International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 06 | June 2020 www.irjet.net p-ISSN: 2395-0072

A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing

Ayisha Tabassum[1], Dr. Rajendra R. Patil[2]

[1]MTech Student, GSSSIETW, Mysore
[2]Professor and Head, Dept. of ECE, GSSSIETW, Mysore

Abstract – *Natural Language Processing (NLP) is a subset of AI that deals with the way machines understand and interpret human Language. Everything that a computer understands is in the form of numbers rather than words. It is therefore worthwhile to study what preprocessing and feature extraction techniques need to be implemented on a human language such that when it is converted to numbers its meaningful enough for the computer to interpret it. One of the major NLP tasks is Text Classification which has found its importance in many use cases such as web search, document classification, chatbots, virtual assistants and so on. Unstructured sentences or texts are inherently difficult to convert to a machine understandable format. Special importance has been given to Pre-processing techniques as they form a precursor on the later stages of information retrieval strategies. Any error in the initial preprocessing techniques is propagated to the later stages of NLP pipeline. Also, the choice and order of the techniques such as tokenization, StopWords removal, lemmatization are carefully studied. Information retrieval systems are particularly concerned with how well the textual data are cleaned and filtered to remove noisy data that do not contribute to increasing the efficiency and moreover lead to erroneous results. This Survey paper emphasizes on the importance of the efficient use of text preprocessing techniques along with Information retrieval using feature extraction techniques of Natural Language Processing.*

***Key Words*: Natural Language Processing, NLP, ML, Text Classification, Tokenization, Punctuations Removal, Stopwords removal, POS Tagging, Lemmatization, NER, TF-IDF, Bag-of-Words ,feature extraction.**

1. INTRODUCTION

Data Pre-processing techniques play a vital role in cleaning up the unwanted words, characters or punctuations that are not useful for the machines to interpret. Thus, Natural Language Processing defines quite a lot of techniques that can be chosen as per the use case. The order in which these techniques are applied is also of utmost importance in some cases. Data or Text pre-processing techniques mainly deal with converting the raw data into an understandable structure where importance is given mostly to the keywords present in the text that highlight the context of the sentence or paragraph.

In today's world where slangs and short forms are the norm of any speech or talk, training a machine to understand these terms needs extra pre-processing. Human Mind can find the keywords in a huge document by perceiving the meaning and thus easily understand to which category the document belongs to. Text or Data Pre-processing techniques help extract these fundamental keywords so that the machine can perform the clustering or classification operations.

There are a variety of techniques available for text pre-processing. This survey paper gives a brief introduction of the text pre-processing and feature extraction techniques, along with the literature survey that highlights the most commonly used methods as well as their applications and drawbacks.

2. TEXT PRE-PROCESSING TECHNIQUES

Pre-processing a text simply means to bring the document to a format that is easily understandable, predictable and analysable by the machine through the various machine learning algorithms

Some of the widely used pre-processing techniques are:

1. Sentence Segmentation
2. Change to Lower Case
3. Tokenisation
4. Parts-of-Speech Tagging
5. Stopwords Removal
6. Removal of Punctuations
7. Stemming
8. Lemmatization

2.1 Sentence Segmentation

Sentence Segmentation is also called as Sentence Boundary detection which refers to the process of breaking a Text document or corpus into individual sentences. This helps in identifying word boundaries so that further processing can be done on each sentence. Segmentation is done at the occurrence of a Full stop or a punctuation using sentence tokenizer.

2.2 Change to Lowercase

Text usually consists of abbreviations and all capital letters. This step is commonly neglected but it is one of the simplest and effective step of text preprocessing especially in cases where dataset is significantly sparse, it has been found that variation in capitalization (e.g. 'India' vs. 'india') gives different results. This means that same words with one in capital and one in lowercase are presumed by the computer as two different words and two different word vectors are formed in the later stages of word embeddings. Thus making all words lowercase has been the best practice in text pre-processing.

2.3 Tokenization

Tokenization refers to splitting of sentences into words, characters, punctuations all of which are called as tokens. The splitting criteria is mainly at the occurrence of a space or a punctuation. This step helps in filtering out unwanted words in further processing steps.

For Example:
"NLP is the future of Speech Recognition Systems!"
This sentence will be tokenized as
"NLP", "is", "the", "future", "of", "Speech", "Recognition" "Systems", "!"

2.4 Parts-of-Speech Tagging

POS Tagging refers to tagging the words in a sentence into the basic grammar classes such as noun, pronoun, verb, adjective, preposition and so on. The POS tagger considers the context of the sentence in assigning the parts of speech much like how a human does. Also, the better the sentence segmentation has been done the better it is for the POS tagger to identify the parts of speech. An incorrect sentence boundary would lead to terrible errors

2.5 StopWords Removal

Words such as "the"," are", "is", "and" and so on do not have significance in Natural Language processing except in certain specific use cases. For example, Text or document classification use case doesn't give weightage to these extra words. Only the keywords that form the topics are extracted. Thus, the more these StopWords are identified and cleaned up the better are the results of classification algorithms. It is also worthwhile to note that in certain use cases like conversational models the usage of certain negation words like "No"," cannot"," wont","not" are of utmost importance to find the context of the sentence and its intent.

2.6 Removal of punctuations

The machine doesn't understand punctuations and thus its existence makes the text noisy. In particular, unstructured documents have numerous such occurrences such as exclamation marks, apostrophe, comma and so on.

Regular expressions (Regex) are the most commonly used pattern searching method applied to strings to identify and replace the punctuations with a common rule.

2.7 Stemming

Stemming is an aggressive word shortening technique that aims to bring a word to its base root. The suffixes are chopped off to bring it to base root while the semantic meaning of all the different forms remains same. However, this doesn't always provide good results as the word loses its meaning [14].

For Example: 'Studying' changes to 'Study' wherein the "ing" is chopped off while retaining the root word.

But stemming applied to words like 'Coding' will result in 'Cod' which doesn't make sense.
Thus, the need for stemming depends again on the problem statement.

2.8 Lemmatization

Lemmatization either removes or replaces the suffix of the word to bring it to its base called as lemma. Lemma is always a meaningful word unlike a stemmed word. Lemmatization is a widely used text preprocessing step in Natural Language Processing and has proven to give great results.

For example: The lemma of the word 'Caring' will be 'Care' which is a meaningful word.

The order in which these steps are implemented is important. For example if StopWords such as "is","the","in" are removed prior to POS tagging then the POS tagger will tag the sentence incorrectly leading to erroneous results.

3. FEATURE EXTRACTION

Feature extraction also refers to the representation of features in vector forms to make it understandable to the machine. Each of the features that these techniques extract is finally represented in vector form that are then fed to the classifier models.

Some of the feature extraction techniques discussed here are

1. Named Entity Recognition (NER)
2. Bag-of-words Model (BoW)
3. TF-IDF

3.1 Named Entity Recognition (NER)

As the name suggests, NER finds the named entities such names of person, organization and location. It works similar to POS tagging. NER is particularly useful to distinguish the chunk of nouns and getting a clear picture of names, countries and organization. This can form a foundation step in use cases where an overview of named entities in document are needed to retrieve as part of Information retrieval.

3.2 Bag of words Model

Bag of words technique plays a fundamental role in extracting features from the text that are later fed to a classifier. It basically groups the features together based on the number of occurrences of the word in the document. It is only concerned with the occurrences of the words rather than where the word occurs in the text.

Thus, the intuition is that texts or documents that contain the same words are similar in context. This model is used in document classification, keyword matching and so on.

A problem with BoW model is that it gives preference to words that occur more frequently thus making it more important[11]. However there might be words that might have a higher occurrence rate but do not have high information content to help in classification or clustering problems. Also lengthier documents give a higher rate compared to smaller ones thus resulting in less accuracy for BoW model. Thus the drawbacks of BoW is taken care by TF-IDF which is discussed next.

3.3 TF-IDF

This TF-IDF is used to penalize the words that are frequent but hold less value in the context of document. It basically has two parts namely Term Frequency and Inverse document frequency.

TF finds the frequency of a word in a document without biasing the size of the document. This is because it is divided by the length of document as a normalization technique thus giving fair accuracy to both small and huge datasets[12].

Its counterpart IDF finds the importance of a word in the document. It basically scales down the words of less importance by computing log of ratio of total documents to the total occurrences of the word in that document.

4. LITERATURE SURVEY

S.P. Paramesh [1] has proposed that in the automation of IT tickets the description data must be cleaned to get better results. The author has used StopWords removal technique using the standard English Stopwords list along with POS tagging, stemming. Porter stemmer has been used to reduce the word to its base form. Additionally, pattern matching regex techniques are used to remove phone numbers and email ids from the dataset to get better results.

Feras Al-Hawari and Hala Barham [2] have found the accuracy by comparing the prediction accuracy of classification models by training it with a dataset that was not preprocessed and the other which was preprocessed and cleaned. The experiment showed that all four classification models had a higher accuracy for a preprocessed and cleaned description data with almost 20 to 30% jump in each case.

Shreedhara K.S [3] has emphasized on the need of efficient text preprocessing in removing noisy data. Stopwords are removed along with date, time and numerical data using regular expressions that are helpful in pattern matching. The author has used random oversampling and under sampling techniques to balance the imbalanced data.

N. Vasunthira Devi [4] discussed the NLP techniques that are used in healthcare and education along with their limitations and usefulness. In education systems the eloquent grammar and corpora are useful in enhancing the skills of students. In contrast to this, in the healthcare systems lack of access to healthcare records pose a challenge.

Of utmost concern is to train a machine with the usage of various healthcare terminologies even though they are referring to the same term. Thus, incorporating NLP techniques in a healthcare system is a challenge and can often lead to high error rates.

Elizabeth D. Trippe[5] discussed that text preprocessing has a significant affect on all the steps relating to right from feature extraction to feature selection in the later stages og Natural Language processing. Tokenization , filtering and stemming are used by the author.

Dr. Balasubramani[6] evaluates Porter's and Krovetz algorithm as part of stemming technique by looking at its applications and limitations. The author concludes that if the document is huge then the algorithm becomes inefficient. This is because of the algorithm's inability in handling words that are not present in its corpora or vocabulary. Also the stemmed words do not give meaning that is required to connect it with the context of the text.

Ranjan Satapathy [7] has proposed two models for microtext language that is most often used in social media such as twitter. The contractions such as 'btw','wrt','lol','2morrow' are normalized using two approaches namely lexicon and phonetic models. The author concludes that sentiment analysis in these contractions are accurately identified using the said models along with a polarity classifier that will classify the intent as being positive or negative.

Mohammad Taher Pilehvar[8] analyzed the effect of text preprocessing techniques on the decisions of a neural text classifier. The author has found that tokenization gives the same level of performance for both small and huge datasets whereas the performance of techniques such as

lemmatization and multi word grouping depends on the domains to which the datasets corresponds to.

Also, multi word grouping has been found to give best results in case word2vec is used as the word embedding technique in the later stages of NLP pipeline.

Sonit Singh [9] emphasizes on the fact that errors in the text preprocessing stages such as tokenization, parsing, stemming get propagated to Information extraction tasks in the later stages. Th author evaluates state-of-the-art systems for IE tasks such as NER, coreference resolution, temporal expression extraction and relation extraction.

The knowledge base that is made as a result of these Information extraction strategies are then acted upon by the subsequent information retrieval methodologies to result in a high semantic understanding of the underlying text.

Resham N. Waykole[10] analyzed bag of words and TF-IDF as potential techniques for feature extraction. Adding word2vec model along with random forest classifier has proven to give better results for text classification on a clinical cancer data.

In contrast to logistic regression and random forest classifier, word2vec scores better since the words with the same meaning are placed close together in forming the vectors so that the actual context of the text is retained when vectorizing texts. Thus, giving better results.

5. CONCLUSION

Natural language Processing has been a boon to make the computers understand speech in its native form. But as the complexity of natural language increases with slangs and internet acronyms it gets difficult to analyze texts especially the unstructured documents. Natural Language is complicated, and 100% accuracy cannot be expected through any machine learning model. Machine learning algorithms have so far made a near close prediction of any entity be it text, image, audio and so on.

In this survey, it was found that text preprocessing techniques form a major contributor in increasing the accuracy of any text-based machine learning algorithm. Also, it is worthwhile to note that not all the preprocessing techniques are used in every use case. The order in which the NLP pipeline is made also dominates the result. It is found that Tokenization, StopWords removal, punctuations removal, lemmatization are the widely used and efficient text pre-processing techniques. Most widely used Feature Extraction techniques are Bag-of-words model and TF-IDF. TF-IDF is by far the best choice for highlighting the prominent features and downscaling the irrelevant features. These are predominantly used in Web Search or Search Engine tools. Depending on the use case these steps can either be enhanced or removed if unnecessary.

REFERENCES

[1] 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

[2] Feras Al-Hawari, Hala Barham, "A machine learning based help desk system for IT service management", Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.04.001

[3] Paramesh S.P, Shreedhara K.S,"Building Intelligent Service Desk Systems using AI",IJESM 2019; Vol: 1, No: 2, pp: 01-08

[4] N.Vasunthira Devi, Dr. R. Ponnusamy, "A Systematic Survey of Natural Language Processing (NLP) Approaches in Different Systems,",IJCSE 2016, vol. 4,issue 7, pg 192-198.

[5] Elizabeth D. Trippe , Krys Kochut , Juan B. Gutierrez, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", arXiv:1707.02919v2 [cs.CL] 28 Jul 2017

[6] Dr. Balasubramani, Arjun Srinivas Nayak, Ananthu P Kanive, Naveen Chandavekar,"Survey on Pre-Processing Techniques for Text Mining" ,IJECS, Volume 5 Issue 6 June 2016, Page No. 16875-16879

[7] Ranjan Satapathy, Claudia Guerreiro, Iti Chaturvedi, Erik Cambria,"Phonetic-Based Microtext Normalization for Twitter Sentiment Analysis", 2017 IEEE International Conference on Data Mining Workshops.

[8] Mohammad Taher Pilehvar ,Jose Camacho-Collados,"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 Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP, pages 40–46

[9] Sonit Singh," Natural Language Processing for Information Extraction", arXiv:1807.02383v1 [cs.CL] 6 Jul 2018

[10] Resham N. Waykole , Anuradha D. Thakare," A Review Of Feature Extraction Methods For Text Classification" ,IJAERD, Volume 5, Issue 04, April -2018

[11] S. Silva, R. Pereira, and R. Ribeiro. Machine learning in incident categorization automation. In 2018 IEEE 13th Iberian Conference on Information Systems and Technologies (CISTI). 2018, pp 1-6.

[12] J. Xu, R. He, W. Zhou, and T. Li, "Trouble ticket routing models and their applications," IEEE Trans. Netw. Service Manage., 2018,vol. 15, no. 2, pp. 530–543.

[13] Parth Suthar, Prof. Bhavesh Oza: "A Survey of Web Usage Mining Techniques" International Journal of Computer Science and Information Technologies, Vol. 6 (6), 2015, 5073-5076

[14] C.Ramasubramanian, R.Ramya: "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm" International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013

[15] Chetan Arora, Mehrdad Sabetzadeh, Lionel Briand & Frank Zimmer, "Automated Checking of Conformance to Requirements Templates Using Natural Language Processing", in IEEE Transactions on Software Engineering, Vol-41, Issue: 10, PP. 944 – 968, 2015.