



www.editada.org

Method for the Evaluation of Similarity Measures using short Texts

Maricela Bravo, Luis Fernando Hoyos Reyes, Domingo Rodríguez Benavides, Leonardo D. Sánchez-Martínez

¹ Universidad Autónoma Metropolitana

mcbc@azc.uam.mx, hrlf@azc.uam.mx, dorobe@azc.uam.mx, ldsm@azc.uam.mx

Abstract. There exist multiple online collections and data bases of scientific articles publicly available, to take full advantage of these resources, it is necessary to process, arrange and correlate texts with respect to a classification or ontology. To achieve an efficient organization and a more relevant correlation between texts, it is necessary to use a similarity measure for short texts. However, determining the best method to calculate the similarity between texts is an arduous task, since there are many similarity measures reported in literature. Additionally, the collection of texts to which the similarity measures are applied should be considered; while some measures are useful for some types of information sources, they fail when the collection of data changes. Therefore, it is necessary to count with a method to evaluate the performance of similarity measures from a statistical perspective and in terms of the accuracy achieved by each measure.

Keywords: Similarity measures, short texts comparison, scientific publishing, evaluation of similarity measures.

Article Info

Received Sep 26, 2022

Accepted Jan 11, 2023

1 Introduction

In the study and development of methods for text processing, one of the most frequent tasks is the calculation of similarities or distances between texts. For example, in the Information Retrieval area, there is a need to correlate research articles with some topic or topics to facilitate document search. The objective of applying these similarity measures is to establish the distance, either semantic or syntactic, between texts. In this article we focus on investigating which similarity measures offer better results considering a set of research articles published on Computer Science in the Semantic Scholar, as well as Computer Science Ontology (CSO)¹. It is important to consider that there are many similarities measures applicable for short texts; among these, the semantic measures based on WordNet that are used and analyzed in this paper are: Wu and Palmer [2], Jiang and Conrath [3], PATH [4], Lin [5], and Resnik [6], additionally the well-known Cosine Similarity distance.

In 2009 [1], there was an estimation of 50 million published research articles. According to the National Science Board of the USA, scientific publications around the world grow at a rate of 4% each year, it is estimated that there are around 30,000 scientific journals, and that approximately two million articles are published each year. Therefore, the task of calculating similarities between research papers and topics represents a problem of dimensionality and scalability.

The problem can be formulated as follows:

Given a collection of P publications, an ontology of topics T , a set of similarity measures S , and a set of evaluation criteria E defined as the 4-tuple $\langle P, T, S, E \rangle$

Where:

P represents a collection of i publications $P = \{p_1, p_2, p_3, \dots, p_i\}, i > 0$

T represents a topics taxonomy arranging j topics $T = \{t_1, t_2, t_3, \dots, t_j\}, j > 0$

S represents a set of k similarity measures $S = \{s_1, s_2, \dots, s_k\}, k > 0$

E represents the evaluation criteria $E = \{e_1, e_2, \dots, e_l\}, l > 0$

The objective is to identify the measurement or measurements that return a better result considering the evaluation criteria.

The number of similarity calculations is defined by $\text{calc} = i * j * k$, and the time required for this number of calculations it is necessary to incorporate the time of execution of each similarity measure, $\text{time} = i * j * (k * \text{time}(k))$. Where $\text{time}(k)$ depends on

¹ <https://cso.kmi.open.ac.uk/home>

the algorithm and computational resources involved in the similarity calculations. Supposing that there is a collection of 2000 publications, 2000 topics, and six similarity measures, the number of calculations is $\text{calc} = 2000 * 2000 * 6$, that is 24 million of calculations multiplied by the time required for each of the six similarity measures. Taking into consideration that more than one evaluation criteria can be used for the selection of similarity measures, the research problem is divided into two objectives.

- a) Create a method to reduce the number of similarity calculations and the time required for their execution.
- b) Define an evaluation function which incorporates the criteria to select the best similarity measure.

In this paper, a method that addresses the afore-mentioned objectives is presented, the goal of this method is to provide evidence for decision making.

2 Description of the Method

Aiming at reducing the number of calculations and time required for the evaluation of similarity measures, the proposed method (shown in Figure 1) consists of the following steps:

- a) Collect input data: select a collection of research publications to work with and select an ontology of topics.
- b) Preprocess the texts: it is important to verify that the titles texts of the publications are not empty, that they do not contain badly formatted characters, and create a bag of words representation for the texts.
- c) Select the similarity measures: the set of similarity calculations are well-documented methods for short texts similarity assessment. However, over the last decade numerous methods of comparison between texts have been reported, among these are those based on the use of the WordNet dictionary. It is important to select the measures that are of interest for the calculation, as these calculations require computational resources.
- d) Experimentation: determine the sample size and randomly select samples of publication titles and topics from the taxonomy to execute the calculations. Repeat the experiment n times.
- e) Evaluation: determine the evaluation criteria that will be used and formulate an evaluation function that incorporates the criteria and possible weighs. Apply the evaluation function and concentrate results to determine the similarity measure that best fits the evaluation criteria.

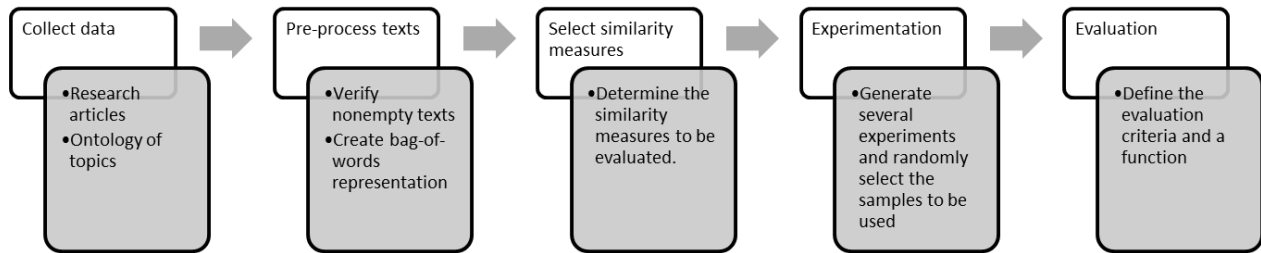


Fig. 1. Proposed method for similarity measures evaluation.

3 Data Collections

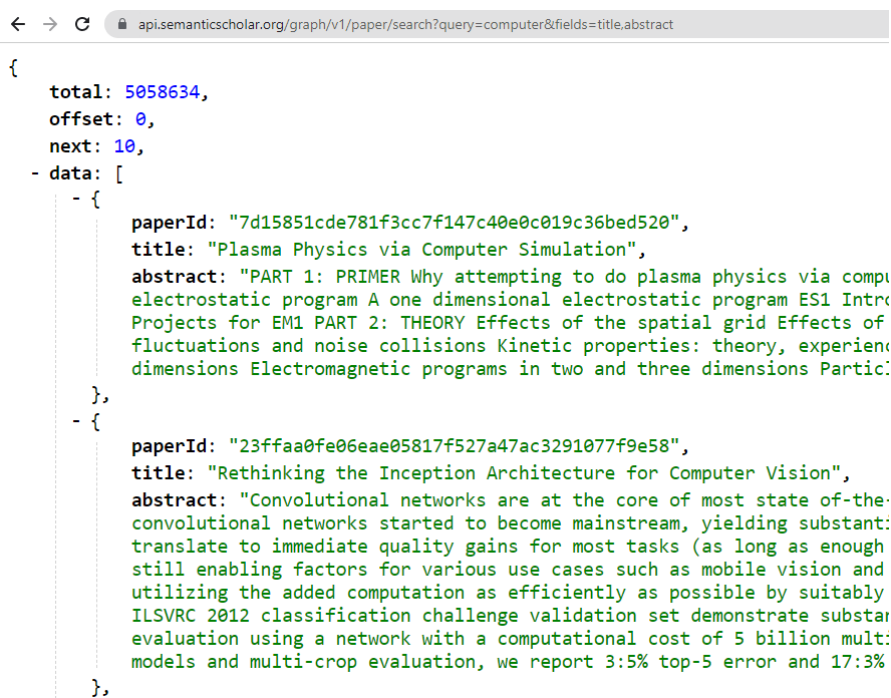
One of the fundamental requirements for calculating similarities is to obtain a very good source of data, together with a classification, taxonomy, or ontology of topics. In this case, we retrieve the data from articles published in the Computer Science domain. Likewise, we use the CSO ontology that provides an organization of topics about Computer Science.

3.1 Research Publications

Regarding the scholarly articles and research publications there are many bibliographic databases available online, we decided to use the Semantic Scholar API², which allows downloading collections of publications by specifying keywords. The execution of the API to obtain a set of articles was configured to retrieve the *id* and the *title* of the publications. It is also possible to recover the abstract and authors of each article; however, as the objective of the experiment was to determine the measurement or measurements that return better results, we only used the titles. In a later phase of classification and correlation of the publications with the topics, we will obtain the abstracts to execute the similarity processing. It is important to point out that some of the similarity measurements that we are using take more computational resources during the calculation, and this is the main reason to first develop a method that supports the decision making to determine the similarity measurements that yield the

² <https://www.semanticscholar.org/>

best results. To develop this method we considered important a light execution, for which we have implemented a method based on statistical sampling, in such a way that titles and topics are taken randomly. Figure 2 presents an example of a request to the Semantic Scholar API, and the JSON formatted response generated.



```

{
  total: 5058634,
  offset: 0,
  next: 10,
  - data: [
    - {
      paperId: "7d15851cde781f3cc7f147c40e0c019c36bed520",
      title: "Plasma Physics via Computer Simulation",
      abstract: "PART 1: PRIMER Why attempting to do plasma physics via comput
electrostatic program A one dimensional electrostatic program ES1 Intro
Projects for EM1 PART 2: THEORY Effects of the spatial grid Effects of
fluctuations and noise collisions Kinetic properties: theory, experienc
dimensions Electromagnetic programs in two and three dimensions Partic:
    },
    - {
      paperId: "23ffaa0fe06eae05817f527a47ac3291077f9e58",
      title: "Rethinking the Inception Architecture for Computer Vision",
      abstract: "Convolutional networks are at the core of most state-of-the-
convolutional networks started to become mainstream, yielding substant:
translate to immediate quality gains for most tasks (as long as enough
still enabling factors for various use cases such as mobile vision and
utilizing the added computation as efficiently as possible by suitably
ILSVRC 2012 classification challenge validation set demonstrate substar
evaluation using a network with a computational cost of 5 billion mult:
models and multi-crop evaluation, we report 3:5% top-5 error and 17:3%
    },
  ],
}

```

Fig. 2. JSON response generated with Semantic Scholar API for paper lookup.

3.2 Representation of Topics

Regarding the representation of topics, there are various representation structures, these range from the use of databases, classifications, taxonomies, knowledge bases, ontologies to knowledge graphs.

According with [7] a taxonomy can be defined as a structured set of names and descriptions used to organize information and documents in a consistent way. Taxonomies are crucial for the management of organizations. According with Pincher [8] all types of management systems in an organization are nearly useless if they do not use taxonomies. Taxonomies are necessary to organize storage and management of resources, and to support better searching of resources.

There are other forms of representation of topics about some domain or discipline, that is the case of ontologies and knowledge graphs. In 1993 Tom Gruber [9] presented the definition of Ontology as an “explicit specification of a conceptualization”. That is, an ontology is a formal representation model based on logic that allows defining concepts and semantic relationships between them. On the other hand, knowledge graphs are also a representation model based on triplets formed by subject, predicate and object. Knowledge graphs have gained enormous popularity since they are very similar to ontologies, but their use and exploitation is much lighter, and they are easily accessible using references or resource identifiers (IRIs).

The following are some examples of knowledge representations:

- a) The ACM Computing Classification System³, is a standard classification system for the Computing Science knowledge field. It is maintained by the ACM Organization.
- b) The Computer Science Ontology (CSO)⁴, is a large-scale, automatically generated ontology of research areas in the field of Computer Science, which includes about 15,000 topics and 70,000 semantic relationships.

³ <https://www.acm.org/publications/class-2012>

⁴ <http://skm.kmi.open.ac.uk/cso/>

- c) In research areas related with Physics and Astronomy, the most popular taxonomy used is the Physics and Astronomy Classification Scheme (PACS). PACS was developed in 1970 by the American Institute of Physics (AIP) for classifying scientific literature using a hierarchical set of codes.
- d) The Mathematics Subject Classification (MSC)⁵ is the main taxonomy used in the field of Mathematics. This taxonomy is maintained by Mathematical Reviews (MRDB) y Zentralblatt MATH (ZMATH).
- e) The Medical Subject Heading (MeSH)⁶ is a controlled vocabulary produced by the National Library of Medicine, it is used for indexing, cataloging, and searching for biomedical and health-related concepts and documents.

To facilitate the task of searching and finding the authors that address a particular topic, or to retrieve a set of publications that are closely related to a particular research topic, it is necessary to correlate the publications with a taxonomy of knowledge, related with the knowledge area of interest. In this study case, the set of publications will be correlated and organized with respect to the Computer Science knowledge area, therefore the CSO Ontology was selected.

4 Similarity Measures for Texts

In recent years there has been a growing interest in the research and development of methods for short-text semantic similarity. According with Prakoso et al. [10] short text similarity (STS) aims at determining the degree of similarity between pairs of texts. Various approaches have been reported in literature: lexical or syntactic, semantic, pragmatic, probabilistic methods, vector-based, among others. Of particular interest are the knowledge-based measures that utilize the WordNet lexical database.

The following measures of relatedness were used to calculate semantic similarity between titles and topics from CSO ontology. These semantic relatedness measures utilize WordNet database and exploit additional non-hierarchical relations.

- a) Wu and Palmer [2] introduced a relatedness measure that finds the path length to the root node from the least common subsumer (LCS) of two concepts, which is the most specific common concept they share as an ancestor. This value is scaled by the sum of the path lengths from the individual concepts to the root.
- b) Jiang and Conrath [3] presented an Information Content (IC) based-distance measure that uses the conditional probability of encountering an instance of a subclass synset given an instance of a superclass synset. Thus, the information content of the two nodes, as well as that of their most specific subsume are considered.
- c) Leacock and Chodorow [4] presented a measure that finds the shortest path length between two concepts, and scales that value by the maximum path length in the is-A hierarchy in which they occur. It considers that the conceptual distance between two nodes is proportional to the number of edges separating the two nodes in the hierarchy.
- d) PATH [4] semantic relatedness is a node-counting scheme (path). The relatedness score is inversely proportional to the number of nodes along the shortest path between the synsets. The shortest possible path occurs when the two synsets are the same, in which case the length is 1. Thus, the maximum relatedness value is 1.
- e) Lin [5] presents a measure that calculates semantic relatedness between two concepts. Lin stated that "the similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are." This measure uses the amount of information needed to state the commonality between the two concepts and the information needed to describe these terms.
- f) Resnik [6] presents a semantic relatedness approach that uses the information of concepts, computed from their frequency of occurrence in a large corpus. Considers that the similarity between a pair of concepts may be judged by "the extent to which they share information", Resnik calculates the semantic relatedness between two lexicalized concepts.
- g) Additionally, we have included the calculation of the Cosine distance. Cosine similarity is a measure between two nonzero vectors of an inner product space, based on the cosine of the angle between them. If two text embedding vectors are similar, the cosine similarity between them produces a value close to 1.

5 Experimentation

For experimentation two data sources were used: a collection of 1,500 publications extracted from Semantic Scholar API, a collection of 14,290 research topics from the CSO ontology, and a set of 7 similarity measures. Considering the size of these collections, the number of similarity calculations is the product of 1,500 titles, 14,290 topics, and 7 similarity measures, that is a total of 150,045,000 similarity calculations. The first objective is to reduce the number of similarity calculations and the time required for their execution. Figure 3 shows the process of generating sample files with similarity calculations. The process

⁵ <https://mathscinet.ams.org/msc/msc2010.html>

⁶ <https://meshb.nlm.nih.gov/search>

starts by randomly selecting a sample of 100 topics from the CSO ontology, and 100 publication titles from the publication dataset. Then the seven similarity measures are calculated between all pairs of topics and titles. To filter representative results, the mean of all measures is used to select those similarities that are higher than a threshold.

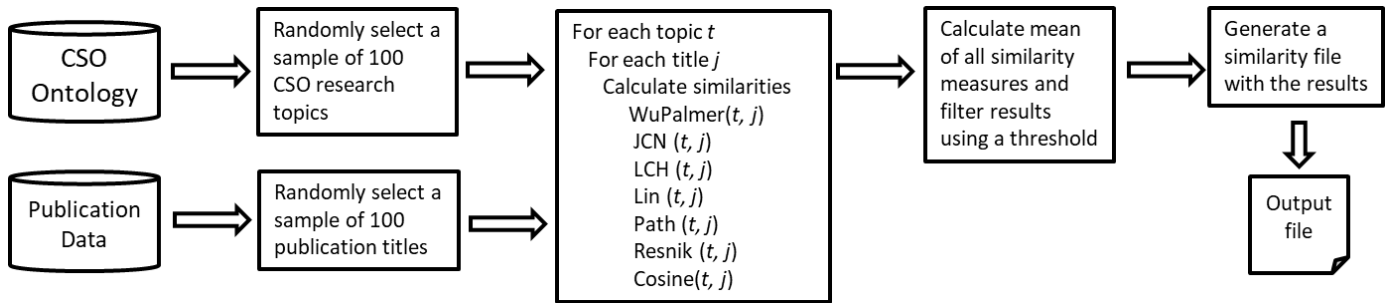


Fig. 3. Process to randomly generate similarity sample files.

To determine which is the similarity measure that generates better results, two evaluation criteria were established: a statistical analysis of measures and the performance of measures. These evaluations are described in the following subsections.

5.1 Statistical Evaluation of Similarity Measures

The objective of this analysis is to determine the stability of the similarity measures under a variance criterion. Table 1 shows the results of statistical analysis of measures for the 10 sample files generated.

StatEval It is a measure that allows calculating the error based on the statistical variance of the data. The purpose of this calculation is to select the measure of similarity that returns the least error.

As can be seen in figure 4, the similarity measurement that returns the highest results is Cosine similarity. However, it is important to evaluate the relative error using a statistical calculation to reliably determine which of the measurements presents more stable results.

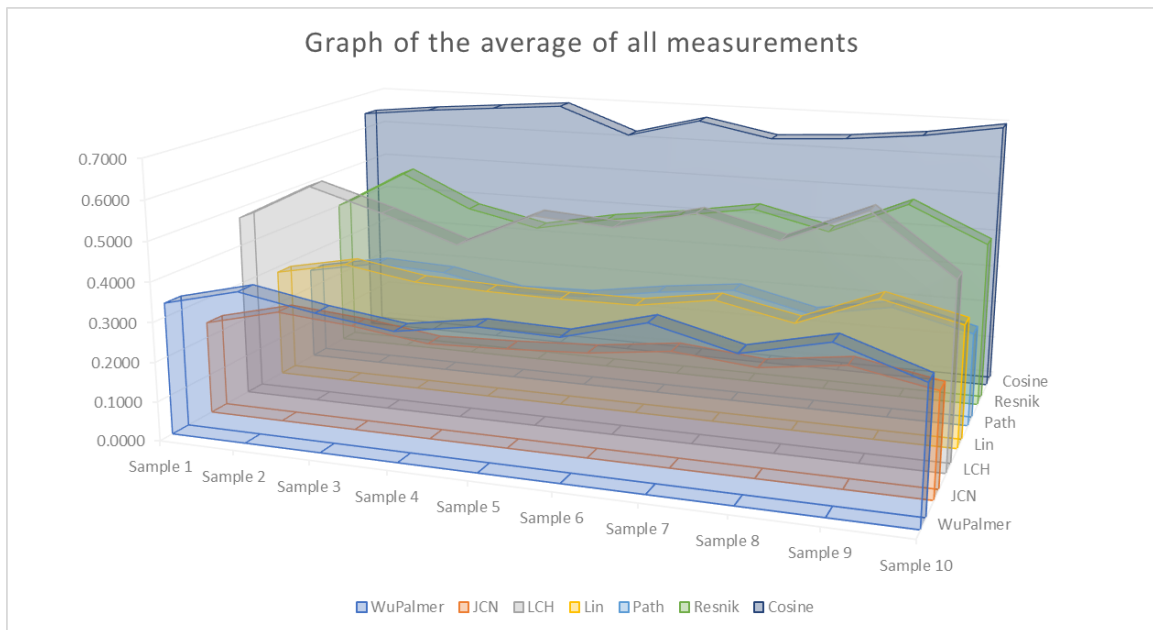


Fig. 4. Graph of the average of all of similarity measures.

Table 1. Statistical analysis of similarity measures.

Similarities results	WuPalmer	JCN	LCH	Lin	Path	Resnik	Cosine
Sample 1	0.3348	0.2381	0.4735	0.2859	0.2490	0.4006	0.6419
Sample 2	0.3812	0.2845	0.5706	0.3240	0.2851	0.5081	0.6623
Sample 3	0.3472	0.2684	0.5146	0.2935	0.2770	0.4198	0.6784
Sample 4	0.3223	0.2427	0.4487	0.2856	0.2391	0.3791	0.6943
Sample 5	0.3528	0.2500	0.5373	0.2831	0.2444	0.4181	0.6233
Sample 6	0.3475	0.2604	0.5235	0.2850	0.2747	0.4438	0.6741
Sample 7	0.4002	0.2849	0.5726	0.3167	0.2977	0.4765	0.6359
Sample 8	0.3507	0.2668	0.5223	0.2762	0.2526	0.4300	0.6477
Sample 9	0.3964	0.2935	0.6077	0.3566	0.2892	0.5170	0.6677
Sample 10	0.3290	0.2561	0.4641	0.3101	0.2398	0.4272	0.6992
Mean	0.3562	0.2646	0.5235	0.3017	0.2649	0.4420	0.6625
Variance	0.0007	0.0003	0.0026	0.0006	0.0005	0.0020	0.0006
Standard Deviation	0.0273	0.0187	0.0512	0.0251	0.0222	0.0451	0.0251
Interval	0.0158	0.0108	0.0297	0.0145	0.0129	0.0261	0.0145
Relative error	4.4491	4.0973	5.6707	4.8177	4.8573	5.9114	2.1936

According with Table 1 results, the most stable similarity measure is the Cosine similarity, followed by the JCN measure because they report the smallest relative errors.

5.2 Performance of Measures

For informed decision making it is relevant to use more than one evaluation approach or criteria, so that the same set of measurements can be evaluated using the same set of input data. Therefore, one of the evaluations that we consider important is that of precision and recall, which we will call performance evaluation, that is, we will verify the certainty of the results it produces.

The evaluation of the performance of measures will be made using the Precision, Recall and F1 measures. Table 2 shows the precision of each semantic similarity measure, applied to each sample file. Accordingly, Lin and LCH measures show better precision results than the others. Table 3 shows the recall results of each measure.

Table 2. Precision of similarities for each sample.

Similarities samples	WuPalmer	JCN	LCH	Lin	Path	Resnik	Cosine
Sample 1	0.0000	0.0000	0.6000	0.0000	0.0000	1.0000	0.2857
Sample 2	1.0000	1.0000	0.5385	1.0000	0.0000	0.7500	0.5263
Sample 3	0.0000	0.0000	0.4167	1.0000	0.0000	0.1667	0.3889
Sample 4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5625
Sample 5	0.0000	0.0000	0.3750	0.0000	0.0000	0.0000	0.3846
Sample 6	1.0000	0.0000	0.2500	0.0000	0.0000	0.0000	0.2778
Sample 7	0.0000	0.0000	0.1538	0.0000	0.0000	0.1429	0.1667
Sample 8	0.5000	0.0000	0.5000	1.0000	0.0000	0.6667	0.4000
Sample 9	0.6667	1.0000	0.6154	1.0000	1.0000	0.3000	0.4211
Sample 10	1.0000	0.0000	0.8333	1.0000	0.0000	0.7500	0.5500
Mean	0.4167	0.2000	0.4283	0.5000	0.1000	0.3776	0.3964

Table 3. Recall of similarities for each sample.

Similarities samples	WuPalmer	JCN	LCH	Lin	Path	Resnik	Cosine
Sample 1	0.0000	0.0000	0.6000	0.0000	0.0000	0.2000	0.8000
Sample 2	0.0909	0.0909	0.6364	0.0909	0.0000	0.5455	0.9091
Sample 3	0.0000	0.0000	0.5556	0.1111	0.0000	0.1111	0.7778
Sample 4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000
Sample 5	0.0000	0.0000	0.4286	0.0000	0.0000	0.0000	0.7143
Sample 6	0.4000	0.0000	0.4000	0.0000	0.0000	0.0000	1.0000
Sample 7	0.0000	0.0000	1.0000	0.0000	0.0000	0.3333	1.0000
Sample 8	0.1250	0.0000	0.3750	0.1250	0.0000	0.2500	1.0000
Sample 9	0.0000	0.1250	1.0000	0.1250	0.1250	0.3750	0.7143
Sample 10	0.0909	0.0000	0.4545	0.0909	0.0000	0.2727	1.0000
Mean	0.0707	0.0216	0.5450	0.0543	0.0125	0.2088	0.8915

As the results of similarity calculations showed that there are many dissimilarities between CSO concepts and titles, then a measure that balances precision and recall is necessary. The *FI* score is the harmonic mean of the precision and recall scores. The *FI* measure penalizes classifiers with unbalanced precision and recall scores. *FI* score is calculated as follows:

$$FI\ Score = 2 * (Precision * Recall) / (Precision + Recall)$$

Table 4 and Figure 5 show that the measures with the best *FI* scores are LCH and Cosine similarities.

Table 4. FI score of similarity measures.

Similarities samples	WuPalmer	JCN	LCH	Lin	Path	Resnik	Cosine
Sample 1	0.0000	0.0000	0.6000	0.0000	0.0000	0.3333	0.4211
Sample 2	0.1667	0.1667	0.5833	0.1667	0.0000	0.6316	0.6667
Sample 3	0.0000	0.0000	0.4762	0.2000	0.0000	0.1333	0.5185
Sample 4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7200
Sample 5	0.0000	0.0000	0.4000	0.0000	0.0000	0.0000	0.5000
Sample 6	0.5714	0.0000	0.3077	0.0000	0.0000	0.0000	0.4348
Sample 7	0.0000	0.0000	0.2667	0.0000	0.0000	0.2000	0.2857
Sample 8	0.2000	0.0000	0.4286	0.2222	0.0000	0.3636	0.5714
Sample 9	0.0000	0.2222	0.7619	0.2222	0.2222	0.3333	0.5298
Sample 10	0.1667	0.0000	0.5882	0.1667	0.0000	0.4000	0.7097
Mean	0.1105	0.0389	0.4413	0.0978	0.0222	0.2395	0.5358

In Figure 5 we can see that the Cosine similarity measure is the one that obtains the best results in relation to the *FI* score.

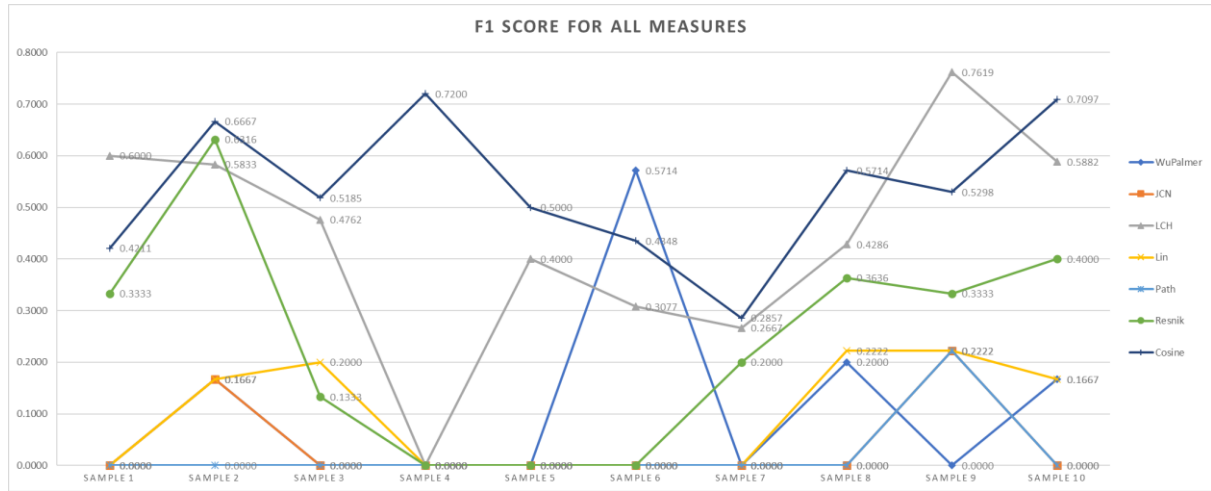


Fig. 5. *F1* scores of all similarity measures.

5.3 Overall Evaluation of Measures

The overall evaluation is calculated as a weighed mean of the measure *F1* and the measure of exploratory analysis. However, the two measurements have a reverse direction. The measure *F1*, the higher the value is returned, the better, while the result of the exploratory analysis, the lower the value, the better. Therefore, the calculation of the average includes the inverse value of the exploratory analysis.

$$OverEval = (F1 * w_1) + \left(\frac{1}{StatEval} * w_2\right)$$

Where:

F1 is the harmonic mean of precision and recall measures.

StatEval is the relative error of the measures.

w_1, w_2 represent the weights.

Table 5. Precision and recall of similarity measures.

	WuPalmer	JCN	LCH	Lin	Path	Resnik	Cosine
Exploratory statistical	4.4491	4.0973	5.6707	4.8177	4.8573	5.9114	2.1936
F1 Measure	0.1105	0.0389	0.4413	0.0978	0.0222	0.2395	0.5358
Overall evaluation	0.1562	0.1210	0.3353	0.1417	0.0957	0.2114	0.5038

Table 5 shows that the best measure is the *Cosine* similarity, considering the exploratory statistical analysis and the *F1* measure. It is pertinent to clarify that the similarity of the cosine is a different measurement from the other 6 based on WordNet, the first 6 measurements are considered of the Information Content type and are supported by the use of the WordNet dictionary and its semantic relationships to determine the similarity, while cosine similarity is not considered semantic.

This result is interesting, since it could be assumed that information-based measures should return better results than a similarity that has a lexical basis. Additionally, any of the measurements that use WordNet take more resources to calculate, while the Cosine is much lighter.

6 Conclusions

There are multiple machine learning tasks that eventually require the execution of similarity measurements. Also considering that there are various measurement approaches and multiple similarity measurements that can be applied to short texts.

This article reports a method to decide which measure of similarity is better with respect to the problem of calculating distances between short texts. This method is especially effective because it reduces the size of the calculations for large amounts of text and allows by means of a sampling to determine the measure of similarity that will offer the best results on the complete set.

The use of an evaluation measure based only on precision and recall does not allow to determine if a similarity measure will result in statistical errors regarding the data being used. Instead, in this paper a combined evaluation method is reported, providing a more adequate reference to decide about the most reliable similarity measure for the given collection of data.

References

1. Jinha, A. E. (2010). Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3), 258-263.
2. Wu, Zhibiao, and Martha Palmer. "Verbs semantics and lexical selection." In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133-138. Association for Computational Linguistics, 1994.
3. Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.
4. C. Leacock, M. Chodorow. "Combining local context and WordNet similarity for word sense identification". *WordNet: An electronic lexical database*. Vol 49. No. 2. 1998. 265-283.
5. Dekang Lin, 'An information-theoretic definition of similarity', in *Proceedings of the 15th International Conf. on Machine Learning*, pp. 296– 304. Morgan Kaufmann, San Francisco, CA, (1998).
6. Philip Resnik. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, Montreal.
7. Lambe, P. (2014). *Organising knowledge: taxonomies, knowledge and organisational effectiveness*. Elsevier.
8. Pincher, M. (2010). *A guide to developing taxonomies for effective data management*. Computer Weekly, 8.
9. Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2), 199-220.
10. Prakoso, D. W., Abdi, A., & Amrit, C. (2021). Short text similarity measurement methods: a review. *Soft Computing*, 25(6), 4699-4723.