Paper Title: An Approach for Effective Text Pre-Processing Using Improved Porters Stemming Algorithm

Link: https://ijiset.com/vol2/v2s7/IJISET_V2_I6_102.pdf

1 Summary:

1.1 Motivation:

The paper "An Approach for Effective Text Pre-Processing Using Improved Porters Stemming Algorithm" proposes an improved version of Porter's stemming algorithm for effective text pre-processing. The algorithm focuses on removing prefixes and suffixes, cleaning noisy data and incorporating user feedback for better results.

1.2 Contribution:

This paper focuses on improving the effectiveness of Porter's stemming algorithm by removing derivational and inflectional affixes, reducing words with the same root to a single form, and incorporating statistical analysis techniques for better results. This paper also emphasizes the importance of pre-processing steps such as stop-word removal, punctuation and digits removal, and cleaning noisy data to achieve better outcomes in information retrieval and natural language processing tasks. The proposed algorithm aims to enhance retrieval effectiveness, reduce the size of indexing files, and save storage, space and processing time by representing key terms of a query or document with stems instead of original words.

1.3 Methodology:

Before beginning the stemming process, the methodology entails removing pre-processing procedures such tokenization, digit removal, punctuation removal, and stop word removal. The algorithm focuses on stripping both the prefix and suffix of a given word encountered in the stemming process. The paper also discusses the use of a POS tagger to determine if the root word obtained from the stemming process is a legitimate real word or not. The paper mentions the use of statistical analysis and techniques in the stemming process to remove affixes. The three modes of the stemming algorithm discussed in the paper include the truncating method, n-gram method, and mixed algorithms. One benefit of the stemming algorithm is that it makes finding documents easier for users by eliminating the need to consider word forms.

1.4 Conclusion:

The study contends that although the method frees up memory, it neglects to focus on word lexical analysis and the dimensionality of their original meaning. The research suggests that, in order to remedy this, antonyms profiling words could aid in matching the original meaning while keeping length in mind. The paper also highlights the importance of pre-processing steps, such as stop-word removal and cleaning noisy data, to attain good results in text pre-processing.

2 Limitations

2.1 First Limitation:

One of the main limitations is that aggressive stemming, particularly when both prefixes and suffixes are removed, can lead to a loss of meaning. Words may be reduced to their root form, but this may result in a lack of specificity, and the stemmed form may not accurately represent the original meaning of the word.

2.2 Second Limitation:

The algorithm might remove prefixes and suffixes indiscriminately, leading to over stemming. Over stemming occurs when different words with distinct meanings are reduced to the same root, making it challenging to distinguish between them.

3 Synthesis:

The pre-processing steps involved in the approach include stop-word removal, punctuation and digits removal, and identification and removal of prefixes using a predefined prefix dictionary lookup. The paper emphasizes the importance of pre-processing in achieving good results with the stemming algorithm. The proposed approach focuses on removing affixes, plurals, and suffixes from words, as well as converting terminal 'y' to 'i' when there is another vowel in the stem.