**Analysis of a novel model of a text-to-graph categorization based on natural language processing and deep learning techniques**

**Author**

**Pascal Muam MAH**

AGH University of Science and Technology, Krakow, Poland, **[mah@agh.edu.pl](mailto:mah@agh.edu.pl)**

Supervisors

**Iwona SKALNA**

AGH University of Science and Technology, Krakow, Poland, [skalna@agh.edu.pl](mailto:skalna@agh.edu.pl)

**Tomasz Pełech-Pilichowski**

AGH University of Science and Technology, Krakow, [tomek@agh.edu.pl](mailto:tomek@agh.edu.pl)

**Abstract**

**Background:** This paper introduces an inductive evaluation of categorizing natural language documents into a graph-like structure to a rich content category. Text categorization has very important function in information retrieval thanks to natural language processing systems.

**Objective**: Aims to develop a text categorization system where each word in a text context matters in the extraction process. Also, aims to eliminate focus on frequently used words (FUW) by most algorithms.

**Problem**: The study realizes that tokenization, classification, text tagging, and summarization have been biased on document context. Many algorithms focus on frequently used words (FUW) which creates a bias to text content categorization present online and on strong databases. Weak databases and less frequently used words (LFUW) with rich content end up being unnoticed.

**Method**: The study used text content, natural language processing, and deep learning techniques to construct a word-to-graph model name (scoring model). Each text word was coded with figures and measured with behavior-oriented drive and influential function.

**Results**: Based on the behavior-oriented drive and influential function, a score grade of 3.61 out of 5 was recorded.

**Conclusion**: The study concluded that content categorization evaluation of words-to-graph-like structure is a perfect model that will achieve unbiased extraction. The study also concluded that this novel model is excellent because it takes into consideration each word in the text content. This approach weighs each word and doesn't focus on popular and frequently used words as other algorithms do.

**Recommendations**: The study recommended professionals in the field of natural language processing and deep learning to review this novel model, train it and implement it as a modern approach to document categorization techniques.

**Keywords:** Text categorization, natural language processing, deep learning, bag-of-words model, scoring model

**Main Contribution**

The study is a proposed model called (scoring model) with the objective to modify the functionality of the bag-of-words model. The study is aimed at simplifying representation used in natural language processing and information retrieval (IR) applied deep learning.

The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier.

This study makes it possible to classify, summarize and categorize text or sentences or documents based on the word frequency or occurrence by weighing its relevance, providing grading scale, aligning a pattern of representation and providing a distinct pattern and format that differentiates it from similar or same content type. Brief explanation of the core objectives.

* Weighing document relevance takes care of grammatical errors unlike bag of words model that doesn't consider grammatical errors
* Providing grading scale make sure two or more documents shouldn't be the same except they received the same score.
* Providing a distinct pattern and format that differentiates one content or document from similar or same content type. The study with the help of a scoring model realized that there's a need for decoding information especially for the purpose of security. A lot of cyber security challenges merits the need to code text. The pattern identified in this study can be confidential between producer and end users.

The (scoring model) enables that when applying algorithms in NLP, which works on numbers, we make sure we automatically feed our text into that algorithm. This model does not require the use of a preprocessing of text by converting it into numbers unlike the bag of words, which keeps a count of the total occurrences of most frequently used words. This model automatically generates numbers to all the words or sentences and begins extraction. This automatic generation of numbers to words makes it easier to evaluate content relevance.

**NB:** The scoring model assumes that information comes from humans and deserves a grade to distinct, and recognize intellectual ability and capabilities