Proposal of a Method for Analyzing the Explainability of Similarity Between Short Texts in Spanish

Isidro Jara Matas, Luis de la Fuente Valentin, Alfonso Ortega de la Puente and Javier Sanz Fayos
Universidad Internacional de La Rioja, Avenida de la Paz 137, 26006 Logroño (La Rioja), Spain

Keywords: Explainability, Similarity Analysis, Short Text Grading.

Abstract: The aim of this project is the design and implementation of a system for analyzing the explainability of similarity between short texts (< 200 words), in Spanish language, with a special focus on the academic domain. For the system implementation, different models based on the BERT architecture will be used. A concise analysis of the explainability of the proposed system will be conducted, aiming to understand the intrinsic functioning of the method and to provide feedback to stakeholders, such as the author of the evaluated text or the professional deciding to use the system. Furthermore, based on the obtained results, an estimation of the system's goodness will be carried out through statistical analysis. This will enable both a comparison with other possible implementations and the proposal of future improvements that could have a positive impact on a more realistic assessment of texts.

1 INTRODUCTION

In recent decades, global access to information, and specifically to certain levels of formal education, often by technological resources, has significantly increased: from 1820 to 2020, the percentage of people aged 15 and older who received some type of formal primary, secondary, or tertiary education increased from 17.2% to 86.3% worldwide (Wittgenstein Center; World Bank; van Zanden, 2023). This has led to a scalability problem in the correction of academic tasks, making it necessary to have appropriate evaluation techniques that save time and effort. In this regard, Artificial Intelligence has produced new advances in the field of automatic text grading, especially for short texts (J. Zhang et al., 2022) (L. Zhang et al., 2022) (Tan et al., 2023), using different approaches and technologies.

The goal is to automatically evaluate a text without the need for human supervision, based on exclusively objective metrics. However, one of the remaining challenges is the explainability of these systems, i.e., why the system assigns a particular grade.

The term “explainability” (Zini & Awad, 2022) is often used to refer to an AI model constituted by a neural network that operates opaquely to its user. That means that, although the neural network is trained with a provided dataset and specific hyperparameters, such as the number of batches or the number and types of layers in the neural network, it can often be seen as a black box during the learning and prediction phases, as it does not provide information about the network's training, such as the weight assignments to each input. Hence, explainability is sought, i.e., the possibility for a human to understand why the machine does what it does. To date, numerous efforts have been made to unravel what happens inside this black box (Oh et al., 2019) (Schwartz-Ziv & Tishby, 2017).

In the field of NLP, the problem of the explainability appears in numerous issues across its different applications. While progress has been made in recent years to address these challenges some of them remain open.

A paradigmatic case is the similarity between short texts, which can be useful, for example, when searching for paraphrases of a sentence or finding insights within a text, i.e., information that may initially seem hidden but can shed light on the semantic field through a similar text, allowing this new information to inspire the system's user.

Another interesting scenario, which has gained significant importance over time (Burrows et al., 2015) and on which this work focuses, is the automatic grading of texts in the academic field, where the aim is to design a model capable of offering

Jara Matas, I., de la Fuente Valentin, L., Ortega de la Puente, A. and Sanz Fayos, J.
Proposal of a Method for Analyzing the Explainability of Similarity Between Short Texts in Spanish.
DOI: 10.5220/0013249500003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1017-1022
ISBN: 978-989-758-737-5; ISSN: 2184-433X

a grade to a student based on their exam or task, comparing each response with the ideally correct answers provided by the teacher.

However, this is where the problem of explainability arises.

The explainability problem of similarity between short texts has already been addressed with English texts (Malkiel et al., 2022) using innovative algorithms with BERT. However, from the state-of-the-art study, no references have been found addressing this issue with Spanish texts. Given its proven utility in various fields, particularly in education, it is necessary to provide a solution that opens the way to new proposals offering differential value.

Explainability is necessary not only for the student to access the grade with proper justification but also for other stakeholders to observe the system's functioning and understand the reasons behind the grading.

Therefore, the main objective is to propose and develop a method to analyze the explainability of the similarity between two sentences in Spanish that will allow understanding why a Natural Language Processing model decides whether two sentences are similar or not, and to what extent they are.

2 EXPERIMENT

For the experiment, three pairs of sentences and four NLP models were used, resulting in a total of twelve combinations.

The proposed sentence pairs cover three distinct and exclusive contexts: sociopolitical context, artistic context, and linguistic-philosophical context, in order to cover a broader range of language topics.

2.1 Pairs of Sentences

First Pair of Sentences: ‘El Gobierno pone en marcha los Presupuestos de 2024 y Alegría le pide al PP una oposición “constructiva”’ - Headline of El País, 11/23/2023; ‘Sánchez entrega una carta a sus ministros contra la oposición: "Niegan la legitimidad de origen a este Ejecutivo."’ - Headline of El Mundo, 11/23/2023.

Second Pair of Sentences: ‘No es tanto un drama erótico como una reflexión psicológica sobre el matrimonio, el deseo, los celos y la paranoia sexual’ - Review by Angie Errigo (Empire) about the film "Eyes Wide Shut"; ‘Fascinante, misteriosa, dura, agresiva, perturbadora, memorable. Cine insólito, magníficamente escrito, desasosegante, sensual, audaz, más que bueno.’ - Review by Carlos Boyero about the film "Eyes Wide Shut".

Third Pair of Sentences: ‘Relación de afecto, simpatía y confianza que se establece entre personas que no son familia’ - Definition of “friendship according” to the RAE; ‘Para Aristóteles, la amistad es un intercambio donde aprender a recibir y a otorgar'. - Definition of “friendship” according to Aristotle.

2.2 NLP Models

Model 1: hiiamsid/sentence_similarity_spanish_es. Developed by Siddhartha Shrestha, its main purpose is sentence similarity. It is based on SentenceTransformers: it maps sentences and paragraphs to a dense vector space of 768 dimensions, to later perform tasks such as clustering or semantic search.

Model 2: sentence-transformers/distiluse-base-multilingual-cased-v1. Also trained with SentenceTransformers for sentence similarity, except that the dimensional space of its vectors is reduced to 512. This model is based on a version of BERT known as DistilBERT, a smaller, faster, cheaper, and lighter Transformer model trained through the “distillation” of the BERT base model. This version of the model is multilingual, including Spanish, and is case-sensitive, so it differentiates between uppercase and lowercase letters.

Model 3: Sentence-transformers/distiluse-base-multilingual-cased-v2. Same definition as Model 2. The only noticeable difference is the higher number of downloads and HuggingFace spaces using each version of the two models, with more in the second case.

Model 4: dccuchile/bert-base-spanish-wwm-cased. The fourth model is BETO, the version of BERT trained with a large corpus of Spanish texts and published in 2020. BETO is one of the most popular BERT-based models, and specifically, one of the most well-known models trained in the Spanish language.

The main properties of each model used in the experiment are summarized in the following table:

Table 1: Comparison of the NLP models.

# Model	1	2	3	4
Purpose	Similarity between sentences			General
Trained with sentence-transformers	Yes	Yes	Yes	No
DistilBERT based	No	Yes	Yes	No
Dimensionality	768	512	512	768

2.3 Method

Once the combination of sentence pair and model is chosen, the experiment proceeds through the following phases:

2.3.1 Token Processing

In this phase, stop-words are removed. Stop-words are words without significant meaning, such as articles, pronouns, prepositions, etc., and they are filtered out before or after the natural language data processing, along with the punctuation symbols of the sentences.

2.3.2 Lemmatization

Each sentence is taken, and each of its tokens is separately lemmatized, excluding punctuation symbols and stop-words that were already classified as such in the SpaCy training process.

This process returns a list of tokens for each sentence.

2.3.3 Calculation of Embeddings

The embeddings of each token or word from the lists resulting from the previous stage are calculated and stored in the corresponding sentence array.

This process is carried out according to the model chosen at the beginning of the method and is applied to each specific word, in isolation and completely independent of its position in the sentence and its proximity to other words in the same sentence.

The result is a list for each sentence of the same size as the original list of tokens, containing a certain number of vectors of different dimensions depending on the characteristics of the embedding of the chosen model.

2.3.4 Calculation of Cosine Similarity

Finally, the cosine similarity between the two lists of embeddings is calculated. The result returns the top k of the original word pairs (i.e., without being lemmatized) from the documents that explain the similarity between the two texts, with k being a parameter set by the user (k = 5 for the experiment).

The result is an array composed of k triplets:

- In the first position, the word from the first document (a).
- In the second position, the word from the second document (b).
- In the third position, the degree of similarity between the two words as returned by the chosen model. The array is ordered based on the last parameter, that is, the similarity between the word pairs, from highest to lowest (s).

In this way, the similarity between short texts is explained by analyzing the similarity between individual words: it can be expected that the greater the similarity between the top k pairs of similar words between two documents, the greater the similarity between those two documents.

2.3.5 Algorithm

The described process can be represented in algorithmic form as follows:

Algorithm 1: Algorithm for analyzing the explainability of similarity.

Data: three pair of sentences
Result: k-ranking of most similar pair of words across each pair of sentences;
for *each model*
  for *each pair of sentences* **and** *each model* **do**
    split sentences into single words;
    remove stop-words;
    lemmatize single words;
    calculate embedding of single words;
      for *each a = embedding of word of the first sentence* **and** *each b = embedding of word of the second sentence* **do**
        s = cosine_similarity(a, b);
        save array(a,b,s);
    end
  full_result = sort array(a,b,s) according to s
  result = first k arrays of full_result
  end
end

3 RESULTS

This section shows the results obtained after the execution of the experiment for the twelve combinations, given by the three pairs of example sentences and the four models presented.

The nomenclature in the results tables for each of the models is the one that has been used throughout the work, and which for greater clarity is now explained:

- Model #1: sentence_similarity_spanish_es
- Model #2: SBERT multilingual cased v1
- Model #3: SBERT multilingual cased v2
- Model #4: BETO cased

Table 2: Result of Model #1.

Pair of sentences #1	Pair of sentences #2	Pair of sentences #3
('oposición', 'oposición', 1.0)	('sexual', 'sensual', 0.6473)	('establecer', 'otorgar', 0.6327)
('Gobierno', 'ministro', 0.6396)	('deseo', 'misterioso', 0.6233)	('confianza', 'amistad', 0.5607)
('pedir', 'negar', 0.6234)	('drama', 'misterioso', 0.5563)	('relación', 'intercambio', 0.5248)
('Gobierno', 'Ejecutivo', 0.5663)	('psicológico', 'misterioso', 0.4884)	('relación', 'amistad', 0.4933)
('oposición', 'negar', 0.5632)	('erótico', 'sensual', 0.4817)	('afecto', 'otorgar', 0.4804)

Table 3: Results of Model #2.

Pair of sentences #1	Pair of sentences #2	Pair of sentences #3
('oposición', 'oposición', 1.0)	('erótico', 'sensual', 0.784)	('afecto', 'otorgar', 0.7256)
('poner', 'negar', 0.753)	('paranoia', 'perturbadoro', 0.7443)	('establecer', 'otorgar', 0.6945)
('poner', 'entregar', 0.7035)	('sexual', 'sensual', 0.7405)	('afecto', 'amistad', 0.593)
('Gobierno', 'ministro', 0.6791)	('erótico', 'insólito', 0.6784)	('afecto', 'recibir', 0.5833)
('Gobierno', 'ministro', 0.6791)	('celo', 'insólito', 0.6523)	('establecer', 'recibir', 0.5738)

Table 4: Results of Model #3.

Pair of sentences #1	Pair of sentences #2	Pair of sentences #3
('oposición', 'oposición', 1.0)	('celo', 'audaz', 0.7922)	('establecer', 'otorgar', 0.7237)
('poner', 'entregar', 0.8012)	('erótico', 'sensual', 0.7539)	('afecto', 'otorgar', 0.7118)
('pedir', 'negar', 0.6956)	('paranoia', 'perturbadoro', 0.7446)	('persona', 'otorgar', 0.6592)
('marcha', 'entregar', 0.6912)	('celo', 'Cine', 0.7175)	('establecer', 'recibir', 0.6554)
('pedir', 'entregar', 0.671)	('celo', 'duro', 0.7115)	('afecto', 'amistad', 0.6369)

Table 5: Results of Model #4.

Pair of sentences #1	Pair of sentences #2	Pair of sentences #3
('oposición', 'oposición', 1.0)	('psicológico', 'escrito', 0.9243)	('simpatía', 'amistad', 0.8733)
('marcha', 'carta', 0.9382)	('sexual', 'escrito', 0.895)	('confianza', 'amistad', 0.8652)
('pedir', 'negar', 0.6956)	('psicológico', 'misterioso', 0.8905)	('simpatía', 'intercambio', 0.8529)
('marcha', 'oposición', 0.9237	('drama', 'misterioso', 0.8899)	('simpatía', 'aprender', 0.8469)
('marcha', 'negar', 0.9075)	('drama', 'escrito', 0.8859)	('familia', 'intercambio', 0.8457)

4 DISCUSSIONS

Based on the results obtained, the following assertions can be derived:

1. The objective evaluation using Word similarity offers very diverse results depending on the model being used.

This may suggest the need to apply a specific bias to each model to center the results and/or a normalization technique for all models, so that the range of output values is more similar and there are no large discrepancies.

2. In line with the above, but from a subjective perspective, Model 1 yields better results in terms of word pairs, despite having the lowest Word similarity among the four models.

3. On the other hand, Models 2 and 3 offer similar and acceptable word pairs and Word similarity values.

However, compared to the word pairs that explain the similarity in Model 1, the pairs in these two models do not seem very relevant in the context of the sentence.

This may be because Model 1 used vectors (embeddings) from a dense vector space of 768 dimensions, while in Models 2 and 3 the number of dimensions is reduced to 512.

4. Conversely, the results from Model 4 (BETO) are worse compared to the previous three models.

Most of the word pairs it marks as highly similar, with a high value in Word similarity, are either not similar from a subjective point of view or do not

reflect the supposed high relevance of these words in the sentences, or both cases occur.

These poor results from BETO could be due to several factors acting together, although the main reason that could explain these results is that BETO is actually a pre-trained model, meaning it has not been specifically trained for a particular task, and specifically not for a text similarity calculation task.

In fact, the model used is "BETO-base," or BETO in its base form. To address this, specific training on BETO for the text similarity calculation task would be required.

Moreover, if the same word appears in both sentences, and if this word is not a stop-word or punctuation mark, its Word similarity (objective evaluation of the word pair) will be equal to 1 and will top the list of word pairs with the highest similarity, regardless of the model used.

Finally, during the lemmatization process with Spacy, the word 'perturbadora' becomes 'perturbadoro', which is a non-existing word in Spanish (the correct form should be 'perturbador')

This does not pose a practical problem since the similarity calculation of the method is performed on the lemmas of the words, and not on the words themselves.

If one wanted to obtain appropriate results, fine-tuning of the lemmatization process would be required.

5 CONCLUSIONS

This work has highlighted the power of Natural Language Processing in developing a method for analyzing the explainability of similarity between short texts in Spanish.

Although the original purpose was to consider that the method would have a primary focus in the academic field, in view of the tests carried out and the results obtained, it has been determined that the proposed method for sentence similarity calculation is effective not only in this field but also in many other areas of Natural Language Processing.

A method has been proposed to analyze the explainability of similarity between short texts in Spanish by evaluating existing technologies that enabled its development, and by comparing four NLP models we conclude that models trained for specific tasks return better results in those activities than models trained with a corpus and a more general purpose. Similarly, it can be determined that there is evidence to suggest that the dimensionality of embeddings may affect the quality of results, with a directly proportional relationship between the number of dimensions and the results obtained.

In addition, by comparing the quality of the results, it has been proven that single objective assessment is not sufficient, and human inspection is necessary to consolidate the model that best performs the explainability of similarity calculation.

Upon completing this research, new possibilities open up for future developments of methods and systems to explain the similarity between short texts in Spanish: manual validation by experts to clarify the quality of the results with the proposed method; expand the scope of the experiment (the experiment conducted in this study considered three specific pairs of sentences and four NLP models based on Google BERT, returning the top k=5 pairs of sentences with the highest similarity). Future work should include a larger number of sentence pairs, i.e., a more extensive corpus that covers a broader spectrum of language; as well as testing other NLP models, whether based on BERT or not, and even architectures not based on Transformers; using and comparing the results with other similarities and distances, such as Jaro-Winkler and Levenshtein; as well as alternative metrics and algorithms like BLEU (Papineni et al., n.d.) and ROUGE (Lin, n.d.). However, it should be noted that such comparisons fall outside the scope of the present position paper.

Finally, the development of an interactive web application that allows the user to input two sentences and return the explanation of their degree of similarity based on the most similar word would increase the corpus size, and collecting user feedback (e.g., through icons or rating buttons), as well as democratizing the value of Artificial Intelligence.

ACKNOWLEDGEMENTS

This research is supported by the UNIR project MLX-PRECARM.

REFERENCES

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/S40593-014-0026-8/TABLES/11

Lin, C.-Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*.

Malkiel, I., Ginzburg, D., Barkan, O., Caciularu, A., Weill, J., & Koenigstein, N. (2022). Interpreting BERT-based

Text Similarity via Activation and Saliency Maps. *WWW 2022 - Proceedings of the ACM Web Conference 2022*, 3259–3268. https://doi.org/10.1145/3485447.3512045

Oh, S. J., Schiele, B., & Fritz, M. (2019). Towards Reverse-Engineering Black-Box Neural Networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11700 LNCS*, 121–144. https://doi.org/10.1007/978-3-030-28954-6_7/COVER

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*.

Schwartz-Ziv, R., & Tishby, N. (2017). *Opening the Black Box of Deep Neural Networks via Information*. https://arxiv.org/abs/1703.00810v3

Tan, H., Wang, C., Duan, Q., Lu, Y., Zhang, H., & Li, R. (2023). Automatic short answer grading by encoding student responses via a graph convolutional network. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2020.1855207

Wittgenstein Center; World Bank; van Zanden, J. et al. (2023). *Wittgenstein Center (2023); World Bank (2023); van Zanden, J. et al. (2014) – with major processing by Our World in Data. "No formal education" [dataset]*.

Zhang, J., Zhang, L., Hui, B., & Tian, L. (2022). Improving complex knowledge base question answering via structural information learning. *Knowledge-Based Systems*, *242*, 108252. https://doi.org/10.1016/j.knosys.2022.108252

Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2022). An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, *30*(1), 177–190. https://doi.org/10.1080/10494820.2019.1648300

Zini, J. El, & Awad, M. (2022). On the Explainability of Natural Language Processing Deep Models. *ACM Computing Surveys*, *55*(5). https://doi.org/10.1145/3529755.