DOI: https://doi.org/10.31341/jios.44.2.2
Open Access
UDC 004.4'414
Original Scientific Paper

A Comparison of Approaches for Measuring the Semantic Similarity of Short Texts Based on Word Embeddings

Karlo Babić *karlo.babic@uniri.hr*
Department of Informatics
Center for Artificial Intelligence and Cybersecurity
University of Rijeka, Rijeka, Croatia

Francesco Guerra *francesco.guerra@unimore.it*
Department of Engineering Enzo Ferrari
University of Modena and Reggio Emilia, Modena, Italy

Sanda Martinčić-Ipšić *smarti@uniri.hr*
Department of Informatics
Center for Artificial Intelligence and Cybersecurity
University of Rijeka, Rijeka, Croatia

Ana Meštrović *amestrovic@uniri.hr*
Department of Informatics
Center for Artificial Intelligence and Cybersecurity
University of Rijeka, Rijeka, Croatia

Abstract

Measuring the semantic similarity of texts has a vital role in various tasks from the field of natural language processing. In this paper, we describe a set of experiments we carried out to evaluate and compare the performance of different approaches for measuring the semantic similarity of short texts. We perform a comparison of four models based on word embeddings: two variants of Word2Vec (one based on Word2Vec trained on a specific dataset and the second extending it with embeddings of word senses), FastText, and TF-IDF. Since these models provide word vectors, we experiment with various methods that calculate the semantic similarity of short texts based on word vectors. More precisely, for each of these models, we test five methods for aggregating word embeddings into text embedding. We introduced three methods by making variations of two commonly used similarity measures. One method is an extension of the cosine similarity based on centroids, and the other two methods are variations of the Okapi BM25 function. We evaluate all approaches on the two publicly available datasets: SICK and Lee in terms of the Pearson and Spearman correlation. The results indicate that extended methods perform better from the original in most of the cases.

Keywords: semantic similarity, short texts similarity, word embedding, Word2Vec, FastText, TF-IDF

1. Introduction

Measuring the semantic similarity of texts has a vital role in various tasks from the field of natural language processing (NLP) such as document classification [1], information retrieval [2], word sense disambiguation [3], plagiarism detection [4], etc. The specific task of measuring the semantic similarity of short texts is of importance in the domain of social media for opinion mining, recommendation [5], event detection [6], news recommendation [7]. Representing short texts may differ from representing long texts due to the sparsity and noisiness [7], [8]. Hence, it is important to develop approaches focused only on short texts such as tweets, comments, or microblogs [9]. Therefore, approaches that are tailored to short texts may not work well with long texts and vice versa.

A large number of approaches is addressing the problem of modeling short texts. Typically, they model short text as an aggregate of words and apply specific metrics to compute the similarity of aggregations [10], [11], [12], [13]. Most of the existing techniques represent text as a weighted set of words (e.g. bag of words), where the order of words in a text and the meaning of words is disregarded [14], [15], [16], [17].

Recently, neural networks have been adopted for the generation of word embeddings. Word embeddings represent a model based on distributional semantics, which describes the context of a word. There is a variety of representation models based on word embeddings. Among them, the most popular are Word2Vec [18], [19], FastText [20], and GloVe [21]. However, word embedding models have certain limitations.

Firstly, word embeddings are derived for one word. In the case of short texts, it is necessary to scale up from word embeddings to text embedding. A large number of techniques that leverage this issue are proposed, and still, there is no consensus in the research community on how to proceed. One possibility is to take the sum or the average (centroid) of the individual word embeddings for all the words in the text. This approach has been widely adopted in many studies, for example, [10], [13], [1], and in general, they perform well. However, by aggregating a set of word embeddings into only one embedding as (averaged or weighted) sum or centroid, we are losing valuable semantic information. This happens is because of reducing the information contained in the set of vectors into one vector.

There are other possible approaches different from this centroid-based approach like for example, in [11] the authors use the Okapi BM25 function and in [12] the authors define the Word Mover's Distance similarity measure to calculate the semantic similarity of short texts based on word embeddings.

Secondly, word embeddings can typically capture only one meaning per word, and this may cause problems with words that have more than one meaning (polysemy). For this reason, other techniques described in [22], [23], [24], [25] have been proposed to extend the approaches above with embeddings of words that can capture more than one meaning associated with the word. In [22] the authors introduced the NASARI dataset that integrates pre-trained word embeddings based on the Word2Vec model with the word sense embeddings reached from the BabelNet. BabelNet is a multilingual dictionary, which contains synsets that can be used to

resolve the vocabulary problems such as synonymy and polysemy [23]. It merges WordNet with other lexical and encyclopedic resources such as Wikipedia and Wiktionary [24], [25], [26].

The motivation for this work stems from the still unsolved problem of representing polysemy for a better assessment of the semantic similarity in short texts. We opt to obtain additional insights into the performance of various word representation models i.e., Word2Vec, FastText, and TF-IDF. In this work, we extend our preliminary research described in [27], in which we preliminary compare the Word2Vec representation with its extension obtained from the NASARI dataset (which includes word sense descriptions in the task of measuring semantic similarity). Here we describe an extended set of experiments in which we evaluate the performance of four word representation models: Word2Vec, Word2Vec+NASARI, FastText, and traditional TF-IDF as a baseline. All the models are combined with five methods for measuring similarity of short texts. More precisely, we experimented with the centroid-based and BM25 methods for calculating text similarity from the set of word embeddings.

Additionally, we proposed three variations of these two methods. The first one is a modified version of the centroid method that uses the inverse document frequency (IDF), and the other two methods modify the BM25 function by leaving out some constants and introducing IDF. To the best of our knowledge, this is the first attempt to experiment with the proposed measures in the task of measuring semantic similarity.

In short, in this paper we address two main open issues relating to the task of measuring semantic similarity: (i) how to aggregate word representations for modeling short texts, and (ii) how to capture more than one meaning per word (polysemy). To resolve (i), we test five methods that aggregate word embeddings and provide the semantic similarity score. To resolve (ii), we apply a technique based on the NASARI dataset by incorporating word senses into word embeddings.

The rest of the paper is organized as follows. In the second section, we present related work. After that, in the third section, we describe word representation models, and we give an overview of various word embeddings based methods for calculating the semantic similarity of short texts. In the fourth section, we provide evaluation results. Finally, in the last section, we provide a conclusion and the possible directions for future work.

2. Related Work

So far, there have been numerous approaches developed for the task of measuring the semantic similarity of words and texts. Generally, they are classified into two groups: knowledge-based and corpus-based [28].

Knowledge-based measures of semantic similarity rely on external sources of knowledge (e.g., ontologies processed as semantic graphs or semantic networks, and/or lexical resources such as WordNet [29], [24], Wikipedia [30], [31], etc.). Commonly, these measures use the formal expression of knowledge, explicitly defining how to compare entities in terms of semantic similarity.

Corpus-based measures enable the comparison of language units of different sizes such as words or texts. They determine the semantic similarity between words or texts using information derived from the statistics of large corpora. These include traditional n-gram measures [32], [16], the bag of words (BoW) model using the TF-IDF weighting scheme [33], [13], [14], [15], [16] or more complex approaches such as Latent Semantic Analysis (LSA) proposed by Landauer [34].

Recent trends in NLP prefer corpus-based over knowledge-based representation models with dense low dimensional vectors in a continuous space such as Word2Vec [18], GloVe [21], FastText [20] and recently ELMo [35].

Continuous-space models for word (document) representation, referred to as Word2Vec (Doc2Vec), embed the words (documents) in a vector space, where the closeness of vectors corresponds to the semantic similarity of words (documents) [18], [19], [36]. Hence, the results of all the proposed models are embeddings (low dimensional vectors in a continuous space) with the property that semantically similar words tend to have vectors that are close in the semantic space [18], [37].

Identifying the degree of semantic similarity of short texts based on word embeddings is a challenging task that has been studied extensively during the past years. Still, only a small number of solutions for the sentence or document embedding has been proposed, as for example [36], [38]. However, in this study, we are focused on approaches that determine the semantic similarity of short texts based only on word embeddings.

Mihalcea et al. proposed assessing the semantic similarity of texts by exploiting the information derived from the similarity of the component words [28]. To this end, they assess two corpus-based and six knowledge-based measures of word semantic similarity. According to the results, the proposed method with a combination of six knowledge-based measures outperforms the vector-based similarity approach in the task of paraphrase detection. However, this approach is rather traditional, and it is not based on word embeddings.

Kusner et al. introduced a new measure, called the Word Mover's Distance (WMD), which quantifies the dissimilarity between two documents [12]. Documents are represented with word embeddings, and the distance is calculated as the minimum amount of distance that the embedded words of the source document need to "travel" to reach the embedding in the target document. The measure is evaluated in the task of text classification, and the results indicate that WMD tends to have lower classification error rates in comparison to other state-of-the-art methods. Furthermore, in [10] the authors used the WMD method for information retrieval in the biomedical domain. These are examples of the indirect evaluation of the WMD method since the measure is not directly applied to measure the semantic similarity. In this work, we propose similar methods and perform a direct evaluation, which enables a better comparison of the measures.

In [7], the authors defined a novel method for the vector representations of short texts. The method uses word embeddings and learns how to weigh each embedding based on its IDF value. The proposed method works with texts of a predefined length but can be extended to any length. The authors showed that their method outperforms other baseline methods that aggregate word embeddings for modeling short texts.

Kenter and De Rijke proposed measuring the semantic similarity of short texts by combining word embeddings with external knowledge sources [11]. They used various text features to train a supervised model. Specifically, they employed a modification of the Okapi BM25 function for document ranking in information retrieval and adjust it to measure the semantic similarity of short texts. They showed that their method outperforms the baseline method in the task of measuring the semantic similarity of short texts.

In our work, we adopt similar principles for measuring the semantic similarity of short texts. However, we propose modifications of the centroid based and Okapi BM25 methods for measuring semantic similarity based on the aggregation of word embeddings into short text embeddings. Next, we perform the evaluations of these methods in combination with four different representation models.

3. Methodology

In this section, we describe the methodology adopted in the paper based on the embeddings and the methods used for the pairwise measurement of the semantic similarity of short texts.

3.1. Word and Word Senses Embeddings

Initially, we describe four word embedding models used in experiments: three neural-network-based models and the TF-IDF model as the baseline.

First, we test the UMBCw2v, set of word embeddings trained on the publically available corpus called UMBC [39]. UMBC trained embeddings are freely available and have already been used in several experiments. The application of embeddings is straightforward: each word is replaced with its corresponding embedding from the Word2Vec set through a lookup table. Word2Vec has two models: continuous bag of words (CBOW) and skip-gram [19]. Both learn word representation through unsupervised learning. The CBOW model scans over the text with a context window around the target word, and it learns to predict the target word from the context words. The skip-gram model learns to predict the context words from the target word. The Word2Vec neural network has only one hidden layer, and word representations are extracted from that layer as dense low-dimensional vector representations of words.

The second model is based on the NASARI set of embeddings [40]. NASARI embeddings incorporate external knowledge by introducing word sense embeddings from the BabelNet synsets [22]. In our experiments, similarly as in [40], we use a NASARI dataset combined with UMBCw2v embeddings, and we call this representation model NASARI+Word2Vec. The application of these embeddings requires the use of the Babelfy system to retrieve the ID of the proper sense associated with a word. The ID is then used to find the embedding for that word or phrase in the NASARI+Word2Vec set of embeddings. For an out of vocabulary word (i.e., not included in Babelfy and does not have any ID), the embedding is extracted from the UMBCw2v set. This way, we enable the disambiguation of different word senses. Note that both sets of embeddings are trained in the same vector space. The resulting

dense vectors have a vector dimensionality of 300. Thus, all the embeddings are numerically, and even more importantly, semantically comparable.

As the third model, we use FastText [15]. The FastText model, like Word2Vec, uses a continuous representation of words, trained on large unlabeled corpora, however, with an important difference - FastText learns word embeddings on the subword level. Each word is represented as a bag of n-gram characters. A vector representation is associated with each n-gram, and each word is represented as a sum of the n-gram vector representations. When training on the subword level, the tokens (character n-grams) occur more frequently throughout the training process, effectively estimating the parameters in the neural model. FastText is very fast to train, and in many cases, it outperforms other models, especially on morphologically rich languages [15]. We use pre-trained word vectors for the English language [41].

The fourth model is the traditional TF-IDF (term frequency-inverse document frequency) and serves as the baseline approach. The TF-IDF proposed in [16], [17] is one of the oldest and yet very efficient approaches for measuring term (word) significance in a document. TF-IDF values are assigned to words, depending on the word frequency in a document, and the inverse occurrence in the corpus.

3.2. Methods for Measuring the Semantic Similarity of Short Texts

In this section, we introduce five methods for calculating the semantic similarity scores between two short texts based on their word embeddings.

The first method is the most basic one and is based on centroids. For a given text represented with the set of word embeddings V, the centroid of V is calculated according to the equation:

$$cent(V) = \frac{\sum_{v \in V} v}{|V|}. \quad (1)$$

The centroid is typically adopted in the literature for synthesizing the meaning of a text. We also experiment with a modified version of the centroid method that uses the inverse document frequency (idf) multiplied with each word embedding (word vector). This variant builds weighted centroids, where the uncommon terms in the collection assume greater importance:

$$cent_{idf}(V) = \frac{\sum_{v \in V} idf(v) \cdot v}{|V|}. \quad (2)$$

By using the centroids, the similarity measure of two documents sim_{cos} is computed as the cosine similarity between centroids of two texts t_1 and t_2, represented with sets of embeddings V_1 and V_2 respectively:

$$sim_{cos}(t_1, t_2) = cos(cent(V_1), cent(V_2)) \quad (3)$$

Analogously, sim_{cos2} is calculated as the cosine similarity between the weighted centroids of two texts (short texts or sentences).

Additionally, we experiment with three other methods based on the Okapi BM25 function. The first one is a reconstruction of the method in [11] and the other two we propose and examine as the possible simplifications of the original method.

The modified version of the Okapi BM25 function that can be applied for measuring the semantic similarity of two texts (short texts or sentences) introduced in [11] is:

$$sts(t_l, t_s) = \sum_{w \in t_l} idf(w) \cdot \frac{sem(w,t_s)\cdot(k_1+1)}{sem(w,t_s)+k_1\cdot(1-b+b\cdot\frac{|t_s|}{avgtl})}, \quad (4)$$

where t_l is the longer text, and t_s is the shorter text. Variables k_1 and b are parameters which can be optimized, the variable $avgtl$ is the average text length.

Function sem for a given word w and text t is defined as:

$$sem(w, t) = max_{w' \in t} cos(w, w'). \quad (5)$$

Next, we introduce two modifications of the equation (5) by leaving out constants k_1 and b. This results in two simplified versions of equation 4.

Equation (6) calculates the average value returned by the function sem (5) multiplied by the idf.

$$sts_s(t_l, t_s) = \frac{\sum_{w \in t_l} idf(w)\cdot sem(w,t_s)}{|t_l|}. \quad (6)$$

Equation (7) is another modification of (5). Here, instead of just calculating the idf of the word from the longer text, idf is calculated for words from both texts (t_s represents the word from the shorter text). One more difference is that the resulting value is passed through the log function, so the most extreme values are reduced.

$$sts_{s2}(t_l, t_s) = log(\frac{\sum_{w \in t_l}(idf(w)+idf(w_s))\cdot sem(w,t_s)}{|t_l|}). \quad (7)$$

3.3. Evaluation

The evaluation of the semantic similarity is standardly performed by the Pearson and/or Spearman correlation coefficient [42]. More precisely, in this task, the Pearson correlation quantifies how correlated the semantic similarities are of pairwise short texts annotated by humans and semantic similarities provided by the system. Thus, the Pearson correlation coefficient measures the linear correlation between two variables. The Spearman correlation measures the rank correlation in terms of the dependence between the human rankings and system rankings. Thus, the Spearman correlation is a nonparametric measure of the rank correlation and it describes how well the relationship between the two variables can be described using a monotonic function.

In order to perform an automatic evaluation, it is necessary to annotate datasets of pairwise texts by a human score (value) that denotes semantic similarity. The

agreement between multiple human annotations is usually calculated as the inter-annotator agreement [43].

4. Results and Discussion

4.1. Datasets

To evaluate the performance of short text similarity measures, we used two datasets of short texts. The first dataset (d1), called the SICK dataset in its original version, is defined within the tasks of the SemEval-2014 International Workshop for the two tasks: determining the degree of relatedness between two sentences and detecting the entailment relation between sentences [44]. The dataset consists of 5,000 English sentence pairs. Each sentence pair is annotated with a score that represents the degree of sentence similarity according to a scale ranging from 1 to 5 (where 1 means that there is no semantic similarity and 5 refers to semantically equivalent sentences). Human annotators annotate the scores.

The second dataset (d2), referred to as the Lee dataset is defined in [45] for the task of evaluating measures for text-to-text similarity. The dataset is composed of 50 short English documents (sentences) presenting news from the Australian Broadcasting Corporations news mail service. Each document pair (2,500 pairs in total) is annotated with the score of relatedness using discrete values from 1 to 5, proposed as an average score based on the annotation of ten annotators with an inter-annotator agreement of 0.61. Note that in the cases where similarity scores are represented with values out of interval [0,1], we normalize the score values.

4.2. Evaluation Results

In this section, we present the evaluation and comparison of the representation models and methods for measuring short text semantic similarity. We compute the pairwise similarity of all the short texts in both datasets and compare the results with human annotations in terms of the Pearson and Spearman correlations. Tables 1 and 2 show the results obtained on the SICK dataset and Tables 3 and 4 show the results obtained for the Lee datasets, respectively.

The rows represent the similarity measures experimented, namely sim_{cos}, sim_{cos2}, sts, sts_s and sts_{s2}. The columns represent the correlation measures computed with the Word2Vec embedding, NASARI+Word2Vec, FastText, and TF-IDF in terms of the Pearson (Table 1, Table 3) and Spearman (Table 2, Table 4) correlations. Additionally, we calculated an average of the correlations across all similarity measures for each model denoted as AVG1 and an average of the obtained correlations across all models for each similarity measure denoted as AVG2.

The set of experiments performed on the SICK datasets show that according to the Pearson correlation the absolutely best performance is achieved using the FastText model in combination with sim_{cos} measure (0,71). On average the best performance have FastText and Word2Vec models (0.60). All three DL models are close in the

performance, while the TF-IDF model has much lower results for all five measures and it is on average two times worse than the DL models. On average, sim_{cos2} provides the best results and it is slightly better than its original, commonly used version sim_{cos}. All measures based on the Okapi BM25 function provide lower results than the cosine-based measures. However, two variations (sts_s and sts_{s2}) of the original version sts improved the results. Almost the same conclusions hold in the case of the evaluation based on the Spearman correlation.

	Word2Vec	NASARI+ Word2Vec	FastText	TF-IDF	**AVG2**
sim_{cos}	0.64	0.59	**0.71**	0.18	**0.53**
sim_{cos2}	**0.66**	0.60	0.70	**0.46**	**0.61**
sts	0.50	0.47	0.43	0.27	**0.42**
sts_s	0.57	0.55	0.54	0.29	**0.49**
sts_{s2}	0.64	**0.61**	0.62	0.31	**0.55**
AVG1	**0.60**	**0.56**	**0.60**	**0.30**	

Table 1: The Pearson (r) correlations of five similarity measures for Word2Vec, NASARI+Word2Vec, FastText, and the TF-IDF approaches for the SICK dataset and the average values of correlations across approaches (AVG1) and across measures (AVG1).

	Word2Vec	NASARI+ Word2Vec	FastText	TF-IDF	AVG2
sim_{cos}	**0.59**	**0.56**	**0.61**	0.31	**0.51**
sim_{cos2}	0.58	0.55	0.59	**0.44**	**0.54**
sts	0.47	0.44	0.40	0.27	**0.39**
sts_s	0.53	0.52	0.51	0.27	**0.46**
sts_{s2}	0.54	0.52	0.52	0.30	**0.47**
AVG1	**0.54**	**0.52**	**0.52**	**0.31**	

Table 2: The Spearman (ρ) correlations of five similarity measures for Word2Vec (a1), NASARI+Word2Vec (a2), FastText (a3), and the TF-IDF (a4) approaches for the SICK dataset and the average values of correlations across approaches (AVG1) and across measures (AVG1).

We repeated the same set of experiments on the Lee dataset. The best performance is achieved using the Word2Vec model, which slightly outperforms the other two DL models (NASARI+Word2Vec and FastText). Again, the worst results are achieved in the case of the TF-IDF model (the correlation is almost two times lower than for DL models). Overall, sim_{cos2} shows the highest values of correlations for all models. Moreover, in the case of the Lee dataset, the sim_{cos2} measure has the best performance not only on average but in all cases: with all models according to both evaluation measures. In the case of the measures based on the Okapi BM25 function, the proposed measures: sts_s and sts_{s2} again perform better than the original

sts version. Exactly the same discussion holds in the case of the Spearman correlation, which confirms our findings.

	Word2Vec	NASARI+ Word2Vec	FastText	TF-IDF	AVG2
sim_{cos}	0.58	0.47	0.46	0.15	**0.42**
sim_{cos2}	**0.59**	**0.50**	**0.52**	**0.49**	**0.52**
sts	0.28	0.28	0.15	0.19	**0.23**
sts_s	0.47	0.50	0.47	0.29	**0.43**
sts_{s2}	0.42	0.44	0.43	0.27	**0.39**
AVG1	**0.47**	**0.44**	**0.41**	**0.28**	

Table 3: The Pearson (r) correlations of five similarity measures for Word2Vec (a1), NASARI+Word2Vec (a2), FastText (a3), and the TF-IDF (a4) approaches for the Lee dataset and the average values of correlations across approaches (AVG1) and across measures (AVG1).

	Word2Vec	NASARI+ Word2Vec	FastText	TF-IDF	AVG2
sim_{cos}	**0.52**	0.46	0.44	0.18	**0.40**
sim_{cos2}	**0.52**	**0.48**	**0.46**	**0.33**	**0.45**
sts	0.19	0.23	0.15	0.16	**0.18**
sts_s	0.29	0.41	0.29	0.18	**0.29**
sts_{s2}	0.29	0.38	0.30	0.18	**0.29**
AVG1	**0.36**	**0.39**	**0.33**	**0.21**	

Table 4: The Spearman (ρ) correlations of five similarity measures for Word2Vec, NASARI+Word2Vec, FastText, and the TF-IDF approaches for the Lee dataset and the average values of correlations across approaches (AVG1) and across measures (AVG2).

The overall comparison of all approaches is illustrated in Figure 1. Here we can see that there is only a slight difference between the models based on the neural networks. However, all three DL models outperform the traditional TF-IDF model. By defining a weighted centroid measure (sim_{cos2}), we managed to improve the other similarity measures in most of the cases.

In comparison to the results reported for the SICK dataset, the presented approaches perform better than few approaches described in [44]. FastText and Word2Vec, in combination with both the centroid-based methods, slightly outperforms the baseline (reported as an overlap of 0.63).

Additionally, we focus on the comparison of the first two approaches: one that uses Word2Vec and the second approach that combines Word2Vec with NASARI embeddings. We expected that introducing a NASARI dataset as an external knowledge resource would improve the performance of the Word2Vec model.

However, it is shown that an approach with classical Word2Vec embeddings slightly outperforms the NASARI variant of embeddings on both datasets regardless of the used measures. Potentially this is caused by the assumption that the Babelfy system does not achieve its full potential since it does not always return the correct word sense embedding for a given word within a context. There is certainly space for improvements in the BabelNet and Babelfy systems, hence of our method as well. Moreover, there are minor deviations compared to the results using the centroid similarity measure on the Lee dataset reported in [17] because recently, a new version of the NASARI dataset was made available. However, the overall results of the centroid-based similarity are in line with the previous study of both approaches (the NASARI approach and Word2Vec approach) [22].

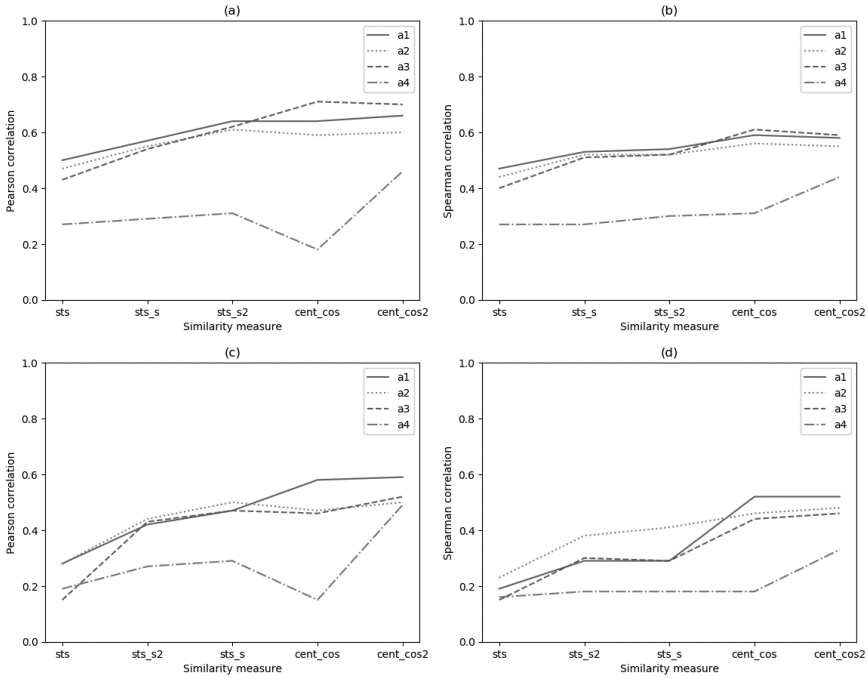

Figure 1. The comparison of the performance of 4 models (Word2Vec (a1), NASARI+Word2Vec (a2), FastText (a3), and TF-IDF (a4)) in combination with 5 similarity measures in terms of the Pearson (left) and Spearman (right) correlations of the SICK (top) and Lee (bottom) datasets

5. Discussion

In this paper, we present research focused on measuring the semantic similarity of short texts. We test and compare four representation models: (i) the Word2Vec model, (ii) its extension with embeddings of word senses NASARI provided by the Babelfy system, (iii) the FastText, and (iv) traditional TF-IDF as the baseline. We combine these representation models with classical centroid-based and BM25-based methods

and their modifications proposed in this paper for calculating the measure of similarity of two short texts.

The evaluation results with two datasets (SICK and Lee) in terms of the Pearson and Spearman correlations indicate that models based on deep learning and neural networks (Word2Vec, NASARI+ Word2Vec, FastText) outperform the traditional TF-IDF model.

Concerning the different methods for measuring similarity, centroid-based methods generally outperform BM25-based methods. According to the Pearson and Spearman correlations, the weighted centroid measure sim_{cos2} proposed in this paper outperforms the traditionally used centroid measure sim_{cos} in both datasets in almost all cases. Both modifications of the sts method that we propose in this paper perform better than sts in all cases.

Regarding our attempt to improve the basic Word2Vec by introducing the NASARI dataset as an external source of knowledge, it seems that the Word2Vec model performs better than its extension. Obviously, the semantics provided by NASARI do not contribute to the improvement of the performance in the results as anticipated. The reason might be that the NASARI dataset is not yet fully developed. Therefore, we expect that with the better version of NASARI and Babelfy, this extended representation model, NASARI + Word2Vec will gain in performance. Still, this remains an open question to be tested in the future.

All of these findings indicate that there is room for improvements and that it is possible to define new approaches for measuring the similarity of short texts.

For future work, we will systematically experiment with new language representation models based on neural networks. Additionally, we will try to incorporate various external knowledge resources, with a special focus on the sources that are based on the networks and graphs [46] by integrating text embeddings with graph embeddings. Moreover, we plan to explore existing external knowledge resources such as Google Knowledge Graph, Wikipedia, and ontologies in general for the purpose of resolving the vocabulary problems (synonymy and polysemy).

Acknowledgements

This work has been supported by the University of Rijeka under the project: **uniri-drustv-18-38**

References

[1] S. Martinčić-Ipšić, T. Miličić and L. Todorovski, "The Influence of Feature Representation of Text on the Performance of Document Classification," *Applied Sciences,* vol. 9, no. 4, p. 743, 2019.

[2] A. Meštrović and A. Calì, "An ontology-based approach to information retrieval," *Semanitic Keyword-based Search on Structured Data Sources,* pp. 150-156, 2016.

[3] A. Raganato, C. Bovi and R. Navigli, "Neural sequence learning models for word sense disambiguation," in *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 2017.

[4] T. Vrbanec and A. Meštrović, "Corpus-Based Paraphrase Detection Experiments and Review," *Information,* vol. 11, no. 5, pp. 1-25, 2020.

[5] N. Jonnalagedda and S. Gauch, "Personalized News Recommendation Using Twitter," *WI-IAT,* vol. 3, pp. 21-25, 2013.

[6] C. De Boom, S. Van Canneyt and B. Dhoedt, "Semantics-driven event clustering in Twitter feeds," *In Making Sense Of Microposts,* vol. 1395, pp. 2-9, 2015.

[7] C. De Boom, S. Van Canneyt, T. Demeester and B. Dhoedt, "Representation learning for very short texts using weighted word embedding aggregation," *Pattern Recognition Letters,* vol. 80, pp. 150-156, 2016.

[8] S. Martinčić-Ipšić, E. Močibob and M. Perc, "Link prediction on Twitter," *PloS one,* vol. 12, no. 7, 2017.

[9] E. Mocibob, S. Martinčić-Ipšić and A. Meštrović, "Revealing the structure of domain specific tweets via complex networks analysis," in *39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2016.

[10] G. Brokos, P. Malakasiotis and I. Androutsopoulos, "Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering," *arXiv preprint,* 2016.

[11] T. Kenter and M. De Rijke, "Short text similarity with word embeddings," in *24th ACM international on conference on information and knowledge management*, 2015.

[12] M. Kusner, Y. Sun, N. Kolkin and K. Weinberger, "From word embeddings to document distances," in *International conference on machine learning*, 2015.

[13] G. Rossiello, P. Basile and G. Semeraro, "Centroid-based text summarization through compositionality of word embeddings," in *Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, 2017.

[14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management,* vol. 24, no. 5, pp. 513-523, 1988.

[15] G. S. and M. J.M., Introduction to Modern Information Retrieval, New York: McGraw-Hill, Inc, 1986.

[16] G. Salton, Automatic text processing: The transformation, analysis, and retrieval of, Reading: Addison-Wesley, 1989.

[17] G. Salton, A. Wong and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM,* vol. 18, no. 11, pp. 613-620, 1975.

[18] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems,* pp. 3111-3119, 2013.

[19] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint,* 2013.

[20] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics,* vol. 5, pp. 135-146, 2017.

[21] J. Pennington, R. Socher and C. Manning, "Glove: Global vectors for word representation," in *Conference on empirical methods in natural language processing (EMNLP)*, 2014.

[22] J. Camacho-Collados, M. Pilehvar and R. Navigli, "Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities," *Artificial Intelligence,* vol. 240, pp. 36-64, 2016.

[23] R. Navigli and S. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence,* vol. 193, pp. 217-250, 2012.

[24] G. Miller, "WordNet: a lexical database for English," *Communications of the ACM,* vol. 38, no. 11, pp. 39-41, 1995.

[25] V. Nastase and M. Strube, "Decoding Wikipedia Categories for Knowledge Acquisition," *AAAI,* vol. 8, pp. 1219-1224, 2008.

[26] N. Matas, S. Martinčić-Ipšić and A. Meštrović, "Extracting domain knowledge by complex networks analysis of Wikipedia entries," in *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015.

[27] B. Karlo, M. Sanda, M. Ana and F. Guerra, "Short Texts Semantic Similarity Based on Word Embeddings," *CECIIS, Central European Conference on Information and Intelligent Systems,* 2019.

[28] R. Mihalcea, C. Corley and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," *Aaai,* vol. 6, no. 2006, pp. 775-780, 2006.

[29] D. Lenat, "CYC: A large-scale investment in knowledge infrastructure," *Communications of the ACM,* vol. 38, no. 11, pp. 33-38, 1995.

[30] V. Nastase and M. Strube, "Decoding Wikipedia Categories for Knowledge Acquisition," *AAAI,* vol. 8, pp. 1219-1224, 2008.

[31] I. Witten and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," pp. 25-30, 2008.

[32] M. Damashek, "Gauging similarity with n-grams: Language-independent categorization of text," *Science,* vol. 267, no. 5199, pp. 843-848, 1995.

[33] C. Manning, P. Raghavan and H. Schütze, "Introduction to information retrieval," *Natural Language Engineering,* vol. 16, no. 1, p. 100–103, 2010.

[34] T. Landauer, P. Foltz and D. Laham, "An introduction to latent semantic analysis," *Discourse processes,* vol. 25, no. 2-3, pp. 259-284, 1998.

[35] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint,* 2018.

[36] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014.

[37] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning* , 2008.

[38] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. John, N. Constant, M. Guajardo-Cespedes, Y. S. C. Tar and Y. Sung, "Universal sentence encoder," *arXiv preprint,* 2018.

[39] L. Han, A. Kashyap, T. Finin, J. Mayfield and J. Weese, "UMBC_EBIQUITY-CORE: Semantic textual similarity systems," in *Second Joint Conference on Lexical and Computational Semantics*, 2013.

[40] R. Sinoara, J. Camacho-Collados, R. Rossi, R. Navigli and S. Rezende, "Knowledge-enhanced document embeddings for text classification," *Knowledge-Based Systems,* vol. 163, pp. 955-971, 2019.

[41] P. Bojanowski, E. G., A. J. and T. Mikolov, "Enriching word vectors with subword information," in *Transactions of the Association for Computational Linguistics 5*, 2017.

[42] J. de Winter, S. Gosling and J. Potter, "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data," *Psychological methods,* vol. 21, no. 3, p. 273, 2016.

[43] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, "Selectivity-based keyword extraction method," *International Journal on Semantic Web and Information Systems (IJSWIS),* vol. 12, no. 3, pp. 1-26, 2016.

[44] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi and R. Zamparelli, "A SICK cure for the evaluation of compositional distributional semantic models," *LREC,* pp. 216-223, 2014.

[45] M. Lee, B. Pincombe and M. Welsh, "An empirical evaluation of models of text document similarity," *Annual Meeting of the Cognitive Science Society,* vol. 27, no. 27, 2005.

[46] H. Paulheim, "Knowledge graph refinement: A survey of approaches and evaluation methods," *Semantic we,* vol. 8, no. 3, pp. 489-508, 2017.