Machine Learning Based Finding of Similar Sentences from French Clinical Notes

Khadim Dramé[1,2](✉), Gayo Diallo[3], and Gorgoumack Sambe[1,2]

[1] Université Assane Seck de Ziguinchor, Ziguinchor, Senegal
{kdrame,gsambe}@univ-zig.sn
[2] Laboratoire d'Informatique et d'Ingénierie pour l'Innovation, Ziguinchor, Senegal
[3] SISTM - INRIA, BPH INSERM 1219, Univ. Bordeaux, 33000 Bordeaux, France
Gayo.Diallo@u-bordeaux.fr

Abstract. Finding similar sentences or paragraphs is a key issue when dealing with text redundancy. This is particularly the case in the clinical domain where redundancy in clinical notes makes their secondary use limited. Due to lack of resources, this task is a key challenge for French clinical documents. In this paper, we introduce a semantic similarity computing approach between French clinical sentences based on supervised machine learning algorithms. The proposed approach is implemented in a system called CONCORDIA, for COmputing semaNtic sentenCes for fRench Clinical Documents sImilArity. After briefly reviewing various semantic textual similarity measures reported in the literature, we describe the approach, which relies on Random Forest (RF), Multilayer Perceptron (MLP) and Linear Regression (LR) algorithms to build different supervised models. These models are thereafter used to determine the degrees of semantic similarity between clinical sentences. CONCORDIA is evaluated using traditional evaluation metrics, EDRM (Accuracy in relative distance to the average solution) and Spearman correlation, on standard benchmarks provided in the context of the DEFT 2020 challenge. According to the official results of this challenge, our MLP based model ranked first out of the 15 submitted systems with an EDRM of 0.8217 and a Spearman correlation coefficient of 0.7691. The post-challenge development of CONCORDIA and the experiments performed after the DEFT 2020 edition showed a significant improvement of the performance of the different implemented models. In particular, the new MLP based model achieves a Spearman correlation coefficient of 0.80. On the other hand, the LR one, which combines the output of the MLP model with word embedding similarity scores, obtains the higher Spearman correlation coefficient with a score of 0.8030. Therefore, the experiments show the effectiveness and the relevance of the proposed approach for finding similar sentences on French clinical notes.

Keywords: Sentence similarity · Machine learning · Random forest · Multilayer perceptron · French clinical notes

1 Introduction

Computing semantic similarity between sentences is a crucial issue for many Natural Language Processing (NLP) applications. Semantic sentence similarity is used in vari-

M. Marchiori et al. (Eds.): WEBIST 2020/2021, LNBIP 469, pp. 26–42, 2023.
https://doi.org/10.1007/978-3-031-24197-0_2

ous tasks including information retrieval and texts classification [11], question answering, plagiarism detection, machine translation and automatic text summarization [6,34]. Therefore, there has been a significant interest in measuring similarity between sentences. To address this issue, various sentence similarity approaches have been proposed in the literature [1,6,7]. The commonly used approaches exploit lexical, syntactic, and semantic features of sentences. In the lexical approaches, sentences are considered as sequences of characters. Therefore, common shared characters [40], tokens/words or terms [18] between the source and the target sentences are usually exploited for measuring sentence similarity. Some other approaches attempt to take into account synonymy issues and/or to capture semantics of sentences using external semantic resources or statistical methods [8]. In statistical approaches, different techniques are used to capture the semantics of sentences, among them latent semantic analysis [22] or words embedding [23,29]. On the other hand, knowledge-based approaches rely on semantic resources such as WordNet [31] for general domain or UMLS (Unified Medical Language System) [4] for the biomedical specific domain.

In recent evaluation campaigns such as SemEval, supervised learning approaches have been shown to be effective for computing semantic similarity between sentences in both general English [2,6] and clinical domains [37,39]. We noted also the emergence of deep learning-based approaches in more recent challenges such as n2c2/OHNLP challenge [42]. Moreover, deep learning-based models have achieved very good performances on clinical texts [42]. However, in the context of French clinical notes, because of the use of domain specific language and the lack of resources, computing effectively semantic similarity between sentences is still a challenging and open research problem. Similarly to international evaluation campaigns such as SemEval [6], BioCreative/OHNLP [37] and n2c2/OHNLP [42], the DEFT 2020 (DÉfi Fouille de Textes - text mining) challenge, aims to promote the development of methods and applications in NLP [5] and provides standard benchmarks for this issue [16,17].

This paper aims to address this challenging issue in the French clinical domain. We propose a supervised approach based on three traditional machine learning (ML) algorithms (Random Forest (RF), Multilayer Perceptron (MLP) and Linear Regression (LR)) to estimate semantic similarity between French clinical sentences. We assume that combining optimally various kinds of similarity measures (lexical, syntactic and semantic) in supervised models may improve their performance in this task. In addition, for semantic representation of sentences, we investigated word embedding in the context of French clinical domain in which resources are less abundant and often not accessible. This proposed approach is implemented in the CONCORDIA system, which stands for COmputing semaNtic sentenCes for fRench Clinical Documents sImilArity. The implemented models are evaluated using standard datasets provided by the organizers of DEFT 2020. The official evaluation metrics were EDRM (Accuracy in relative distance to the average solution) and Spearman correlation coefficient. According to the performance and comparison from the official DEFT 2020 results, our MLP based model outperformed all the other participating systems in the task 1 (15 submitted systems from 5 teams), achieving an EDRM of 0.8217. In addition, the LR and MLP based models obtained the higher Spearman correlation coefficient, achieving respectively 0.7769 and 0.7691. An extension of the models as proposed in the context of the DEFT 2020 challenge has significantly improved the achieved performance. In particular, the MLP-based model achieved a Spearman correlation coefficient of 0.80. On the other hand, the LR based model combining the predicted similarity scores of the MLP model with the word embedding similarity scores obtained the higher Spearman correlation coefficient with 0.8030.

The rest of the paper is structured as follows. Section 2 gives a summary of related work. Then, Sect. 3 presents our supervised approach for measuring semantic similarity between clinical sentences. Next, the official results of our proposal and some other experimental results on standard benchmarks are reported in Sect. 4 and discussed in Sect. 5. Conclusion and future work are finally presented in Sect. 6.

2 Related Work

Measuring similarity between texts is an open research issue widely addressed in the literature. Many approaches have been proposed particularly for computing semantic similarity between sentences. In [14], the author reviews approaches proposed in the literature for measuring sentence similarity and classifies them into three categories according to the used methodology: word-to-word based, structure-based, and vector-based methods. He also distinguishes between string-based (lexical) similarity and semantic similarity. String-based similarity considers sentences as sequences of characters while semantic similarity take into account the sentence meanings.

In lexical approaches, two sentences are considered similar if they contain the same words/characters. Many techniques based on string matching have been proposed for computing text similarity: Jaccard similarity [18,32], Ochiai similarity [33], Dice similarity [10], Levenstein distance [24], Q-gram similarity [40]. These techniques are simple to implement and to interpret but fail to capture semantics and syntactic structures of the sentences. Indeed, two sentences containing the same words can have different meanings. Similarly, two sentences which do not contain the same words can be semantically similar.

To overcome the limitations of these lexical measures, various semantic similarity approaches have been proposed. These approaches use different techniques to capture the meanings of the texts. In [7], authors describe the methods of the state of the art proposed for computing semantic similarity between texts. Based on the adopted principles, methods are classified into four categories: corpus-based, knowledge-based, deep learning-based, and hybrid methods.

The corpus-based methods are widely used in the literature. In general, they rely on statistical analysis of large corpus of texts using techniques like Latent semantic analysis (LSA) [22]. The emerging word embedding technique is also widely used for determining semantic text similarity [13,21]. This technique is based on very large corpus to generate semantic representation of words [29] and sentences [23].

The knowledge-based methods rely on external semantic resources. WordNet [31] is usually used in general domain [15] and sometimes even in specific domains like medicine [39]. UMLS (Unified Medical Language System) [4], a system that includes and unifies more than 160 biomedical terminologies, is also widely used in the biomedical domain [37,39]. Various measures have been developed to determine semantic similarity between words/concepts using semantic resources [19,25,38]. In [27], an open source tool (called UMLS-Similarity) has been developed to compute semantic similarity between biomedical terms/concepts using UMLS. Many approaches are based on these word similarity measures to compute semantic similarity between sentences [26,35]. The knowledge-based methods is sometimes combined with corpus-based methods [28,35] and especially with word embedding [13]. One limitation of the knowledge-based methods is their dependence on semantic resources that are not available for all domains.

However, in recent evaluation campaigns such as SemEval, supervised approaches have been the most effective for measuring semantic similarity between sentences in general [2,6] and clinical domains [37,39].

Recently, we noted the emergence of deep learning-based approaches in semantic representation of texts, particularly the word embedding techniques [23,29,30,36]. These approaches are widely adopted in measuring semantic sentence similarity [8,13] and are increasingly used. More advanced deep learning-based models have been investigated in the most recent n2c2/OHNLP (Open Health NLP) challenge [42]. Transformer-based models like Bidirectional Encoder Representations from Transformers (BERT), XLNet, and Robustly optimized BERT approach (RoBERTa) have been explored [43]. In their experiments on the clinical STS dataset (called MedSTS) [41], authors showed that these models achieved very good performance [43]. In [9], an experimental comparison of five deep learning-based models have been performed: Convolutional Neural Network, BioSentVec, BioBERT, BlueBERT, and ClinicalBERT. In the experiments on MedSTS dataset [41], BioSentVec and BioBERT obtained the best performance. In contrast to these works which deal with English data where resources are abundant, our study focuses on French clinical text data where resources are scarce or inaccessible.

3 Proposed Approach

In this section, we present the approach followed by CONCORDIA. Overall, it operates as follows. First, each sentence pair is represented by a set of features. Then, machine learning algorithms rely on these features to build models. For feature engineering, various text similarity measures are explored including token-based, character-based, vector-based measures, and particularly the one using word embedding. The top-performing combinations of the different measures are then adopted to build supervised models. An overview of the proposed approach is shown in Fig. 1.

3.1 Feature Extraction

Token-Based Similarity Measures. In this approach, each sentence is represented by a set of tokens/words. The degree of similarity between two sentences depends on the number of common tokens into these sentences.

The **Jaccard similarity** measure [18] of two sentences is the ratio of the number of tokens shared by the two sentences and the total number of tokens in both sentences. Given two sentences S1 and S2, X and Y respectively the sets of tokens of S1 and S2, the Jaccard similarity is defined as follows [12]:

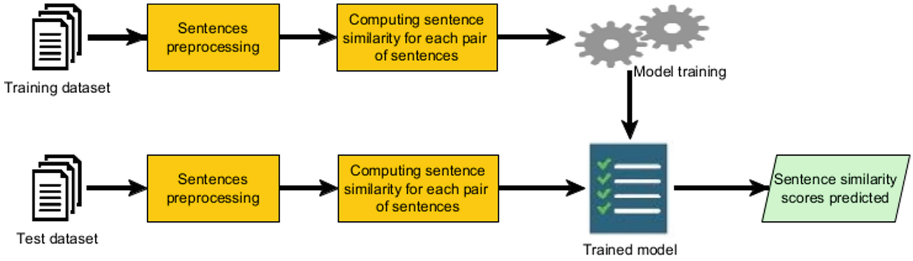
Training dataset
Sentences preprocessing
Computing sentence similarity for each pair of sentences
Model training
Test dataset
Sentences preprocessing
Computing sentence similarity for each pair of sentences
Trained model
Sentence similarity scores predicted

Fig. 1. Overview of the proposed approach [12].

$$sim_{Jaccard}(S1, S2) = \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

The **Dice similarity** measure [10] of two sentences is the ratio of two times the number of tokens shared by the two sentences and the total number of tokens in both sentences. Given two sentences S1 and S2, X and Y respectively the sets of tokens of S1 and S2, the Dice similarity is defined as [12]:

$$sim_{Dice}(S1, S2) = \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{2}$$

The **Ochiai similarity** measure [33] of two sentences is the ratio of the number of tokens shared by the two sentences and the square root of the product of their cardinalities. Given two sentences S1 and S2, X and Y respectively the sets of tokens of S1 and S2, the Ochiai similarity is defined as [12]:

$$sim_{Ochiai}(S1, S2) = \frac{|X \cap Y|}{\sqrt{|X| \times |Y|}} \tag{3}$$

The **Manhattan distance** measures the distance between two sentences by summing the differences of token frequencies in these sentences. Given two sentences S1 and S2, n the total number of tokens in both sentences and Xi and Yi respectively the frequencies of token i in S1 and S2, the Manhattan distance is defined as [12]:

$$d_{Manhattan}(S1, S2) = \sum_{i=1}^{n} |X_i - Y_i| \tag{4}$$

Character-Based Similarity Measures. The **Q-gram similarity** [40] is a character-based measure widely used in approximate string matching. Each sentence is sliced into sub-strings of length Q (Q-grams). Then, the similarity between the two sentences is computed using the matches between their corresponding Q-grams. For this purpose, the Dice similarity (described above) is applied using q-grams instead of tokens.

The **Levenshtein distance** [24] is an edit distance which computes the minimal number of required operations (character edits) to convert one string into another. These operations are insertions, substitutions, and deletions.

Vector-Based Similarity Measures. The Term Frequency - Inverse Document Frequency (TD-IDF) weighting scheme [20] is commonly used in information retrieval and text mining for representing textual documents as vectors. In this model, each document is represented by a weighted real value vector. Then, the cosine measure is used to compute similarity between documents. Formally, let $C = \{d_1, d_2, \ldots, d_n\}$, a collection of n documents, $T = \{t_1, t_2, \ldots, t_m\}$, the set of terms appearing in the documents of the collection and the documents d_i and d_j being represented respectively by the weighted vectors $d_i = (w_1^i, w_2^i, \ldots, w_m^i)$ and $d_j = (w_1^j, w_2^j, \ldots, w_m^j)$, their cosine similarity is defined as [12]:

$$Sim_{COS}(d_i, d_j) = \frac{\sum_{k=1}^{m} w_k^i w_k^j}{\sqrt{\sum_{k=1}^{m} \left(w_k^i\right)^2} \sqrt{\sum_{k=1}^{m} \left(w_k^j\right)^2}} \tag{5}$$

where w_k^l is the weight (TF.IDF value) of the term t_k in the document d_l. In the context of this work, the considered documents are sentences.

The **word embedding**, specifically the word2vec model [30], on the other hand, allows to build distributed semantic vector representations of words from large unlabeled text data. It is an unsupervised and neural network-based model that requires large amount of data to construct word vectors. Two main approaches are used to training, the continuous bag of words (CBOW) and the skip gram model. The former predicts a word based on its context words while the latter predicts the context words using a word. Considering the context word, the word2vec model can effectively capture semantic relations between words. This model is extended to sentences for learning vector representations of sentences [23]. Like the TF.IDF scheme, the cosine measure is used to compute the semantic sentence similarity.

Before applying token-based, vector-based and Q-gram similarity algorithms, preprocessing consisting of converting sentences into lower cases is performed. Then, the pre-processed sentences are tokenized using the regular expression tokenizers of the Natural Language Toolkit (NLTK) [3]. Therafter, the punctuation marks (dot, comma, colon, ...) and stopwords are removed.

3.2 Models

We proposed supervised models which rely on sentence similarity measures described in the previous section. For feature selection, combinations of different similarity measures (which constitute the features) were experimented. These supervised models require a labeled training set consisting of a set of sentence pairs with their assigned similarity scores. First, each sentence pair was represented by a set of features. Then, traditional machine learning algorithms were used to build the models, which were thereafter used to determine the similarity between unlabeled sentence pairs. Several machine learning algorithms were explored: Linear Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Extreme Learning Machine (ELM) and Multilayer Perceptron (MLP). Based on their performance on the validation set, we retained RF and MLP which outperformed the other models. In addition, we proposed a Linear Regression (LR) model taking as inputs

Table 1. Sample annotated sentence pairs. **Vote** indicates the gold similarity score between the two sentences [12].

Sentence 1	Sentence 2	Vote
La plupart des biberons d'étain sont de type balustre à tétine vissée sur pied (ou piédouche)	On a ensuite fait des biberons en étain et en fer blanc	0
La proportion de résidents ayant des prothèses dentaires allait de 62% à 87%	Dans toutes les études, la plupart des participants avaient des dentiers (entre 62% et 87%)	1
Les essais contrôlés randomisés, les essais cas-témoins et les études de cohorte comprenant des enfants et des adultes soumis à n'importe quelle intervention pour l'hématome aigu de l'oreille	Nous avons recherché des essais portant sur des adultes ou des enfants ayant subi un hématome	2
Les agents de déplétion du fibrinogène réduisent le fibrinogène présent dans le plasma sanguin, la viscosité du sang et améliorent donc le flux sanguin	Ils réduisent également l'épaisseur du sang (ou la viscosité), ce qui permet d'améliorer le flux sanguin jusqu'au cerveau	3
Refermez le flacon immédiatement après utilisation	Refermez l'embout du flacon avec le bouchon immédiatement après utilisation	4
La dose d'entretien recommandée est également de 7,5 mg par jour	La posologie usuelle est de 7,5 mg de chlorhydrate de moexipril par jour	5

the predicted similarity scores of both models and the average score of the different similarity measures.

An extension of the models proposed in the DEFT 2020 challenge were performed using several techniques. For this, we considered this sentence similarity computation task as regression problem. Thus, we used regressors to predict real values and then converted these values into integer values in the range [0–5] rather than multi-class classifiers. Furthermore, we used grid search technique to determine the optimal values of the models hyper-parameters. In addition, the LR model, instead of taking as inputs the predicted scores of the other models, combines scores predicted by the MLP model with the word2vec semantic similarity scores. The motivation is to better take into account the meanings of the sentences. For this purpose, we created a French clinical corpus of 70 K sentences partially from previous DEFT datasets.

4 Evaluation

In order to assess the proposed semantic similarity computing approach, we used benchmarks of French clinical datasets [16,17] provided by the organizers of the DEFT 2020 challenge. The EDRM (Accuracy in relative distance to the average solution) and the Spearman correlation coefficient are used as the official evaluation metrics [5].

We additionally used the Pearson correlation coefficient and the accuracy metrics. The Pearson correlation coefficient is commonly used in semantic text similarity evaluation [6,37,42], while the accuracy measure enables to determine the correctly predicted similarity scores.

4.1 Datasets

In the DEFT 2020 challenge, the organizers provided annotated clinical texts for the different tasks [16,17]. The task 1 of this DEFT challenge aims at determining the degree of similarity between pairs of French clinical sentences. Therefore, an annotated training set of 600 pairs of sentences and a testing set of 410 are made available. In total, 1,010 pairs of sentences derived from clinical notes are provided. Each sentence pair is manually annotated with a numerical score indicating the degree of similarity between the two sentences. The clinical sentence pairs are annotated independently by five human experts that assess the similarity scores between sentences ranging from 0 (that indicates the two sentences are completely dissimilar) to 5 (that indicates the two sentences are semantically equivalent). Then, scores resulting from the majority vote are used as the gold standard. Table 1 shows examples of sentence pairs in the training set with their gold similarity scores. The distribution of the similarity scores in the training set is highlighted in Fig. 2.

During the challenge, only the similarity scores associated with the sentence pairs in the training set are provided. Thus, the training set is partitioned into two datasets: a training set of 450 and a validation set of 150 sentence pairs. This validation set was then used to select the best subset of features but also to tune and compare machine learning models.

4.2 Results

The CONCORDIA proposed approach is experimented with different combinations of similarity measures as features for building the models. For each model, the results of the best combination are reported. The results of the proposed models on the validation set (please see Sect. 4.1) are presented in Table 2. According to the Pearson correlation coefficient, the MLP-based model got the best performance with a score of 0.8132. The MLP-based model slightly outperforms the RF-based model, while the latter yielded the highest Spearman correlation coefficient with a score of 0.8117. The LR-based model using predicted scores of the two other models as inputs got the lowest performance in this validation set.

Table 2. Results of the proposed models over the validation dataset [12].

Models	Pearson correlation	Spearman correlation
Random Forest model	0.8114	**0.8117**
Multilayer Perceptron model	**0.8132**	0.8113
Linear Regression model	0.8083	0.7926

Thereafter, the models were built on the entire training set using the best combinations of features, which yielded the best results in the validation set. Table 3 shows the official CONCORDIA results during the DEFT 2020 challenge [5,12]. According to the EDRM, the MLP model got significantly better results. We also note that the RF model performed better than the LR model, which combines the predicted similarity scores of the two other models. However, the latter yielded the highest Spearman correlation coefficient over the official test set.

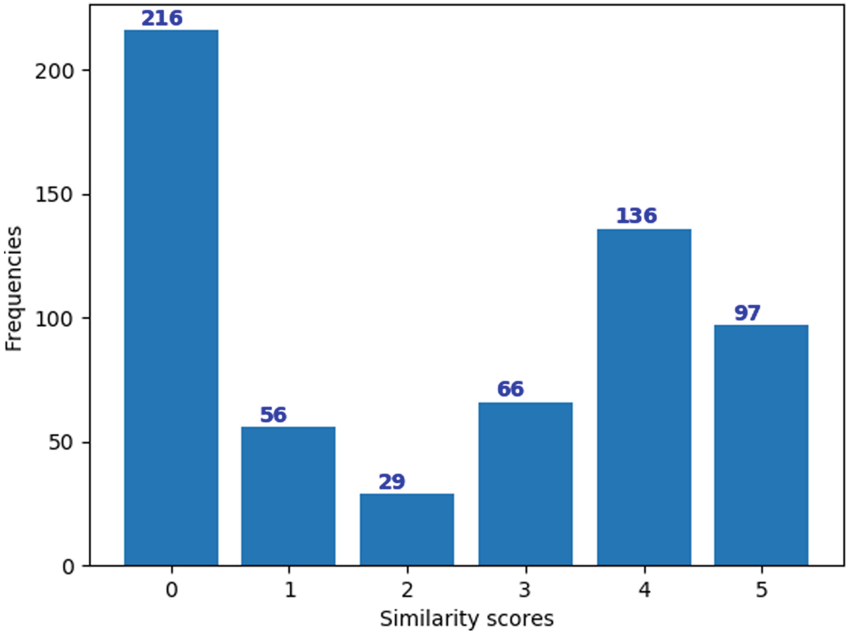

Fig. 2. Distribution of similarity scores in the training set [12].

Table 3. Results of the proposed models over the official test set of the DEFT 2020 [12].

Models	EDRM	Spearman correlation
Random Forest model	0.7947	0.7528
Multilayer Perceptron model	**0.8217**	0.7691
Linear Regression model	0.7755	**0.7769**

Compared to the other participating systems in the task 1 of the DEFT 2020 challenge, the proposed MLP model got the best performance (achieving an EDRM of 0.8217) [5]. Overall, CONCORDIA obtained EDRM scores higher than the average EDRM (0.7617). In addition, the two CONCORDIA best learning models, respectively the MLP model and the RF model, obtained EDRM scores greater than (for MLP) or equal to (for RF) the median score (0.7947). According to the Spearman correlation, the LR-based and MLP-based learning models got the best performance (respectively 0.7769 and 0.7691) out of all the other methods presented at the task 1 of the DEFT 2020 challenge.

Extension of the models proposed in the DEFT 2020 challenge are performed using several techniques. Table 4 shows the post challenge results of our improved models. The performances of the different models are significantly increased. In particular, the MLP based model now achieves a Spearman correlation of 0.80. On the other hand, the LR based model combining the predicted similarity scores of the MLP model and the word embedding similarity scores obtains the higher Spearman correlation with 0.8030.

Table 4. Results of the improved models over the official test set of the DEFT 2020.

Models	Pearson correlation	Spearman correlation
Random Forest model	0.8004	0.7948
Multilayer Perceptron model	0.8054	0.80
Linear Regression model	**0.8056**	**0.8030**

5 Discussion

5.1 Findings

The official results of the DEFT 2020 challenge showed that our approach is effective and relevant for measuring semantic similarity between sentences in the French clinical domain. Experiments performed after the challenge demonstrated also that word embedding semantic similarity can improve the performance of supervised models.

In order to estimate the importance of the different features in predicting the similarity between sentence pairs, the Pearson correlation coefficient of each feature is computed over the entire training dataset (please see Table 5). The findings show that the 3-gram and 4-gram similarity measures obtained the best correlation scores (respectively, 0.7894 and 0.7854). They slightly outperformed the semantic similarity measure based on the word embedding (0.7746) and the 5-gram similarity (0.7734). In addition, we noted that the Dice, Ochiai and TF.IDF based similarity measures performed well with correlation scores higher than 0.76. Among the explored features, the Levenshtein similarity was the less important feature (with a correlation score of 0.7283) followed by the Jaccard similarity (0.7354) and the Manhattan distance (0.7354). These results are consistent with those of the related work [8,39] although the word embedding based measure got the highest Pearson correlation coefficient in [39].

Table 5. Importance of each feature according to the Pearson correlation coefficient over the entire training set [12].

Feature	Pearson correlation
Q-gram similarity (Q = 3)	0.7894
Q-gram similarity (Q = 4)	0.7854
Word2vec similarity	0.7746
Q-gram similarity (Q = 5)	0.7734
Dice similarity	0.7644
TF-IDF similarity	0.7639
Ochiai similarity	0.7630
Jaccard similarity	0.7354
Manhattan distance	0.7354
Levenshtein similarity	0.7283

Using of together all these various similarity measures as features to build the models did not allow to increase their performance. On the contrary, it led to a drop of their performance. Thus, combinations of several similarity measures were experimented. The top-performing combination (which yield results presented in Sect. 4.2) was achieved with the following similarity measures: Dice, Ochiai, 3-gram, 4-gram, and Levenshtein. These findings show that these similarity measures complement each other and their optimal combination in supervised models allows to improve the models performance.

5.2 Comparison with Other Participating Systems

Most of systems submitted on the task 1 of the DEFT 2020 challenge mainly used string-based similarity measures (e.g. Jaccard, Cosine) or distances (Euclidean, Manhattan, Levenshtein) between sentences. Various machine learning models (e.g. Logistic Regression, Random Forest) were trained using these features [5]. Models of multilingual word embeddings derived from BERT (Bidirectional Encoder Representations from Transformers), in particular Sentence M-BERT and MUSE (Multilingual Universal Sentence Encoder) were equally developed but their performance were limited on this task. Compared with these systems, CONCORDIA explores more advanced features (e.g. word embedding) to determine the degree of similarity between sentences. In addition, instead of combining all the explored similarity measures as features, feature selection method were used to optimize the performance of our models. Furthermore, CONCORDIA is based on traditional ML algorithms for computing semantic sentence similarity.

5.3 Analysis of CONCORDIA Performance

Evaluation of the CONCORDIA semantic similarity approach on the DEFT 2020 dataset showed its effectiveness in this task. The results also demonstrated the relevance of the features used to measure similarity between French clinical sentences. Thus, all the CONCORDIA's learning models allowed to correctly estimate the semantic similarity between most of the sentence pairs of the official dataset. However, an analysis of the prediction errors using the Mean squared error (MSE) highlight variations of the models performance according to the similarity classes. Figure 3 shows the performance of our models over the official test set of the DEFT 2020 challenge. Overall, the LR model significantly made fewer errors. Moreover, the MLP model performed slightly better than the RF model in all similarity classes except class 4. These findings are consistent with the official results (Table 3) based on the Spearman correlation coefficient. The results also show that the RF and MLP models made fewer errors in predicting classes 5 and 0 but they performed much worse in predicting classes 2 and 3. We equally note that the proposed models, especially the RF model and the MLP model, struggled in predicting the middle classes (1, 2 and 3). Indeed, in the official test set, classes 1 and 2 are respectively 37 and 28. The RF model did not predict any value in both classes, while the MLP model predicted only 9 values of the class 1. The low performance in predicting these classes may be also attributed to the fact that they are less representative in the training dataset.

5.4 Limitations and Future Work

An extensive analysis of the results reveals limitations of CONCORDIA in predicting semantic similarity of some sentence pairs. The similarity measures used (Dice, Ochiai, Q-gram, and Levenshtein) struggle to capture the semantics of sentences. Therefore, our methods failed to correctly predict similarity scores for sentences having similar terms, but which are semantically not equivalent. For example, for sentence pair 224 (id = 224 in Table 6) in the test set, all methods estimated that the two sentences are roughly equivalent (with a similarity score of 4) while they are completely dissimilar according to the human experts (with similarity score of 0). On the other hand, our methods are limited in predicting the semantic similarity of sentences that are semantically equivalent but use different terms. For example, the sentences of pair 127 (id = 127 in Table 6) are considered completely dissimilar (with a similarity score of 0) while they are roughly equivalent according to the human experts (with a similarity score of 4). To address these limitations, we proposed a semantic similarity measure based on words embedding. But the combination of this semantic measure with the other similarity measures in supervised models led to a drop in performance.

Several avenues are identified to improve the performance of the proposed approach. First, we plan to explore additional similarity measures, especially those capable to capture the meanings of sentences. A post challenge experiment performed with word embedding on medium French corpus slightly improved the performance. Using a larger corpus could enable to increase significantly the performance. Furthermore, to overcome the limitation related to semantics, we plan the use of specialized biomedical resources, such as the UMLS (Unified Medical Language System) Metathesaurus. The latter contains various semantic resources, some of which are available in French (MeSH, Snomed CT, ICD 10, etc.). Another avenue would be to investigate the use of deep learning models such as BERT in the French clinical domain.

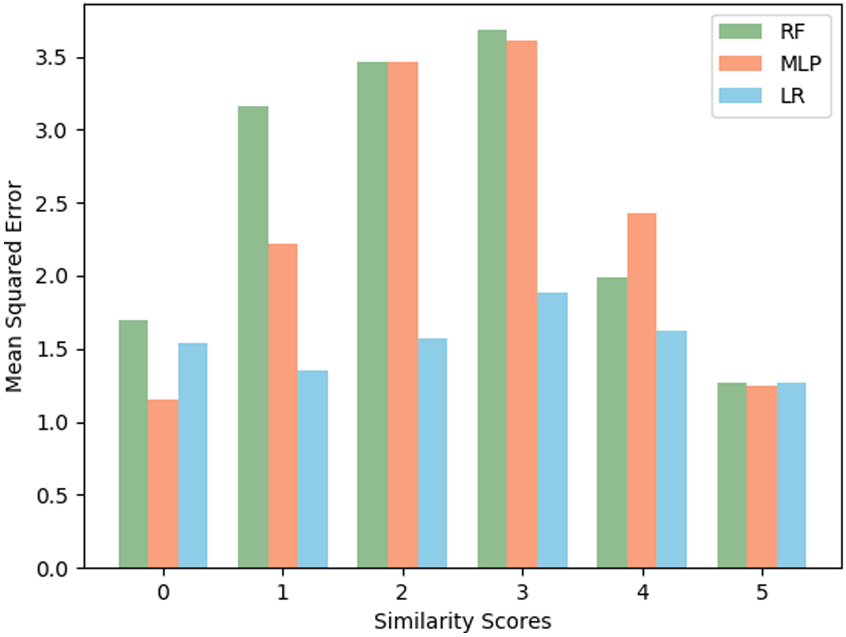

Fig. 3. The mean squared error of the proposed models according to the similarity classes over the test set [12].

Table 6. Sample similarity scores prediction of sentence pairs. **Vote** indicates the gold similarity scores while **Pred** indicates the predicted similarity scores.

Id	Sentence pair	Vote	Pred
42	Sentence 1: Ce médicament est contre-indiqué en cas d'hypersensibilité aux anesthésiques locaux ou à l'un des composants, et dans les situations suivantes Sentence 2: N'utilisez jamais Septanest 40 mg/ml adrenalinee au 1/200 000, solution injectable à usage dentaire en cas d'hypersensibilité (allergie) aux anesthésiques locaux ou à l'un des composants et dans les situations suivantes	4	4
127	Sentence 1: Eviter la prise de boissons alcoolisées et de médicaments contenant de l'alcool Sentence 2: La prise d'alcool est formellement déconseillée pendant la durée du traitement	4	0
224	Sentence 1: La persistance du canal artériel (PCA) est associée à une mortalité et une morbidité chez les nouveau-nés prématurés Sentence 2: Administration prophylactique d'indométacine intraveineuse pour prévenir la mortalité et la morbidité chez les nouveau-nés prématurés	0	4
338	Sentence 1: Nous avons évalué les bénéfices et les risques cliniques des agents stimulant l'érythropoïèse contre l'anémie dans la polyarthrite rhumatoïde Sentence 2: Qu'est-ce que l'anémie dans la polyarthrite rhumatoïde et que sont les agents stimulant l'érythropoïèse	0	4

6 Conclusion

In this paper, we presented the CONCORDIA approach which is based on supervised models for computing semantic similarity between sentences in the French clinical domain. Several machine learning algorithms have been explored and the top-performing ones (Random forest and Multilayer perceptron) retained. In addition, a Linear regression model combining the output of the MLP model with word embedding similarity were proposed. CONCORDIA achieved the best performance on a French standard dataset, provided in the context of an established international challenge, DEFT 2020 challenge. An extension of this approach after the challenge let to improve significantly the models performance. Several avenues to improve the effectiveness of the models are considered.

References

1. Agirre, E., et al.: SemEval-2015 task 2: semantic textual similarity, English, Spanish and pilot on interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 252–263. Association for Computational Linguistics, Denver (2015). https://doi.org/10.18653/v1/S15-2045, https://www.aclweb.org/anthology/S15-2045
2. Agirre, E., et al.: SemEval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 497–511. Association for Computational Linguistics, San Diego (2016). https://doi.org/10.18653/v1/S16-1081, https://www.aclweb.org/anthology/S16-1081
3. Bird, S., Loper, E.: NLTK: the natural language toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 214–217. Association for Computational Linguistics, Barcelona (2004). https://www.aclweb.org/anthology/P04-3031
4. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. **32**(Database issue), D267–D270 (2004). https://doi.org/10.1093/nar/gkh061, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC308795/
5. Cardon, R., Grabar, N., Grouin, C., Hamon, T.: Presentation of the DEFT 2020 Challenge: open domain textual similarity and precise information extraction from clinical cases. In: Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes, pp. 1–13. ATALA et AFCP, Nancy (2020). https://www.aclweb.org/anthology/2020.jeptalnrecital-deft.1
6. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 1–14. Association for Computational Linguistics, Vancouver (2017). https://doi.org/10.18653/v1/S17-2001, https://www.aclweb.org/anthology/S17-2001
7. Chandrasekaran, D., Mago, V.: Evolution of semantic similarity-a survey. ACM Comput. Surv. 54(2) (Feb 2021). https://doi.org/10.1145/3440755, https://doi.org/10.1145/3440755, place: New York, NY, USA Publisher: Association for Computing Machinery
8. Chen, Q., Du, J., Kim, S., Wilbur, W.J., Lu, Z.: Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. BMC Med. Inform. Decis. Making **20**(1), 73 (2020). https://doi.org/10.1186/s12911-020-1044-0

9. Chen, Q., Rankine, A., Peng, Y., Aghaarabi, E., Lu, Z.: Benchmarking effectiveness and efficiency of deep learning models for semantic textual similarity in the clinical domain: validation study. JMIR Med. Inform. **9**(12), e27386 (2021). https://doi.org/10.2196/27386
10. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945). https://doi.org/10.2307/1932409, https://app.dimensions.ai/details/publication/pub.1069656769, http://pdfs.semanticscholar.org/2304/5299013e8738bc8eff73827ef8de256aef66.pdf
11. Dramé, K., Mougin, F., Diallo, G.: Large scale biomedical texts classification: a kNN and an ESA-based approaches. J. Biomed. Semant. **7**, 40 (2016). https://doi.org/10.1186/s13326-016-0073-1
12. Dramé, K., Sambe, G., Diallo, G.: CONCORDIA: computing semantic sentences for French clinical documents similarity. In: Proceedings of the 17th International Conference on Web Information Systems and Technologies - WEBIST, pp. 77–83. INSTICC, SciTePress (2021). https://doi.org/10.5220/0010687500003058
13. Farouk, M.: Sentence semantic similarity based on word embedding and WordNet. In: 2018 13th International Conference on Computer Engineering and Systems (ICCES), pp. 33–37 (2018). https://doi.org/10.1109/ICCES.2018.8639211
14. Farouk, M.: Measuring sentences similarity: a survey. Indian J. Sci. Technol. **12**(25), 1–11 (2019). https://doi.org/10.17485/ijst/2019/v12i25/143977, http://arxiv.org/abs/1910.03940, arXiv: 1910.03940
15. Fernando, S., Stevenson, M.: A semantic similarity approach to paraphrase detection. In: Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics, pp. 45–52. Citeseer (2008)
16. Grabar, N., Cardon, R.: CLEAR - simple corpus for medical French. In: Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA), pp. 3–9. Association for Computational Linguistics, Tilburg (2018). https://doi.org/10.18653/v1/W18-7002, https://www.aclweb.org/anthology/W18-7002
17. Grabar, N., Claveau, V., Dalloux, C.: CAS: French corpus with clinical cases. In: Lavelli, A., Minard, A.L., Rinaldi, F. (eds.) Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, Louhi@EMNLP 2018, Brussels, Belgium, 31 October 2018, pp. 122–128. Association for Computational Linguistics (2018). https://aclanthology.info/papers/W18-5614/w18-5614
18. Jaccard, P.: The distribution of the flora in the alpine zone. 1. New Phytol. **11**(2), 37–50 (1912). https://doi.org/10.1111/j.1469-8137.1912.tb05611.x, https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x, _eprint: https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x
19. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th Research on Computational Linguistics International Conference, pp. 19–33. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Taipei (1997). https://aclanthology.org/O97-1002
20. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. (2004). https://doi.org/10.1108/00220410410560573, https://www.emerald.com/insight/content/doi/10.1108/00220410410560573/full/html
21. Kenter, T., de Rijke, M.: Short text similarity with word embeddings. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM 2015, pp. 1411–1420. Association for Computing Machinery, New York (2015). https://doi.org/10.1145/2806416.2806475
22. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discour. Process. **25**(2–3), 259–284 (1998). https://doi.org/10.1080/01638539809545028, _eprint: https://doi.org/10.1080/01638539809545028

23. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. arXiv:1405.4053 [cs] (2014)
24. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Sov. phys. Dokl. **10**, 707–710 (1965)
25. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning, ICML 1998, pp. 296–304. Morgan Kaufmann Publishers Inc., San Francisco (1998)
26. Liu, H., Wang, P.: Assessing sentence similarity using WordNet based word similarity. J. Softw. **8**(6), 1451–1458 (2013). https://doi.org/10.4304/jsw.8.6.1451-1458
27. McInnes, B.T., Pedersen, T., Pakhomov, S.V.: UMLS-interface and UMLS-similarity : open source software for measuring paths and semantic similarity. In: AMIA Annual Symposium Proceedings 2009, pp. 431–435 (2009). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815481/
28. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI 2006, vol. 1, pp. 775–780. AAAI Press, Boston (2006)
29. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv:1301.3781 [cs] (2013)
30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. arXiv:1310.4546 [cs, stat] (2013)
31. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995). https://doi.org/10.1145/219717.219748
32. Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S.: Using of jaccard coefficient for keywords similarity. In: Proceedings of The International MultiConference of Engineers and Computer Scientists 2013, pp. 380–384 (2013)
33. Ochiai, A.: Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions-II. Bull. Jpn. Soc. scient. Fish. **22**, 526–530 (1957). https://ci.nii.ac.jp/naid/10024483079
34. P, S., Shaji, A.P.: A survey on semantic similarity. In: 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), pp. 1–8 (2019). https://doi.org/10.1109/ICAC347590.2019.9036843
35. Pawar, A., Mago, V.: Calculating the similarity between words and sentences using a lexical database and corpus statistics. arXiv:1802.05667 [cs] (2018)
36. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014). http://www.aclweb.org/anthology/D14-1162
37. Rastegar-Mojarad, M., et al.: BioCreative/OHNLP challenge 2018. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB 2018, p. 575. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3233547.3233672
38. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, vol. 1, pp. 448–453. Morgan Kaufmann Publishers Inc., San Francisco (1995)
39. Soğancıoğlu, G., Öztürk, H., Özgür, A.: BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. Bioinformatics **33**(14), i49–i58 (2017). https://doi.org/10.1093/bioinformatics/btx238
40. Ukkonen, E.: Approximate string-matching with q-grams and maximal matches. Theor. Comput. Sci. **92**(1), 191–211 (1992). https://doi.org/10.1016/0304-3975(92)90143-4, https://www.sciencedirect.com/science/article/pii/0304397592901434
41. Wang, Y., et al.: MedSTS: a resource for clinical semantic textual similarity. Lang. Resour. Eval. **54**(1), 57–72 (2018). https://doi.org/10.1007/s10579-018-9431-1

42. Wang, Y., Fu, S., Shen, F., Henry, S., Uzuner, O., Liu, H.: The 2019 n2c2/OHNLP track on clinical semantic textual similarity: overview. JMIR Med. Inform. **8**(11), e23375 (2020). https://doi.org/10.2196/23375, https://medinform.jmir.org/2020/11/e23375. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics Publisher: JMIR Publications Inc., Toronto, Canada
43. Yang, X., He, X., Zhang, H., Ma, Y., Bian, J., Wu, Y.: Measurement of semantic textual similarity in clinical texts: comparison of transformer-based models. JMIR Med. Inform. **8**(11), e19735 (2020). https://doi.org/10.2196/19735, http://www.ncbi.nlm.nih.gov/pubmed/33226350