Measuring Brazilian Portuguese Product Titles Similarity using Embeddings

Alan da Silva Romualdo[1], **Livy Real**[2], **Helena de Medeiros Caseli**[1]

[1]Federal University of São Carlos – Computing Department
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brazil

[2]Digital Lab – americanas s.a.
São Paulo – SP – Brazil

alan.romualdo@b2wdigital.com, helenacaseli@ufscar.br

livy.coelho@b2wdigital.com

Abstract. *Textual similarity deals with determining how similar two pieces of texts are, considering the lexical (surface forms) or semantic (meaning) closeness. In this paper we applied word embeddings for measuring e-commerce product title similarity in Brazilian Portuguese. We generated some domain-specific word embeddings (using Word2Vec, FastText and GloVe) and compared them with general-domain models (word embeddings and BERT models). We concluded that the cosine similarity calculated using the domain-specific word embeddings was a good approach to distinguish between similar and non-similar products, but the multilingual BERT pre-trained model proved to be the best one.*

1. Introduction

The importance of e-commerce for companies and the general population has grown in recent years and even more in 2020. According to the Brazilian Association of Electronic Commerce (ABComm)[1], there was a growth of 56.8% in the first half of 2020 compared with the first eight months of 2019, with a turnover of approximately 8 billion dollars. With the global situation of the pandemic due to the SARS-CoV-2 virus, there was a major migration from physical stores to digital media. According to Mastercard's Global Outlook 2021 report[2], it is expected that 20-30% of the operations that migrated to digital media during social isolation become permanent.

According to [Rodrigues et al. 2014], the fast growth of the internet and e-commerce had made many companies to see them as a very interesting way to expand their business. In addition to the several marketing advantages, such as the dynamic trading and the reduction of marketing costs, in the online environment there is a direct large-scale exposure of products for sale. These characteristics favor communication and assortment global dissemination and contribute to the evolution of logistics, tending to reach a broader population.

[1]https://www.ecommercebrasil.com.br/noticias/faturamento-do-e-commerce-brasileiro-2020/

[2]https://www1.folha.uol.com.br/mercado/2021/01/ate-30-do-aumento-do-comercio-eletronico-relacionado-a-covid-deve-ser-permanente.shtml

In e-commerce, advertisements are usually composed of images and texts used to illustrate and to describe the products for sale. In marketplaces[3], typically the products information come from the vendors. In this case, the information is normally presented in a non-standard format and with high variability in terms of specifications and characteristics described for similar products, making it difficult to group them even for human beings.

For the automatic matching of similar products, it is necessary to use text and/or image processing techniques that are capable of extracting relevant characteristics for measuring the similarity between products. In this paper, we only deal with the **textual similarity** and, therefore, we apply natural language processing (NLP) techniques capable of finding similar products based on their textual information.

To illustrate this problem, consider the products shown in Figure 1[4]. In this figure there are: a pair of similar products (the first two) which are both ink cartridge of the same color (magenta) for the same printer, an in-class product (a kit of cartridges) and an out-class product (a printer).

Figure 1. Example of products that should be considered similar (the first two), another one from the same product category (the third one) and a non-similar product (the fourth one).

This paper examines the hypothesis that the textual similarity calculated based on the semantic distance between product titles can be applied for finding similar products in a marketplace such as Americanas[5]. By proving this hypothesis, similar products could be matched together before being offered as options in response to a customer query, thus improving shopping experience. In this paper we investigate product titles similarity based on **word embeddings** and BERT pre-trained models.

The main contributions of this work are: (i) the evaluation of the applicability of different word embedding and contextualized language models in measuring textual similarity in the e-commerce domain; and (ii) the addressing of a poorly explored scenario of e-commerce for Brazilian Portuguese.

This work is organized as follows. Section 2 presents some Related Works for calculating textual similarity in specific and general domains. Section 3 describes the investigated approaches, the *corpus* used in our experiments and the experimental setup.

[3]Marketplaces are online platforms that gather sellers offering different products or services.

[4]Image taken from `https://www.americanas.com.br/` accessed in 06/15/2021.

[5]`http://www.americanas.com.br`

Section 4 presents the results which pointed out that cosine similarity calculated using multilingual pre-trained BERT model achieved the best discrepancy ability. Finally, section 5 closes this paper with some conclusions and proposals for future work.

2. Related Works

In [Alam et al. 2020], the authors present several relevant approaches for calculating textual similarity in the field of biomedicine, including cosine similarity using word embeddings generated by GloVe [Pennington et al. 2014], Word2Vec [Mikolov et al. 2013] and FastText [Bojanowski et al. 2017]. They concluded that the general-domain word embeddings built by those tools did not work well at the sentence or paragraph level in the field of biomedicine because they did not capture medical terms neither optimized the word embeddings for the specific domain. According to these authors, similarity measuring techniques for a specific domain must take into account the semantic relevance of the information in that domain since misinterpretations about the content can lead the experts to bad decisions.

In [Lo 2017], word embeddings were also used for calculating the lexical and structural similarity for all language pairs. By means of Word2Vec [Mikolov et al. 2013] and other topic analysis tools, the authors concluded that their new version of MEANT was a more accurate alternative to BLEU [Papineni et al. 2002] in evaluating translation quality for low-resource languages.

In [Rosa da Silva et al. 2017], the problem of categorizing offers in the context of price comparison sites was investigated. They compared two techniques for generating word embeddings: one that learns unsupervised word embeddings from millions of offer descriptions (using BOW), and another that learns supervised word inclusion using a convolutional neural network (CNN). The CNN model substantially outperformed their best BOW model.

According to [Aryal et al. 2019] and [Zhang et al. 2020], there are several effective ways to calculate textual similarity using word embeddings, but the most traditionally used measures are the **cosine similarity** and the Euclidean distance. These measures calculate the degree of similarity between two objects based on the coordinates of these objects in a vector space [Alam et al. 2020, Arts et al. 2017].

Recently, a new measure for automatic evaluation in text generation was proposed: the BERTScore [Zhang et al. 2020]. Similar to other measures, BERTScore calculates a similarity score for each token in the candidate sentence with each token in the reference sentence using previously trained contextualized representations from a BERT model. According to [Zhang et al. 2020], BERTScore showed a better correlation with human judgments and a better model selection performance than other measures used in comparison.[6] Also according to these authors, BERTScore proved to be more robust in challenging examples compared to other evaluated measures.

In this paper, we present experiments carried out to evaluate how word embeddings and contextualized language models perform in measuring the similarity between

[6]The evaluation was made by comparing BERTScore with the following measures: BLEU, METEOR, ROUGE-L, CIDER, SPICE, LEIC, BEER, EED, CHRF ++ and CHARACTER. See [Zhang et al. 2020] for details.

product titles in Brazilian e-commerce.

3. Experiments

In this work, we investigated the most applied approaches for textual similarity measurement: word embeddings and contextualized language models. For that, different word embeddings models for the specific domain of e-commerce were generated using Word2Vec [Mikolov et al. 2013], FastText [Bojanowski et al. 2017] and GloVe [Pennington et al. 2014]. Pre-trained general domain word embeddings and BERT [Devlin et al. 2019] models available for Portuguese were also used to compare the results.

3.1. Experimental setup

For training the domain-specific (e-commerce) word embedding models, a *corpus* granted by Americanas was used, containing about 7.490 million products, with titles and descriptions totaling approximately 8 billion words. A vocabulary of $455,031$ words was extracted from this *corpus* containing the words that occur at least 2 times in the whole *corpus*.

Using this *corpus*, we trained five **domain-specific WEs** using 30 training epochs, a learning rate of 0.025 and word embeddings dimension equal to 64[7]:

1. `FastText-spec SKIPGRAM` – FastText word embeddings trained using SkipGram, Americanas *corpus* and character ngram maximum size of 6;
2. `Word2Vec-spec SKIPGRAM` – Word2Vec word embeddings trained using SkipGram and Americanas *corpus*;
3. `FastText-spec CBOW` – FastText word embeddings trained using CBOW, Americanas *corpus* and character ngram maximum size of 6;
4. `Word2Vec-spec CBOW` – Word2Vec word embeddings trained using CBOW and Americanas *corpus*;
5. `Glove-spec` – GloVe word embeddings trained with Americanas *corpus*.

In addition to these five domain-specific word embeddings, six other **general-domain** were used in comparison, all of them with dimension equal to 300 and trained by NILC[8] [Hartmann et al. 2017]:

7. `FastText-NILC SKIPGRAM` – FastText word embeddings trained using SkipGram;
8. `Word2Vec-NILC SKIPGRAM` – Word2Vec word embeddings trained using SkipGram;
9. `FastText-NILC CBOW` – FastText word embeddings trained using CBOW;
10. `Word2Vec-NILC CBOW` – Word2Vec word embeddings trained using CBOW;
11. `Glove-NILC` – GloVe word embeddings;

Finally, we also used **BERT models** for Portuguese: the multilingual BERT[9] and the BERTimbau [Souza et al. 2020] Large and Base[10] models :

[7]It is worth mentioning that we also trained word embeddings with dimension equal to 300 but the results were worse.

[8]Available at: `http://www.nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc`

[9]Available at: `https://github.com/google-research/bert`

[10]Available at: `https://github.com/neuralmind-ai/portuguese-bert`

12. `mBERT` – multilingual BERT model trained for 104 languages, including Portuguese.
13. `BERTimbau Base` – BERT model trained for Portuguese, with 12 layers and $110M$ of parameters.
14. `BERTimbau Large` – BERT model trained for Portuguese, with 24 layers and $335M$ of parameters.

It is worth mentioning that it was not possible to train a BERT model for our e-commerce big *corpus* with the available hardware. On the machine available for the experiments, which has 126GB of RAM and 16 processing cores, but no GPU, the estimated training time was more than 120 days.

3.2. Test *corpus*

As previously mentioned, the experiments presented in this paper aim to find titles of similar products by means of static or dynamic, domain-specific or general-domain embeddings. To assess this task, we chose to work with different levels of similarity by dividing our test *corpus* into four sets each one with 100 pairs of product titles:

- `manual` – this set contains 100 pairs of product titles that have been manually marked as similar;
- `automatic` – this set contains 100 pairs of product titles that were marked as similar by a simple automatic pattern matching system;
- `in-class` – this set contains 100 pairs of product titles that are not similar, but belong to the same product category;
- `out-class` – this set contains 100 pairs of product titles selected at random and manually checked to ensure they were not similar and were not even in the same category.

In Table 1 we present some examples of pairs of product titles in each of these classes.

Table 1. Examples of product titles for each of the test classes.

Product title	Class
Cartucho de tinta epson t196320 magenta xp204/xp401 -t196320 Cartucho de tinta epson T196320 magenta P/XP104/XP204/XP401	`manual`
Cartucho Epson 196 magenta T196 320BR 5 ml Cartucho Epson 196 Preto 5ml T196120	`automatic`
Kit Refil Tinta Com 04 Cores Epson L3110 L3150 T544 Epson Original 544 K M Y C Cartucho de Tinta HP 664 Preto - F6V29AB	`in-class`
Cartucho de Tinta HP 662 Preto - CZ103AB - HP Impressora Multifuncional HP Ink Advantage 2776 Jato de Tinta Wi-Fi - Impressora + Copiadora + Scanner	`out-class`

The product titles in the `manual` class are both for Epson cartridge (*cartucho*), with the same model (t196320) and color (*magenta*). The product titles in the `automatic` class are also of a Epson cartridge 196 but for different colors (magenta and black, *preto*). The `in-class` products are, respectively: a ink refill kit (*kit refil tinta*) and a cartridge. Finally, the `out-class` products are, respectively: a cartridge and a multifunctional printer (*impressora*).[11]

[11]The dataset was built by Americanas and it is a proprietary dataset.

4. Results

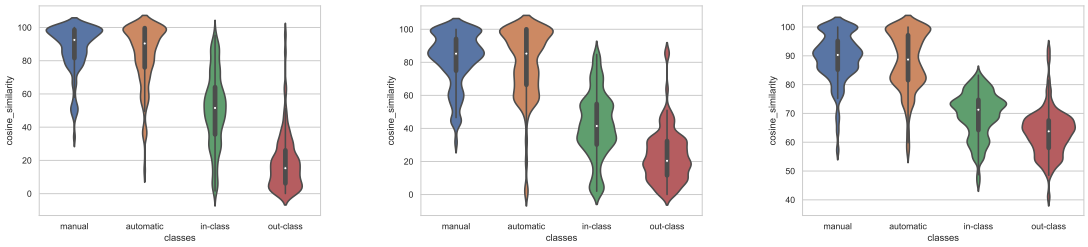
First, the average values of cosine similarity calculated using domain-specific word embeddings were compared with each other. Table 2 sumarizes the average cosine similarity values calculated using each domain-specific word embedding for each test set.

Table 2. Average values of cosine similarity calculated using domain-specific word embeddings

		positive class		negative class	
		manual	automatic	in-class	out-class
FastText-spec	SKIPGRAM	92.76	94.10	75.37	47.82
Word2Vec-spec		92.78	93.82	75.02	47.13
FastText-spec	CBOW	85.59	88.20	51.01	18.74
Word2Vec-spec		85.24	87.88	58.79	23.34
Glove-spec	–	90.76	89.48	66.36	39.83

Using domain-specific word embeddings we can see that the separation between positive and negative classes is very clear, with the pairs of titles from manual and automatic classes far from the other classes (by at least 18 points). All domain-specific models were able to differentiate well between positive (manual and automatic) and negative (in-class and out-class) classes. The **FastText model trained using CBOW** (FastText-spec CBOW) was the one with the largest margin between positive and in-class (about 34 points) product titles. The same model was also the one with the best largest distance between in-class and out-class (about 32 points) product titles.

Figure 2a shows the cosine distance values distribution, in each class, generated by FastText-spec CBOW model. From these values it is possible to set a threshold for similar products as, for instance, those with cosine similarity above 80.



(a) FastText-spec CBOW (b) Glove-NILC (c) mBERT

Figure 2. Cosine similarity values distribution, in each class, generated by the best models in each category: domain-specific (a), general-domain (b) and BERT (c).

In our second experiment, we evaluated the performance of the general-domain word embeddings in the same task, obtaining the average cosine similarity values presented in Table 3.

As expected, the average cosine similarity values calculated using general-domain word embeddings were lower than those calculated using domain-specific word embeddings. In this case, the model that best separated the classes was the Glove-NILC (with

Table 3. Average values of cosine similarity calculated using general-domain word embeddings

		positive class		negative class	
		manual	automatic	in-class	out-class
FastText-NILC	SKIPGRAM	88.80	89.72	66.90	55.86
Word2Vec-NILC		81.76	81.63	45.15	29.56
FastText-NILC	CBOW	83.42	82.93	50.52	38.62
Word2VeC-NILC		79.11	79.97	40.37	25.70
Glove-NILC	–	80.46	79.31	40.50	23.14

about 39 points between positive and in-class). However, all general-domain models were not so good in distinguishing in-class from out-class: Glove-NILC separating them by only 17 points. This fuzzy boundary between in and out-class products is easily observed in Figure 2b. Furthermore, in this case a threshold of 80 for similar products would label many of those products in automatic class as non-similar ones. Thus, the domain-specific word embeddings perform better than the general ones in calculating product title similarity.

Finally, we also evaluated how well the general-domain pre-trained BERT models – mBERT, BERTimbau Base and BERTimbau Large – can distinguish between title products from the four classes of similarity.

Table 4. Average values of cosine similarity calculated using BERT models

	positive class		negative class	
	manual	automatic	in-class	out-class
mBERT	86.01	85.35	62.93	54.51
BERTimbau Base	89.28	90.45	70.96	59.27
BERTimbau Large	93.88	94.81	85.57	77.60

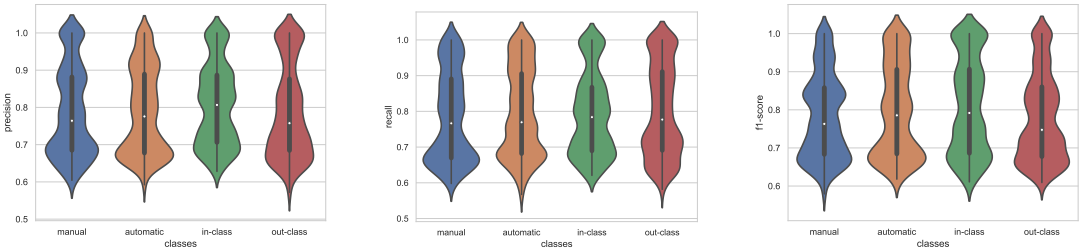
It can be noticed from the values in Table 4, that **mBERT** was the BERT model which best separated between positive and negative classes (with about 22 points between them).

From Figure 2 it is possible to notice that mBERT seems to be the best choice for separating between positive and negative classes. This insight is confirmed when we take a look at the numbers. For example, if we set a threshold of 80 for the cosine similarity, the number of instances incorrectly classified as similar are: 13 for FastText-spec CBOW, 5 for Glove-NILC and only 1 for mBERT. With the same threshold, the number of instances incorrectly classified as non-similar are: 55 for FastText-spec CBOW, 78 for Glove-NILC and 56 for mBERT. So, the best model for calculating product title similarity was the **general-domain pre-trained mBERT model**.

We also investigated if BERTScore [Zhang et al. 2020] calculated using mBERT would lead to better results. However, as can be noticed from the values in Table 5, the BERTScore calculated using mBERT was not so good as the other models in separating the similar products from those not similar. The inadequacy of BERTScore for this task is easily noticed when we analyse the graphs in Figure 3, where it is impossible to clearly separate the classes.

Table 5. Average values for BERTScore calculated using general-domain `mBERT` model

	positive class		negative class	
	manual	automatic	in-class	out-class
Precision	89.02	89.48	71.96	66.12
Recall	88.71	88.59	72.54	66.64
F1	88.80	88.99	72.20	66.35



(a) `mBERT`'s precision (b) `mBERT`'s recall (c) `mBERT`'s F1

Figure 3. BERTScore values distribution, in each class, generated by `mBERT`.

5. Conclusions and Future Work

From the results of the experiments presented in this paper, we can conclude that domain-specific word embeddings are effective in measuring the similarity between product titles. Among the domain-specific models we trained, the FastText with CBOW showed the best results. However, the best approach for distinguishing between similar and non-similar products was calculating the cosine similarity using the multilingual pre-trained general-domain BERT model.

As future work we intend to fine-tune the Brazilian Portuguese BERTimbau model [Souza et al. 2020] for our task and measure how well a domain-specific fine-tuned BERT model, for Portuguese, would perform in calculating product title similarity. Another proposal for future work is to expand our product title similarity task by including image processing techniques in order to develop a multimodal system.

Finally, although the experiments present in this paper were carried out for Brazilian Portuguese, the product title similarity measuring approach evaluated here is language independent and can be easily replicated for other idioms.

Acknowledgments

This paper and the research behind it would not have been possible without the support of americanas s.a. Digital Lab, specially José Pizani, Ester Campos and Allan Batista, who closely followed this research. This work is is part of the project "Dos dados ao conhecimento: extração e representação de informação no domínio do e-commerce" (Projeto de extensão - UFSCar #23112.000186/2020-97).

References

Alam, F., Afzal, M., and Malik, K. M. (2020). Comparative analysis of semantic similarity techniques for medical text. In *2020 International Conference on Information Networking (ICOIN)*, pages 106–109.

Arts, S., Cassiman, B., and Gomez, J. C. (2017). Text matching to measure patent similarity. *Strategic Management Journal*, 39.

Aryal, S., Ting, K. M., Washio, T., and Haffari, G. (2019). A new simple and effective measure for bag-of-word inter-document similarity measurement. *arXiv preprint arXiv:1902.03402*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Hartmann, N. S., Fonseca, E. R., Shulby, C. D., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Proceedings of Symposium in Information and Human Language Technology*, pages 122–131, Uberlândia, MG, Brasil. SBC.

Lo, C.-k. (2017). MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rodrigues, E. L., Fernandes, L. A., Rodrigues, E. F., de Arruda, I. P., and Moia, R. P. (2014). A importância da distribuição no comércio eletrônico. *INOVAE-Journal of Engineering, Architecture and Technology Innovation (ISSN 2357-7797)*, 1(1):24–38.

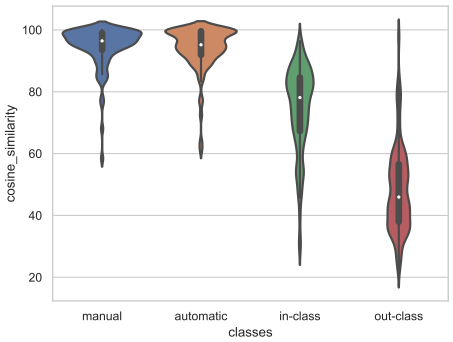
Rosa da Silva, R., Fernandes, E., Motta, E., Akira, E., Guarino, R., and Alvim, L. (2017). Offer categorization for price comparison websites: Word embedding approaches. In Martí, L. and Sánchez Pi, N., editors, *Anais do 13 Congresso Brasileiro de Inteligência Computacional*, pages 1–12, Curitiba, PR. ABRICOM.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In Cerri, R. and Prati, R. C., editors, *Lecture Notes in Computer Science*, volume 12319, pages 403–417, Cham. Springer International Publishing.

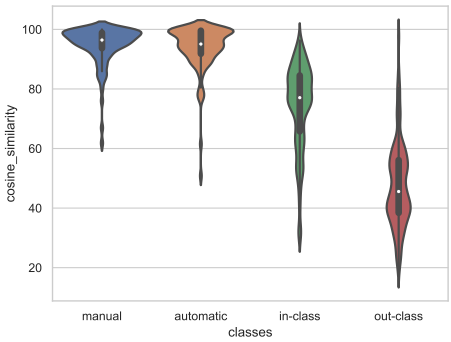
Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations*.

A. Violin plots

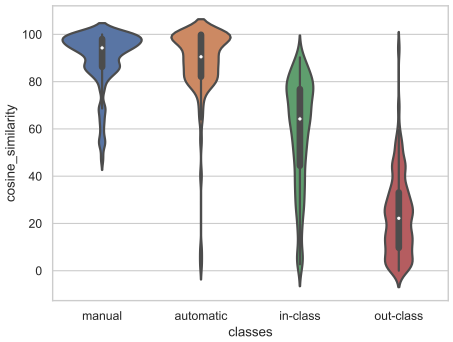
In this Appendix we group all the violin plots for the cosine similarity values calculated using all the domain-specific and general-domain word embeddings.



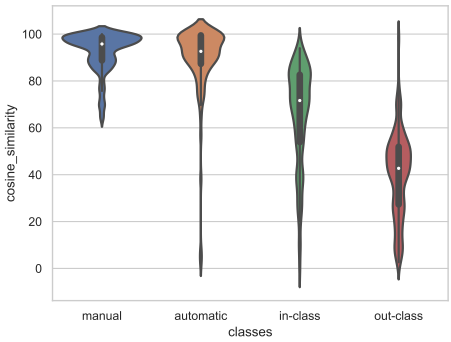
(a) `FastText-spec SKIPGRAM`



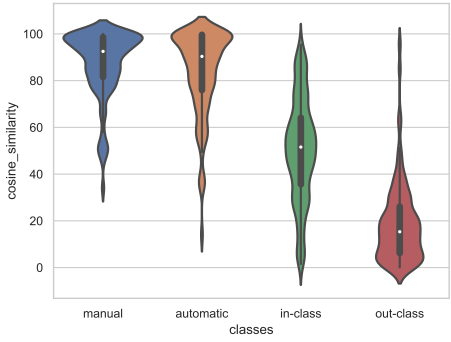
(b) `Word2Vec-spec SKIPGRAM`



(c) `Word2Vec-spec CBOW`

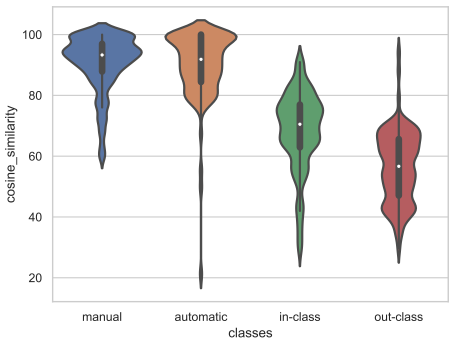


(d) `Glove-spec`

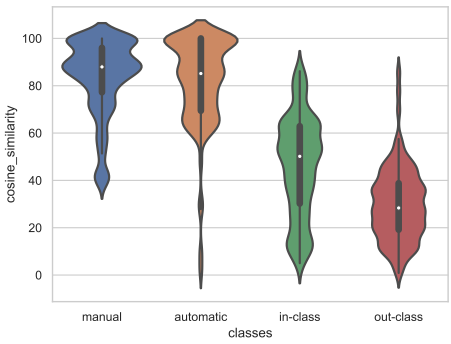


(e) `FastText-spec CBOW`

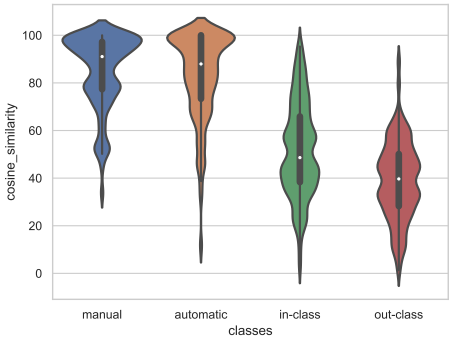
Figure 4. Cosine similarity values distribution, in each class, generated by domain-specific word embeddings.



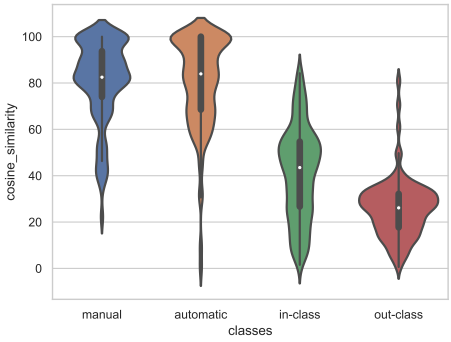
(a) **FastText-NILC SKIPGRAM**



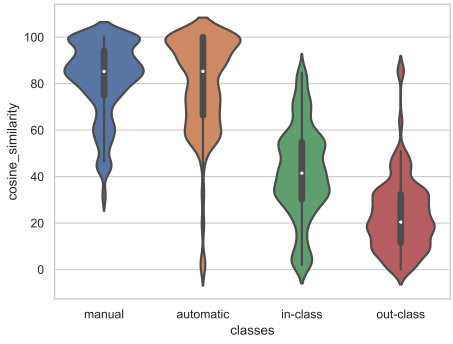
(b) **Word2Vec-NILC SKIPGRAM**



(c) **FastText-NILC CBOW**

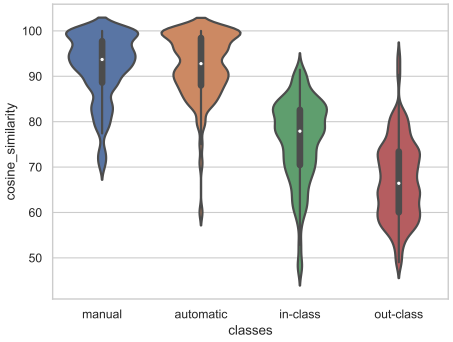


(d) **Word2Vec-NILC CBOW**

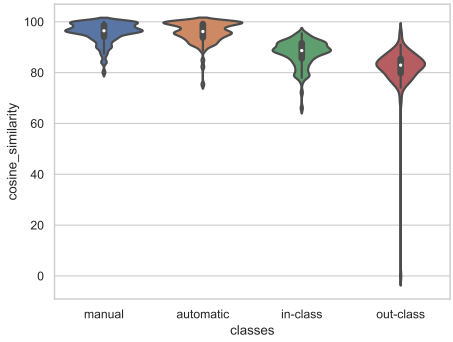


(e) **Glove-NILC**

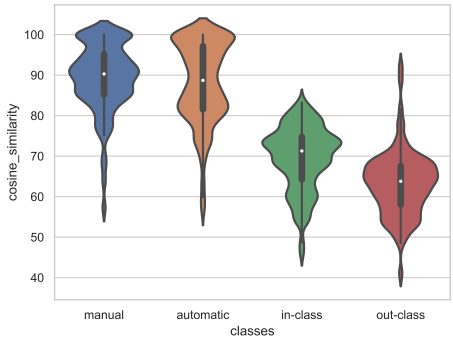
Figure 5. Cosine similarity values distribution, in each class, generated by general-domain word embeddings.



(a) **BERTimbau Base**



(b) **BERTimbau Large**



(c) **mBERT**

Figure 6. Cosine similarity values distribution, in each class, generated by general-domain BERT models.