unless they are given the right training to do so. Hence the pre-processing techniques help in preserving the semantic meaning of the text by identifying the right keywords.

It is necessary to know the techniques, their usage and the order in which they can be applied well in advance. This paper guides you through the various techniques that exist and along with their advantages and drawbacks.

## 2. TECHNIQUES FOR TEXT PREPROCESSING

The various techniques available for text pre-processing are listed as follows. The order in which they should be applied is also the order in which they are listed below:

1. Segmentation
2. Removal of punctuations, special characters and URLs
3. Lowercasing
4. Tokenization
5. Parts-of-Speech Tagging
6. Removing Stopwords
7. Text Normalization
8. Stemming
9. Lemmatization

### 2.1 Segmentation

Segmentation mainly refers to sentence segmentation or sentence tokenization wherein the continuous dataset or a block of text is broken down into meaningful sentences to ease the process of extracting features from text. The breaking point or the point of tokenization for a sentence would commonly be a full stop or a comma in some cases. The text is broken into another one when a full stop or any significant punctuation mark is encountered which makes the remaining part of the sentence look meaningful. Segmentation isn't always carried out for every kind of use case. It depends on the size of every text in the corpus which determines the necessity to use this technique.

### 2.2 Removal of punctuations, special characters and URLs

Raw data has lots of instances of punctuations or special characters (@, $, *,) which are not of much importance nor is understood by the machine. Therefore its existence in data just contributes to the noise in it and should be removed.

This is done by using regular expressions to eliminate all kinds of punctuations and special characters that are encountered. Regular expressions are also used to get rid of URLs in the text that aren't really important usually.

### 2.3 Lowercasing

After successful removal of punctuations and URLs, the text must be converted to lowercase. Lowercasing the text is important as the machine might consider the same word again if it is written in uppercase or any other manner. For example the word 'love' and Love' might be considered differently and would be assigned different vectors when the text is vectorized for feature extraction.

This step might be neglected at times but it is the quite efficient and the simplest method in preprocessing which enhances the possibility of accurate results.

### 2.4 Tokenization

Unlike sentence tokenization, this step primarily breaks up or splits the sentence into an array of words which are referred to as tokens. The sentence is usually split up on space between two words or even when a punctuation is encountered depending on what condition might be applied. For example the sentence:

Text preprocessing is an important step in NLP.

After tokenization will appear as:

"Text", "preprocessing", "is", "an", "important", "step", "in", "NLP", "."

### 2.5 Parts-of-Speech Tagging

To understand the semantic meaning of the text, POS tagging proves to be useful. POS tagging is the classification of text into the various parts of speech such as noun, adjective, verbs, preposition and so on. POS tagging can be done after sentence segmentation as well so that the correct context of the sentence is understood and the words are tagged in a better way.

This technique should be applied before Stopwords removal for the provision of accurate results. It can also be considered to apply this technique before removing the punctuations as well for more accurate results. If POS tagging has to be carried out then these factors should definitely be considered.

### 2.6 Removing Stopwords

NLP is extracting keywords about a particular topic depending on the use case. Hence for text classification or other problems, word such as 'the', 'a', 'is', 'are', 'an', etc. are not of importance and are quite often discarded. Such words are known as stopwords and need to be identified as efficiently as possible. Tokenization of text helps in identifying such words easily without much hassle.

This technique need not be applied always. The kind of problem must be considered before its application.

### 2.7 Text Normalization

Text Normalization can prove to be very useful when analyzing social media comments. It is used to standardize the text which helps in the elimination of noise that is contributed to the data when people usually express themselves on social media. Using text normalization the words 'loveeeeee' and 'luv' can be transformed to its righteous equivalent 'love'.

### 2.8 Stemming

Stemming is a technique used to shorten the word and bring it to its root form. The technique results in shortening the word to an extent where the semantics are preserved but the meaning is lost in some cases. For example the word 'connect' will be stemmed as 'connect' but the word 'trouble' will be stemmed as 'troubl'. The problem should be well studied and then it should be decided whether or not to use stemming for preprocessing. Stemming might prove useful in google searches as all possible searches related to the typed words need to be shown. Porter Stemmer is the commonly used algorithm for stemming.

### 2.9 Lemmatization

Lemmatization is quite similar to stemming but the meaning of the word is very well preserved with it. The word that is transformed to its root form is known as lemma which preserves the semantics as well the meaning of the word. An example for how lemmatization is different from stemming would be that the words 'bullying' and 'bullied' would be stemmed as 'bulli' but it would be lemmatized as 'bully'.

Lemmatization in at times preferred over stemming because of the fact that the meaning of the word is preserved.

## 3. LITERATURE SUREVEY

Rathi Megha [1] has preprocessed text for the purpose of performing sentiment analysis on tweets. The author has suggested lowercasing, converting URLs and usernames to some predefined words. The author also goes on to trim the tweets, removing emoticons, converting words having higher frequency of a letter together and does stemming at the end.

Vateekul Peerapon [2] also uses preprocessing techniques for using data mining on Thai tweets. The author has applied tokenization and removed the unnecessary emoticons, sequence of duplicate characters and single character words as it does not make sense in Thai.

Z. Jianqiang [3] provides a comparative study on the effect of various preprocessing techniques on sentiment analysis. The author concludes that omitting stopwords, URls and numbers does not affect the accuracy much but helps in minimizing the amount of noise in data and hence are quite important steps to be applied.

Ramasamy [4] has surveyed various techniques for preprocessing, majorly the algorithms for stemming. The author agrees to the fact that tokenization is an important step in preprocessing whereas also discusses the efficiency of various stemming alogrithms.

S.P. Paramesh [5] has used preprocessing on data that described IT incidents inorder to build an automated IT helpdesk system. The author uses stopword removal, removal of punctuations and special characters, POS tagging and stemming to achieve the best results.

## 4. CONCLUSIONS

The above techniques specified make it very clear why they can prove to be of utmost importance in the process of natural language processing. As the field of artificial intelligence has been a blessing to the human era, so is its subset natural language processing. To make this process useful it is necessary that the techniques mentioned above are applied in the right manner and right order.

This paper is a survey that very well indicates how using the various preprocessing techniques can have an impact on the accuracy level of the results targeted to be produced. Furthermore it is seen that not all these techniques need to be applied on all kinds of data. For example text normalization is popularly used only on a corpus related to comments on social media whereas lowercasing can be applied on all kinds of data. For most of the cases Lowercasing, Tokenization, Stopwords Removal and Lemmatization or Stemming is always carried out. These preprocessing techniques when combined with the right feature extraction technique can do wonders and provide unexpectedly.

## REFERENCES

[1] Rathi, Megha, et al. "Sentiment analysis of tweets using machine learning approach." 2018 Eleventh international conference on contemporary computing (IC3). IEEE, 2018.

[2] Vateekul, Peerapon, and Thanabhat Koomsubha. "A study of sentiment analysis using deep learning techniques on Thai Twitter data." 2016 13th International joint conference on computer science and software engineering (JCSSE). IEEE, 2016.

[3] Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," in IEEE Access, vol. 5, pp. 2870-2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[4] Ramasamy, Balasubramani & Chandavekar, Naveen. (2016). Survey on Pre-Processing Techniques for Text Mining. 5. 16875-16879.

[5] S.P. Paramesh, K.S. Shreedhara," IT Help Desk Incident Classification Using Classifier Ensembles", ICTACT Journal On Soft Computing, July 2019, Vol: 09, Issue: 04.