A Synonym-based Approach for the Semantic Indexing of Texts

Georgios Grigoriadis–Kotsalis
Department of Computer Science and Engineering
University of Ioannina
Ioannina, Greece
cs03209@uoi.gr

Stavros D. Nikolopoulos
Department of Computer Science and Engineering
University of Ioannina
Ioannina, Greece
stavros@cs.uoi.gr

Iosif Polenakis
Department of Computer Science and Engineering
University of Ioannina
Ioannina, Greece
ipolenak@cs.uoi.gr

ABSTRACT

In this work, we present an algorithmic technique for text indexing based on the utilization of classes of synonyms. The method proposed in this study utilizes a set of synonym classes in order to develop a more abstract representation of a given text focusing on the indexing of texts that express semantic similarity, according to the terms utilized. The content of the texts under consideration is represented by a set of terms that correspond to the class of synonyms substituting each term of the sentences of the text. In the proposed approach the terms are stored into vectors where the uniqueness and the multiplicity of their appearance inside the text are considered to deploy a corresponding similarity metric. Through the development of our model, we omit words that consist of monograms, di-grams and tri-grams, where a novel approach is deployed considering the optimally discriminating words over each class of synonyms that characterize each thematic area on which a text is indexed according to its relevance with semantically similar texts. We describe thoroughly the proposed approach and perform a series of evaluation experiments utilizing an adequate number of text samples from specific thematic areas, such as business, politics, sports, entertainment and technology, intending to attest the potentials of our proposed model to index texts from specific areas.

CCS CONCEPTS

• **Information systems** → **Document topic models**; **Content analysis and feature selection**; *Document structure*; Dictionaries; **Document representation**; **Similarity measures**; Structured text search.

KEYWORDS

Text Indexing, Similarity, Semantics, Synonym.

ACM Reference Format:
Georgios Grigoriadis–Kotsalis, Stavros D. Nikolopoulos, and Iosif Polenakis. 2022. A Synonym-based Approach for the Semantic Indexing of Texts. In *International Conference on Computer Systems and Technologies 2022 (CompSysTech '22), June 17–18, 2022, University of Ruse, Ruse, Bulgaria.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3546118.3546119

CompSysTech '22, June 17–18, 2022, University of Ruse, Ruse, Bulgaria

ACM ISBN 978-1-4503-9644-8/22/06…$15.00
https://doi.org/10.1145/3546118.3546119

1 INTRODUCTION

Text similarity is divided into five main approaches [16], namely, the *Character-based* that utilizes max sub-strings and sequences, the *Term-based*, that determines the distance of two texts based on the distance of its vectors/words, the *Corpus-based* that tackles the semantic similarity between words, the *Knowledge-based* that utilizes the information derived from existing lexical databases(e.g. WordNet), and the *Hybrid Similarity Measures* that are developed over a mixture of the aforementioned approaches. The similarity can be considered as a measurement for the investigation of relations among words or sentences that can be represented through various structures. In particular, semantic similarity measures compute a quantitative heuristic of similarity over a particular representation focusing on ontology-based methods [15]. Across the recent literature [10, 16] the procedure of text indexing (or text classification) is mainly developed initially by the retrieval of data referencing the available data-set of texts,the analysis of the collected data, proceeding to the phase where the features are investigated in order to select the most significant ones, concluding with the states of model training and the procedure of experimental evaluation utilizing the deployed classifier. In particular, as referred in [1], the traditional approaches for text classification utilize mostly the so-called Bag-of-Words method or equivalently the Vector Space model, where the terms (words) are collected and processed independently of their position in the text/sentences (e.g. as word frequencies stored in vectors), omitting hence their semantic aspect. Most importantly, as referred in [1], the traditional text classification approaches exhibit several drawbacks that reduce their accuracy in text categorization, mainly by omitting complex syntactic expressions and not distinguishing among polysemous or synonymous words.

1.1 Related Work

Gu *et. al.* in [5] investigate an efficient algorithm for Dynamic Text Indexing, by pre-processing the text aiming to extract chunks and patterns on substrings, aiming to find all occurrences within a chunk and gather partial occurrences overlapping segment borders. In [19] the authors compare two well-known indexing methods the signature file method and the inverted file method.

Through the literature, several approaches have been proposed to face the procedure of text indexing according to its content. Recently, Hazem and Daille *et. al.* in [6] aim to introduce a new approach for the acquisition of synonyms based on word-embedding. The authors follow two main approaches: a compositional method and a semi-compositional method. The full-compositional method is based upon the idea that phrases can be represented by an element-wise sum of the word embeddings of semantically related words

of its parts, while the semi-compositional one is based on distributional analysis. In [2] Biswas *et. al.* take a closer look at the graph based keyword extraction model and propose the collective node weight as a solution, creating a model named KECNW, meaning that after the pre-cleaning the algorithm takes into consideration Distance from the central node,Selectivity Centrality, Importance of neighboring nodes, Position of a node and Term frequency. In [13] Shehab *et. al.* explore the possibility of an auto-grading system in essay-typed questions using Text Similarity Algorithms. Comparing DL (Damera-Levelnshtein), N-Gram, LSA, Disco 2 with manual grading, the authors split the text categories into short and long answers. The short answers require strictly a semantic approach, while the longer answers require taking more variables into consideration. Ferrada *et. al.* in [4] provide a thorough investigation of Hybrid Indexing improving the time effectiveness and extending its usability to text (not just DNA), as well as comparing it with other compressed indexes in highly repetitive text collections. In [18] Fengqi *et. al.* suggest that word embedding can improve the semantic similarity retrieval, where using the cosine similarity to calculate the semantic distance of two words, the author tries to show the improvement, while using a high threshold for word similarity. In [17] Uysal analyzes the two-stage feature selection methods for text classification, evaluating various classification settings utilizing a linear SVM classifier.

Later, Mallick *et. al.* in [8] propose an effective graph-based text summarization method utilizing a modified version of TextRank, where the representation of the text is achieved through a weighted graph structure where the vertices correspond to the sentences of the text and the weight of the edges among them is computed utilizing a modification of the inverse sentence frequency-cosine similarity. Shahmirzadi *et. al.* in [12] provide a comparative study on automatic measurement of semantic text similarity when measuring the similarity exhibited between patents utilizing vector space models. In [20] Zouaoui and Rezeg propose a multi-agents indexing system for plagiarism detection investigating over this aspect the impact of Arabic ontology on the effectiveness of plagiarism detection. In [7] Mahmoud and Zrigui propose an effective deep learning model utilizing global word embedding and recurrent convolutional neural network (RCNN) in order to represent the contextual dependencies between words-vectors, considering their semantic meanings.

Recently, Chamidah *et. al.* in [3] explore the ability to grade Short Essays using Word Expansion, a method used for improving accuracy in text-based classification, using as string-based similarity methods the Coefficient provided via the Cosine, the Jaccard and the Dice methods. The authors compare the similarities with or without the use of the word expansion grading them using the Pearson Correlation for 4 separate categories of text (Tech,Lifestyle,Politics and Sports), while the results of the experiment prove that Word Expansion is indeed a valid approach. Singh *et. al.* in [14] attempt to enhance an existing method for streaming Similarity Search. The algorithm in review is called Approximate Nearest Neighbor Search (ANNS) which is the go-to method when dealing with a fundamental building block in information retrieval with graph-based indices, while the proposed enhancement is called FreshDiskANN and is using a two-pass algorithm called Streaming-Merge, which merges the in-memory index with the SSD-index efficiently and proportionally to the time and space complexity of the merge.

1.2 Our Approach

In this work, we design and implement an integrated framework for text indexing based on its content. Throughout our approach, we aim to investigate the utilization of synonym-based text abstraction where the semantics of the text are retained through a paraphrase procedure by substituting over each sentence of the text each word term in the text with another word, that is the representative of its synonym class. We investigate the effect of preserving the multiplicity of each term utilized in the text considering the uniqueness or the multiplicity on the appearance of each term. Moreover, we utilize a set of pre-processing approaches concerning the omission of words of length that is less than four letters, i.e., monograms, di-grams and tri-grams, the computation of the most popular words utilized among each text category resulting therefore to the optimally discriminating word terms, augmenting the similarity measurement between pairs of texts by adding a positive score over the presence of word terms that are uniquely used over each category that potentially could characterize the relevance of text content.

We explore a set of known similarity metrics, proposing additionally one metric that fitted the requirements of text indexing, and evaluate the proposed model through various settings deploying combinations of the proposed text pre-processing procedures over each similarity metric. The contribution resulting from the development of our approach is mostly derived from the provision of an integrated framework for text indexing concerning its content deploying a set of various text pre-processing techniques and similarity metrics. Moreover, throughout the design of our proposed approach, we designed and implemented a similarity metric that fulfills the requirements posed throughout the concept of text indexing. Moreover, we provide a thorough evaluation of our proposed model for text indexing and examine its potentials to index articles into specific categories according to the content.

1.3 Road-map

In Section 2 we discuss the theoretical background behind the representation of the text objects, the methodology deployed for the utilization of the classes of synonyms in order to abstract the content of the text and the utilization of the underlying similarity metrics deployed to measure the similarity among the term-vector representations of text objects. In Section 3 we present the proposed model regarding the proposed approach on synonym-based text indexing, the integration of the proposed approach into a unified framework for text indexing, as also the architecture of the proposed system. Next, in Section 4, we provide an evaluation of our proposed model conducting a series of evaluation experiments in order to present the potentials of our proposed approach in text indexing, we discuss the experimental design followed, the achieved results, as also the analysis that reveal further insights about the accuracy of the proposed approach. Finally, in Section 5, we conclude our work providing a further discussion over the aspects of the proposed model, its potentials and its limitations, and cite our aims for future work.

2 THEORETICAL BACKGROUND

In this section, we discuss the theoretical background that consists the basis of our proposed approach, and present the techniques utilized to represent and measure the similarity over the structure of the text objects.

2.1 Text Representation and Paraphrasing

In our approach, in order to represent text objects, we follow the traditional Bag-of-Words method by aggregating the word terms appearing in the text object under consideration and performing a set of pre-processing and post-processing techniques in order to filter out elements that are valuable for the text indexing procedure. As we will discuss later, word terms of minor importance (e.g. monograms, di-grams, and tri-grams) or words commonly used and are not indicative of a special content category are excluded and the remaining of the terms are stored into a `term-vector`.

This object is then processed again, throughout the text preprocessing procedure that we will discuss more extensively in later sections, in order to be transformed into a state that matches the particular requirements of the deployed similarity metrics utilized for the indexing of the text. Throughout this approach, all the terms are collected into a unified term-vector focusing mostly on the significant points of the text under consideration, i.e., the utilization of specific words, rather than their relations exhibited among them across their co-existence in the same sentence.

Throughout the text representation procedure, in our approach, we focus on the abstraction of the text objects in order to achieve the preservation of the semantics of the text by the aspect of its content as it is expressed by the word terms utilized across its extent and additionally eliminate the redundant information resulting by synthetic or commonly used words. The main idea behind this approach comes from a similar work deployed for the classification of malicious samples [11], where the System-call functions are substituted by their corresponding group developed based on their functional commonalities, where similarly to this approach we consider that a text in order to retain the semantics of its content should utilize terms from the same categories, i.e., similar words, expressed by the synonym classes. Thus, regarding the case where the same text content can be expressed by the utilization of synonym word terms retaining its semantics, we adopt the approach of text paraphrasing by utilizing the synonyms class of each word term and replacing each term with the `representative term` of its class of synonyms it belongs to.

Thus, considering the `term-vector` that represent the text object under consideration, we perform a post-processing procedure, by linearly processing its contents and term-by-term we replace each element with the `representative term` of the class of synonyms it belongs to, resulting to a paraphrasing of the content of the text object. In our approach, for the utilization of the synonym classes, we deployed the `WordNet` lexical database [9], in order to retrieve the list of word-terms that belong to the same synonym class, selecting as the `representative term` of each class the term that appears lexicographically first in the corresponding synonym list. To this point, we should recall that the `term-vector` that results after the substitution of each word-term by the `representative term` of the synonym class it belongs to, may include multiple or unique occurrences of each term, considering the similarity metric that will be deployed later, in order for a pair of text objects to be compared.

2.2 Comparing Term-Vectors

In our approach, in order to measure the similarity between a pair of `term-vectors`, we utilize a set of similarity metrics, that rely either only on the existence of each word-term (i.e., term uniqueness), that operate similarly over vectors that contain bits, or on the weight of each word-term (i.e., term multiplicity), that operate similarly over vectors that contain integers. In each case, we investigated various types of similarity metrics, including the utilization of the Cover similarity metric developed to suit the requirements and the settings of text indexing presented in this work, the Jaccard Index, the Tanimoto Coefficient, the Cosine similarity, and the computation of Cross Entropy similarity. Next, we discuss the similarity metrics utilized in our model to measure the similarity between a pair of `term-vectors`, let $V_{\tau_1}[\cdot]$ and $V_{\tau_2}[\cdot]$ representing the text objects τ_1 and τ_2, respectively, containing the word-terms that are denoted as e, appearing over the corpus of τ_1 and τ_2.

Regarding the first type of similarity metrics, i.e., the ones that investigate only the existence of each term regardless of its multiplicity, based on the Jaccard Index, we defined the Cover similarity metric for the computation of the similarity among non foreign sets regarding a pair of `term-vectors`, which is computed as:

$$Cv(V_{\tau_1}, V_{\tau_2}) = \frac{|e \in V_{\tau_1} \wedge e \notin V_{\tau_2})| + |e \in V_{\tau_2} \wedge e \notin V_{\tau_1})|}{|e \in V_{\tau_1}| + |e \in V_{\tau_2}|} \quad (1)$$

The Jaccard Index computes the similarity between `term-vectors` regarding the rate of the number of word-terms existing in both text objects over the number of word-terms existing either in the one or in the other, or in both text objects as follows:

$$J(V_{\tau_1}, V_{\tau_2}) = \frac{|e \in (V_{\tau_1} \cap V_{\tau_2})|}{|e \in V_{\tau_1}| + |e \in V_{\tau_2}| - |e \in V_{\tau_1} \cap V_{\tau_2}|} \quad (2)$$

Accordingly, the Tanimoto Coefficient between two `term-vectors` is calculated as:

$$T(V_{\tau_1}, V_{\tau_2}) = \frac{\sum_{\ell=1}^{k} (V_{\tau_1}[\ell] \times V_{\tau_2}[\ell])}{\sum_{\ell=1}^{k} V_{\tau_1}[\ell]^2 + \sum_{\ell=1}^{k} V_{\tau_2}[\ell]^2 - \sum_{\ell=1}^{k} (V_{\tau_1}[\ell] \times V_{\tau_2}[\ell])}, \quad (3)$$

where ℓ denotes the index of each term and k is the size of the corresponding `term-vectors` V_{τ_1} and V_{τ_2} representing texts τ_1 and τ_2, respectively.

For the second type of similarity metrics, i.e., the ones that investigate the multiplicity (in terms of weight, denoted as $w(e)$) of each term, the Cosine similarity computes the similarity between a pair of `term-vectors` as follows:

$$C(V_{\tau_1}, V_{\tau_2}) = \frac{\sum_{\ell=1}^{k} w(e_\ell \in V_{\tau_1}) \times w(e_\ell \in V_{\tau_2})}{\sqrt{\sum_{\ell=1}^{k} w(e_\ell \in V_{\tau_1})^2} \sqrt{\sum_{\ell=1}^{k} w(e_\ell \in V_{\tau_2})^2}}, \quad (4)$$

where $w(e_\ell \in V_{\tau_1})$ represents the number of multiple occurrences of the ℓ-th word-term e_ℓ in registered in V_{τ_1}, or, equivalently, $w(e_\ell \in V_{\tau_1}) = |e_\ell \in V_{\tau_1}|$.

Finally, another metric we utilized in our model focusing on the measurement of the frequency of the words appearing in both texts, i.e., a text sample under consideration and a categorized one, is the cross-entropy which is defined as follows:

$$E(V_{\tau_1}, V_{\tau_2}) = \frac{\sum_{\ell=1}^{k} -log(\frac{|e_\ell \in V_{\tau_1}|}{|V_{\tau_1}|} + \frac{|e_\ell \in V_{\tau_2}|}{|V_{\tau_2}|})}{|V_{\tau_1} \cap V_{\tau_2}|} : e_\ell \in V_{\tau_1} \cap V_{\tau_2}, \quad (5)$$

where $\frac{|e_\ell \in V_{\tau_i}|}{|V_{\tau_i}|}$ denotes the frequency that the word term e_ℓ appears across the corpus of text V_{τ_i}

3 THE MODEL

In this section, we present the deployment of the synonym-based text indexing procedure and discuss the integration of our proposed model, providing further insights about the architecture of the derived framework.

3.1 Synonym-based Text Indexing

The integration of the proposed approach has its basis on the cooperation of the procedures deployed for the processing of the text sample under consideration, namely, the pre-processing, the post-processing, and the grading procedure. Initially, during the *pre-processing* phase of the deployment of our model, we transform every character in the text under consideration into lower case, so that the consistency among the texts is retained. Then, we replace each word with the representative word of the synonym group it belongs to. To this point, it is worth noting that, according to the grading procedure following next, we can distinguish between the retaining of the multiplicity and the uniqueness of the terms appearing in the text, since particular similarity metrics require the multiplicity of the words (i.e., a weighted object), while others require only the existence of each term (i.e., unique appearance) in order to handle such objects. Hence, it is necessary to distinguish such cases during the pre-processing phase before we proceed to the next phases. Next, through the *post-processing* phase we deploy the following techniques:

- **Trimming Small Words.** Through this technique we remove (or "`trim`") the monograms, di-grams, tri-grams, (i.e., word terms of length $\ell \leq 3$, e.g., "a", "the", "of", etc.) since most do not affect directly the semantics of the text.
- **Filtering Non-Characterizing Terms.** Utilizing this technique we proceed with a procedure similar to the dimensionality reduction, filtering out and removing specific word terms that appear in all text samples of several categories, do not consisting thus characterizing elements for the classification process. In particular, we compute the word terms that appear in more than 80% of the text samples inside each text category, the so-called `Non-Characterizing Terms` (i.e., *Most Popular Words*) and retain the ones that do not appear in any other category's `Terms`, the so-called `Optimally Discriminating Terms`.
- **Emphasize on Optimally Discriminating Terms.** By this technique an addition is retained to be added then over the computation of the similarity metric between the text sample under consideration and texts of this category, in order to point out its likeness to this category if the corresponding `Optimally Discriminating Terms` are detected inside the terms of the text under consideration.

Finally, during the grading procedure deployed in order to compute the similarity between a given text sample and known categorized text samples as to index the text sample under consideration into a category of texts, we utilize the similarity metrics presented in Section 2.2 investigating the commonalities exhibited throughout the structural characteristics of a pair of text samples with respect to the word terms utilized as they are depicted by their corresponding `term-vectors`. To this point, it is important to note that, regarding the post-processing procedure conducted over the derived `term-vectors`, in this approach there are examined several combinations over the processing techniques deployed, i.e., there are investigated the initial accuracy of the deployed similarity metrics without any post-processing procedure conducted. Moreover it is investigated the effect caused by the removal of the monograms, digrams and tri-grams, the effect of the `Non-Characterizing Terms` and the computation of an increase in the exhibited similarity value added in presence of common `Optimally Discriminating Terms` between the text sample under consideration and a text sample from a specific category, as also combinations of these post-processing techniques. Then, regardless of the grading procedure utilized to compute the similarity value between a given text sample and the categorized text samples, in this approach the text sample under consideration is indexed into the category that contains the categorized text sample that exhibited the maximum similarity value among all the text samples from all categories.

3.2 System Architecture and Deployment

The proposed model for synonym-based text indexing incorporates the pre-processing, the post-processing, and the grading procedures discussed in Section 3.1 in order to conduct the synonym-based text indexing. In Figure 1, we present an illustrative view of the overall architecture of our proposed model and discuss the procedural flow followed over its deployment. Next, it is discussed step-by-step the overall procedure followed by our model. Once a text sample is about to be indexed into a specific category by our proposed system, in the first step, a set of pre-processing techniques are deployed in order to construct for the given text sample its corresponding `term-vector` that includes the word terms appeared across its corpus. In the proposed approach, WordNet is utilized in order to achieve the development of a proper set of synonym classes and conduct a paraphrasing procedure in order to construct an abstract representation of the content of the text sample with respect to the contained word terms. In the proposed model, we selectively retain either multiple or unique instances of the utilized word-terms, regarding the similarity metric that is computed in the next steps, to produce single or weighted `term-vectors`, respectively.

Next, throughout the post-processing procedure, there are applied various filtering techniques in order to refine the content of the `term-vectors` and augment the information provided regarding the content of the text sample under consideration that is represented. Namely, the post-processing procedures deployed are the removal (or "`trimming`") of the monograms, di-grams, and tri-grams, the filtering of the `Non-Characterizing Terms`, and the

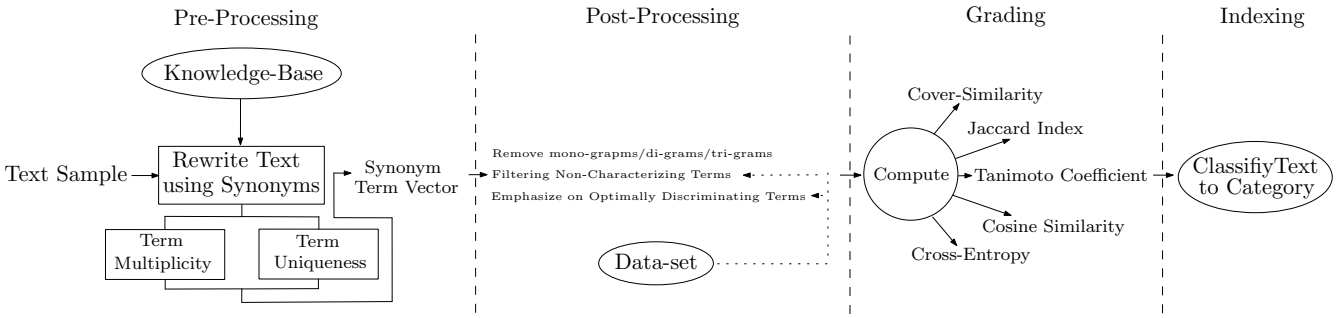

Figure 1: Architecture and work-flow of the proposed model.

emphasis through an additive value selected to be added on the similarity measurement in presence of `Optimally Discriminating Terms` in both the categorized text and the text under consideration, augmenting thus the indexing procedure depicting a strong set of characteristics that indicate directly the category of a text. To this point, we ought to underline that the last two post-processing procedures have their basis on the training procedure achieved through computations performed over the data-set that constitutes the knowledge base of text categories alongside their text samples.

Having already refined the derived single or weighted synonym-based `term-vectors` we proceed with the similarity measurement between the `term-vector` that represent the text sample under consideration and all the `term-vectors` that represent the categorized text samples in our knowledge base. In the proposed model, for the case of single `term-vectors`, i.e., each word term is depicted as an element in the `term-vectors` only once, it is utilized the Cover-similarity, or the Jaccard Index, or the Tanimoto Coefficient, while for the case of weighted `term-vectors`, i.e., each word term is depicted as element in the `term-vectors` multiple times depicting the number of its occurrences in the corpus of the text, there are utilized the Cosine Similarity or the Cross-Entropy similarity metrics. Finally, in order to index the text sample under consideration into one of the text categories from our knowledge base, the text sample is classified into to category that includes the text sample that exhibited the maximum similarity metric with the test sample.

4 EVALUATION

In this section, it is presented the evaluation of the proposed model on the indexing of given text samples into specific categories.

4.1 Experimental Design

For the evaluation of our proposed model for synonym-based text indexing it is utilized a data-set of 500 articles from BBC pre-classified into five thematic categories according to their content, namely *Sports, Politics, Entertainment, Business,* and *Technology* including 100 articles per category. Through the evaluation experiments it is compared each text sample from every category against the rest 499 samples, i.e, 99 samples from the same category and 400 articles from the rest four categories. In order to evaluate the accuracy of the proposed model there are conducted several evaluation experiments attesting the accuracy of each similarity metric deployed,

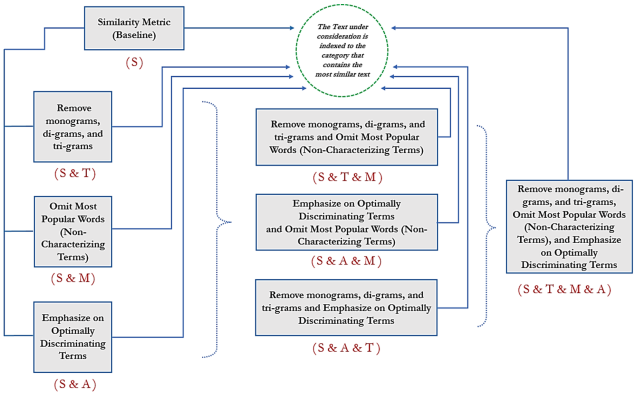

Figure 2: Experimental setup followed for the evaluation of the text indexing accuracy of the proposed model.

as also the effectiveness and the potentials of the corresponding post-processing techniques. The experimental setup is partitioned into two series of experiments. In the first series of experiments, there is investigated the indexing accuracy of each similarity metric when applied directly to the `term-vector` without applying any post-processing technique. Next, regarding the post-processing procedures it is conducted a second series of experiments in order to investigate the impact of each `term-vector` filtering technique and explore further insights of them on the indexing accuracy of the proposed model.

In Figure 2 it is illustrated the design followed for the evaluation of the indexing accuracy of the proposed model. In Figure 2 there are both depicted the flows over the two series of experiments, i.e., the utilization of the similarity metric with no post-processing techniques, and the utilization of the post processing techniques one-by-one utilizing only one in each experiment for each similarity metric, utilizing their two-by-three combinations and finally all of them. In particular, each case is denoted by the a representative letter for each approach, i.e., `S` for the utilization of a similarity metric, `S&T` for the deployment of the "trimming" of mono-grams, di-grams, and tri-grams, `S&M` for the removal of the `Most Popular Words` (i.e., `Non-Characterizing Terms`), and `S&A` for the incorporation the advantage in similarity value in presence of common significant word terms (i.e., `Optimally Discriminating Terms`),

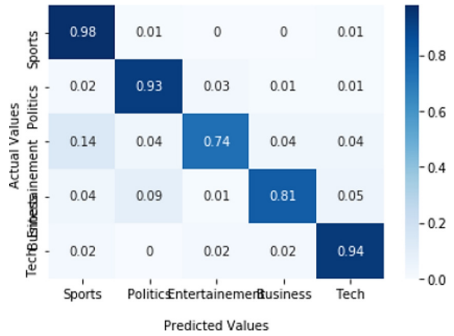

Figure 3: Confusion matrix of indexing accuracy deploying Cover similarity metric.

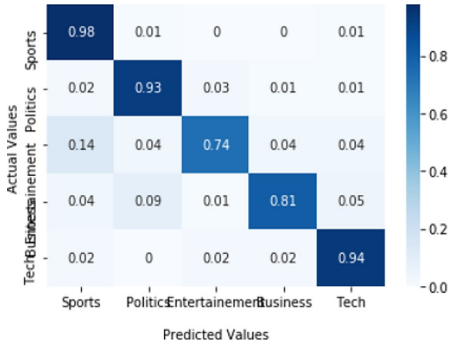

Figure 4: Confusion matrix of indexing accuracy deploying the Jaccard similarity metric.

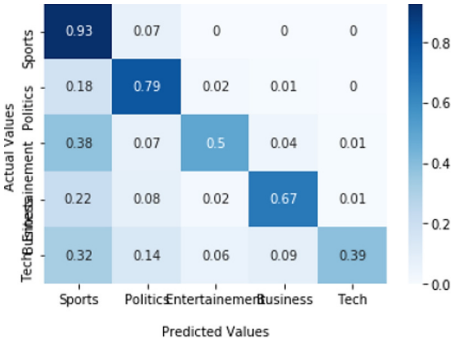

Figure 5: Confusion matrix of indexing accuracy deploying the Tanimoto similarity metric.

while S&T&M, S&A&M, S&A&T denote their two-by-three combinations and S&T&M&A denotes the utilization of all three post-processing techniques.

4.2 Discussion over the Exhibited Results

In Figures 3–7 there are presented the confusion matrices depicting the indexing results for each similarity metric deployed without the utilization of any post-processing techniques i.e., the baseline, over the unprocessed `term-vectors`. In particular, in Figure 3 there

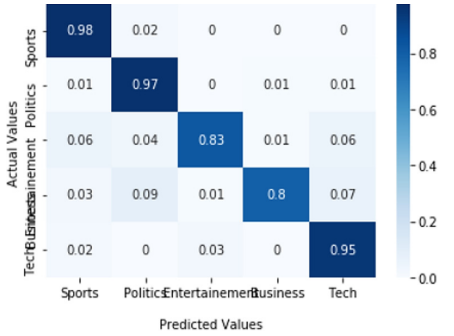

Figure 6: Confusion matrix of indexing accuracy deploying the Cosine similarity metric.

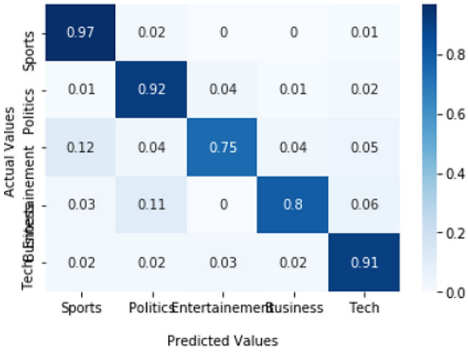

Figure 7: Confusion matrix of indexing accuracy deploying the Cross-Entropy similarity metric.

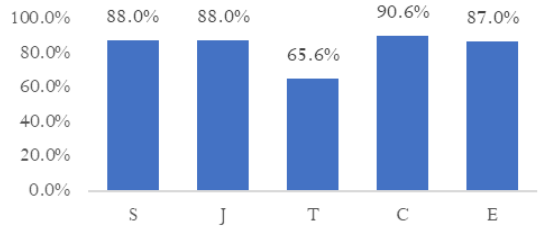

Figure 8: Comparative indexing results averaged over the five categories deploying only each similarity metric with no post-processing techniques.

are presented the indexing results achieved utilizing the Cover-Similarity, in Figure 4 there are presented the indexing results achieved utilizing the Jaccard Index, in Figure 5 there are presented the indexing results achieved utilizing the Tanimoto Coefficient, in Figure 6 there are presented the indexing results achieved utilizing the Cosine similarity, while in Figure 7 there are presented the indexing results achieved utilizing and the Cross-Entropy similarity metric. As we can observe through the achieved indexing results, our proposed model obtained a decent classification accuracy over each text category in all the similarity metrics deployed, where in many cases the variance was almost eliminated. Indicative examples

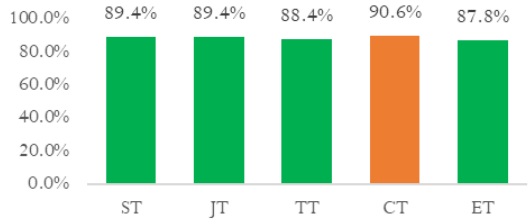

Figure 9: Comparative indexing results averaged over the five categories deploying each similarity metric and removing the monograms, di-grams, and tri-grams (i.e., `Trimming Small Words`).

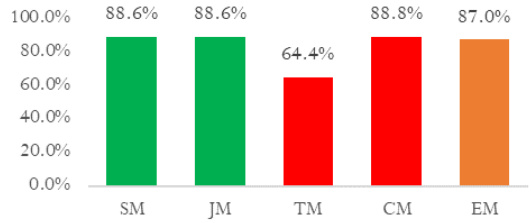

Figure 10: Comparative indexing results averaged over the five categories deploying each similarity metric and filtering the most popular words (i.e., `Omitting Non-Characterizing Terms`).

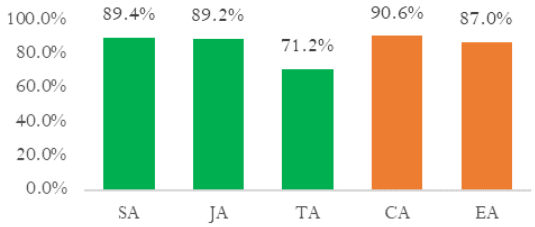

Figure 11: Comparative indexing results averaged over the five categories deploying each similarity metric and emphasizing on characterizing terms incorporating the advantage in similarity value in presence of common significant words (i.e., `Emphasizing on Optimally Discriminating Terms`).

of high classification accuracy are demonstrated by the results achieved over the categories of *Sports* and *Technology*, where on the other hand, categories such *Politics*, *Entertainment*, and *Business* seem to exhibit higher contextual overlap, which is a fact that is depicted over a decrease on the corresponding indexing results.

In Figures 8–15 there are demonstrated the exhibited results achieved through the second series of experiments where the post-processing techniques were utilized, averaged over the five categories. In Figure 8 there are demonstrated the achieved indexing results when deploying no post-processing techniques (i.e., baseline) where each similarity metric is denoted by its first initial letter.

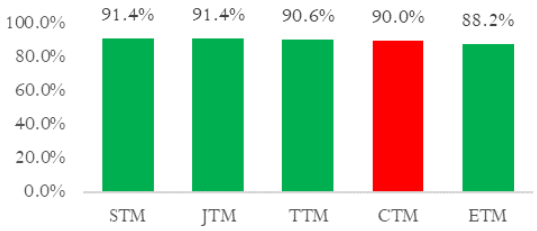

Figure 12: Comparative indexing results averaged over the five categories deploying each similarity metric, `Trimming Small Words` and `Omitting Non-Characterizing Terms`.

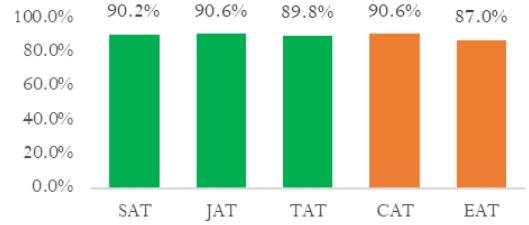

Figure 13: Comparative indexing results averaged over the five categories deploying each similarity metric, `Trimming Small Words` and `Emphasizing on Optimally Discriminating Terms`.

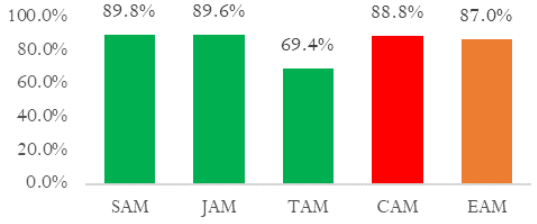

Figure 14: Comparative indexing results averaged over the five categories deploying each similarity metric, `Omitting Non-Characterizing Terms` and `Emphasizing on Optimally Discriminating Terms`.

As we can observe from the corresponding indexing results averaged over the five categories presented in Figure 8 depicting the accuracy of the baseline, i.e., the utilization of the similarity metric without post-processing techniques, the Cosine similarity exhibits the maximum accuracy achieving a 90.6% correct classifications.

In Figures 9–15, the exhibited results are marked with red, orange and green colors, denoting the reduced, the equal and the improved results achieved by the post-processing techniques incorporated in each case, respectively. In Figure 9 there are demonstrated the achieved indexing results when deploying each similarity metric along side the "trimming" post-processing technique denoted as the first initial letter of the similarity metric, let `S` and `T`, in which compared to the baseline (see, Figure 8) we observe that in all the cases the results exhibit an improvement. On the other hand,

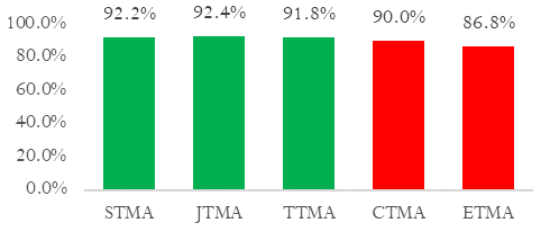

Figure 15: Comparative indexing results averaged over the five categories deploying each similarity metric, `Trimming Small Words`**,** `Omitting Non-Characterizing Terms` **and** `Emphasizing on Optimally Discriminating Terms`**.**

when deploying the post-processing techniques of the utilization of the `Optimally Discriminating Terms` and the advantage in similarity value in presence of common significant word term (see Figures 10 and 11) only the Cover, the Jaccard and the Tanimoto similarity metrics exhibit an improvement.

Next, regarding the deployment of the two-by-three combinations of the post-processing techniques, as we can observe from Figures 12, 13, and 14, respectively, again the Cover, the Jaccard and the Tanimoto similarity metrics exhibit a significant improvement compared to the baseline (see, Figure 8). Finally, concerning the utilization of all three post-processing techniques, as we can observe from Figure 15, the proposed Cover similarity and the Jaccard similarity metrics exhibit both a maximum indexing rate, i.e., 92.2% and 92.4%, respectively, which validates the potentials of the underlying post-processing techniques deployed to augment the indexing procedure.

5 CONCLUSION

In this work we presented an integrated model for synonym-based text indexing, utilizing several pre-processing and post-processing techniques and various similarity metrics in order to investigate the potentials of our approach to indexing texts into specific categories. In order to evaluate the abilities of our model we conducted a series of evaluation experiments utilizing a data-set of 500 articles from BBC pre-classified into five distinct thematic achieving a 92.4% classification accuracy, proving the potentials of our proposed synonym-based approach in text indexing.

The decent results exhibited highlight the potentials of our proposed synonym-based indexing approach, however, further investigation should be conducted over the semantic relations exhibited among the word terms inside the sentences of the text and across the greater extent of its corpus. Our aims for future research, considering further improvement on the exhibited classification rates, are focusing to the representation of these sentence-based semantic relations of these words incorporating a graph-based approach that could represent the abstract structure of the text samples.

ACKNOWLEDGMENTS

We acknowledge support of this work by the project "Dioni: Computing Infrastructure for Big-Data Processing and Analysis." (MIS No. 5047222) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Co-financed by Greece and the European Union

REFERENCES

[1] Berna Altınel and Murat Can Ganiz. 2018. Semantic text classification: A survey of past and recent advances. *Information Processing & Management* 54, 6 (2018), 1129–1153.

[2] Saroj Kr Biswas, Monali Bordoloi, and Jacob Shreya. 2018. A graph based keyword extraction model using collective node weight. *Expert Systems with Applications* 97 (2018), 51–59.

[3] Nurul Chamidah, Mayanda Mega Santoni, Helena Nurramdhani Irmanda, Ria Astriratma, Lomo Mula Tua, and Trihastuti Yuniati. 2021. Word Expansion using Synonyms in Indonesian Short Essay Auto Scoring. In *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS.* IEEE, 296–300.

[4] Héctor Ferrada, Dominik Kempa, and Simon J Puglisi. 2018. Hybrid indexing revisited. In *2018 Proceedings of the Twentieth Workshop on Algorithm Engineering and Experiments (ALENEX).* SIAM, 1–8.

[5] Ming Gu, Martin Farach, and Richard Beigel. 1993. An Eficient Algorithm for Dynamic Text Indexing. (1993).

[6] Amir Hazem and Béatrice Daille. 2018. Word embedding approach for synonym extraction of multi-word terms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

[7] Adnen Mahmoud and Mounir Zrigui. 2021. Semantic similarity analysis for corpus development and paraphrase detection in arabic. *Int. Arab J. Inf. Technol.* 18, 1 (2021), 1–7.

[8] Chirantana Mallick, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das, and Apurba Sarkar. 2019. Graph-based text summarization using modified TextRank. In *Soft computing in data analytics.* Springer, 137–146.

[9] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[10] Marcin Michał Mirończuk and Jarosław Protasiewicz. 2018. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications* 106 (2018), 36–54.

[11] Stavros D Nikolopoulos and Iosif Polenakis. 2015. Malicious software classification based on relations of system-call groups. In *Proceedings of the 19th Panhellenic Conference on Informatics.* 59–60.

[12] Omid Shahmirzadi, Adam Lugowski, and Kenneth Younge. 2019. Text similarity in vector space models: a comparative study. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA).* IEEE, 659–666.

[13] Abdulaziz Shehab, Mahmoud Faroun, and Magdi Rashad. 2018. An automatic Arabic essay grading system based on text similarity Algorithms. *International Journal of Advanced Computer Science and Applications* 9, 3 (2018), 263–268.

[14] Aditi Singh, Suhas Jayaram Subramanya, Ravishankar Krishnaswamy, and Harsha Vardhan Simhadri. 2021. FreshDiskANN: A Fast and Accurate Graph-Based ANN Index for Streaming Similarity Search. *arXiv preprint arXiv:2105.09613* (2021).

[15] Luke T Slater, Sophie Russell, Silver Makepeace, Alexander Carberry, Andreas Karwath, John A Williams, Hilary Fanning, Simon Ball, Robert Hoehndorf, and Georgios V Gkoutos. 2022. Evaluating semantic similarity methods for comparison of text-derived phenotype profiles. *BMC medical informatics and decision making* 22, 1 (2022), 1–12.

[16] P Sunilkumar and Athira P Shaji. 2019. A survey on semantic similarity. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3).* IEEE, 1–8.

[17] Alper Kursat Uysal. 2018. On two-stage feature selection methods for text classification. *IEEE Access* 6 (2018), 43233–43251.

[18] Fengqi Yan, Qiaoqing Fan, and Mingming Lu. 2018. Improving semantic similarity retrieval with word embeddings. *Concurrency and Computation: Practice and Experience* 30, 23 (2018), e4489.

[19] Justin Zobel, Alistair Moffat, and Kotagiri Ramamohanarao. 1998. Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems (TODS)* 23, 4 (1998), 453–490.

[20] Samia Zouaoui and Khaled Rezeg. 2020. Multi-agents indexing system (MAIS) for plagiarism detection. *Journal of King Saud University-Computer and Information Sciences* (2020).