Detection and Censorship of Offensive Language in Extended Texts in Portuguese

Lucas Lenoch de Souza[1,2], **Franciele Beal**[1,3],
André Roberto Ortoncelli [1,2,4], **Marlon Marcon**[1,2,4]

[1]Federal University of Technology - Paraná - UTFPR

[2]Software Engineering Coordination - Dois Vizinhos - Paraná - Brazil

[3]Academic Department of Informatics - Pato Branco - Paraná - Brazil

[4]Programa de Pós-graduação em Informática (PPGI)

lucasouza191141@gmail.com

{fbeal, ortoncelli, marlonmarcon}@utfpr.edu.br

Abstract. *This article addresses the problem of detecting and censoring offensive language in extensive Brazilian Portuguese texts on the web. This paper proposes a pipeline for classifying and censoring extensive texts, focusing on comments, posts, and articles using NLP techniques. The results include an in-depth review of current methods for offensive content classification in Portuguese and the implementation of a BERTimbau-based pipeline for offense detection. This work represents a significant advancement in the state-of-the-art NLP in Portuguese, promoting safer and more respectful online environments for users, especially children.*

1. Introduction

In recent years, the Internet has been growing at an impressive rate in terms of users and the data generated by web page publications. This increasing Internet use often introduces many children to virtual environments from a very early age. Consequently, offensive language in online texts becomes a concern for ethical reasons [Economist 2019]. Issues such as cyberbullying, hate speech, and various forms of offensive content in social media posts are also relevant [Cook 2024].

Regarding the intelligent interpretation of textual data from social networks, Natural Language Processing (NLP) is commonly used. To address the problem of offensive language in web posts, articles in the field of NLP, combined with Machine Learning (ML) and Deep Learning (DL) techniques, have been developed [Hajibabaee et al. 2022]. These efforts include creating research pipelines and developing high-quality annotated datasets by professionals [Leite et al. 2020].

Given the differences in languages and cultures, it is only possible to formalize a model for some languages. However, efforts are being made to learn various languages [Husain and Uzuner 2021]. The HateBR [Vargas et al. 2022] corpus is an example for Brazilian Portuguese. Such an article highlights the lack of academic production related to offensive language in the national language, along with the dataset and its results. Despite existing academic contributions to addressing offensive language in Brazilian

Portuguese, applying NLP and ML/DL techniques to classify longer texts (such as news articles or blog posts) remains challenging. These techniques are often limited to social media posts or news comments.

Based on this information, the present work focuses on detecting and subsequently censoring or filtering offensive language online. The goal is to ensure that web pages that are not explicitly focused on adult content can be suitable environments for children. Additionally, efforts are made to reduce prejudice and offensiveness in online posts, benefiting users' emotional well-being and promoting mutual respect. Specifically, we propose identifying offensive words using DL techniques, which can then be filtered or censored. We define small texts as comments, posts, and sentences, while paragraphs and entire pages constitute more extended texts, with the latter being the focus of this study.

As the main contributions of this work, we have: 1) An in-depth review of state-of-the-art methods applied to offensive content classification for the Portuguese language; 2) An update to the state-of-the-art results on the HateBR dataset [Vargas et al. 2022]; and 3) A Deep-Learning-based pipeline that effectively classifies and censors extensive Brazilian Portuguese texts based on their offensiveness.

2. Related Works

When considering works that address text processing with offensive language in the Brazilian context, the range of existing works is very small, including ToLD-BR [Leite et al. 2020], OffComBr [Pelle and Moreira 2017], HateBR [Vargas et al. 2022], and OLID-BR [Trajano et al. 2023]. Such methods are explained in the following sections.

2.1. ToLD-BR

This dataset [Leite et al. 2020], presents an unspecified number of posts extracted from the Twitter platform. A total of 42 individuals, chosen from 129 volunteers, were tasked with annotating each post, classifying them into various categories of prejudice: homophobia, obscene language, misogyny, and xenophobia.

The posts were classified using BERT-style algorithms [Devlin 2018], which achieved optimal results for such a complex and subjective task. The researchers also explored the possibility of building models for this task in multiple languages, but their results indicated that monolingual data is still preferable for more accurate classifications.

2.2. OffComBr

This paper presents the development of a dataset comprising comments on news articles derived from the website g1.globo.com, named "OffComBr" [Pelle and Moreira 2017]. The researchers obtained around 10,000 comments, but given the manual annotation process done by three experts in detecting offensive language, they included only 1,250 comments in the final dataset.

The authors performed two classification algorithms (SMO and Naive Bayes) to evaluate the dataset, with different assessments depending on different data preprocessing methods. Two versions of the dataset were developed, OffComBR-2 and OffComBR-3, with the difference being the size, as the latter retained from the former only the annotations agreed upon by all three experts.

2.3. HateBR

The work of [Vargas et al. 2022] presented the first large annotated corpus of offensive language in Instagram comments in Brazilian Portuguese. Motivated by the presence of hate speech on social media and the lack of studies on the subject in Portuguese, the project collected 7,000 Instagram comments, annotated by experts regarding the presence, degree, and category of offensiveness. The process involved data collection, selection of accounts of Brazilian political figures (three left-wing and three right-wing), and the selection of 30 posts from which 15,000 comments were extracted, with 7,000 being balanced between offensive and non-offensive.

The comments were labeled into three levels of offensiveness: whether they were offensive or not, the degree of offensiveness (mild, moderate, or high), and whether they contained hate speech, categorized into nine types such as xenophobia, racism, and homophobia. From the 7,000 comments, 3,500 were offensive, with 778 highly offensive, 1,044 moderately offensive, and 1,678 mildly offensive. Among the offensive comments, 727 contained some type of hate speech. Table 1 presents samples of data from HateBR.

Table 1. Examples of comments extracted form the HateBR dataset.

Class	Comments
Offensive	Essa **besta humana** é o **câncer** do País, tem que voltar para a jaula, urgentemente! E viva o Presidente Bolsonaro.
Non-Offensive	Quem falou isso para você deputada? O Sergio Moro está aprovado pela maioria dos brasileiros.
With hate speech	**Vagabunda. Comunista. Mentirosa.** O povo chileno não merece uma desgraça dessa.
Without hate speech	Pois é, deveria devolver o dinheiro aos cofres públicos do Brasil. **Canalha**.

Finally, after a detailed explanation of their entire annotation system, as well as evaluations to judge the annotations of each of the three experts and decide the most appropriate annotations for each comment, the study presents the test results with some ML models trained on the HateBR corpus, comparing the best result obtained with the best results of two other reference works. In this work, we seek to replicate the results obtained by HateBR, following the same training procedure and using the same models for comparison with our trained model.

2.4. OLID-BR

The work of [Trajano et al. 2023] also developed an annotated dataset of offensive comments in Portuguese, similar to HateBR. However, the main advantage of this dataset lies in its application to various NLP tasks, including binary classification of offensiveness, multi-category prediction of the type of toxicity, identification of targeted toxic comments, prediction of the target of toxicity, and identification of toxicity spans in comments.

The primary focus of the work was on the task of identifying toxicity spans, which involves detecting sequences of characters containing offensive language. To collect data, OLID-BR used various sources such as Twitter, YouTube, and other datasets with different annotation schemes.

The annotation was conducted in three stages: detection of offensive language, categorization of offensive language, and identification of the target of the offense. Com-

pared to HateBR, OLID-BR distinguishes between offensiveness against an individual, a group, or another type of target, while HateBR focuses on categorizing hate speech.

Data annotation in OLID-BR was not exclusively done by humans but also with the assistance of the Perspective API[1], allowing human annotators to correct the classifications. The entire corpus was divided into three datasets for training and testing, with a similar distribution of classifications in each. This work replicated the part of OLID-BR related to the identification of toxicity spans, using the code available on GitHub to train the model and apply it to the HateBR dataset for detecting offensive phrases and to the OLID-BR for identifying offensive spans.

3. Main Technologies

For the development, training, and testing of techniques for offensive language detection and censorship, we primarily used two libraries available for the Python language for developing ML algorithms: Transformers and spaCy.

3.1. Transformers

The Transformers library is a Python tool that offers state-of-the-art architectures for NLP tasks, featuring over 32 pre-trained models in more than 100 languages. It provides deep interoperability between TensorFlow 2.0 and PyTorch. The library is named after the Transformer architecture introduced by Google Brain in 2017, which is based on the "attention mechanism." This mechanism allows the model to focus on important parts of the input data, leading to superior performance in NLP tasks like sentence classification, named entity recognition (NER), and natural language generation compared to previous models like recurrent neural networks (RNNs) [Vaswani et al. 2017].

3.2. BERT and BERTimbau

The algorithm used in the first stage of our pipeline was a fine-tuned version BERTimbau [Souza et al. 2020], a Brazilian model based on BERT (Bidirectional Encoder Representations from Transformers) [Devlin 2018]. BERT, introduced by Google in 2018, is a language model that generates numerical representations for words based on their surrounding context and is used for various NLP tasks. BERTimbau adapts BERT for Brazilian Portuguese using transfer learning, where a BERT model was trained on a Portuguese corpus (brWaC) [Wagner Filho et al. 2018] and evaluated on tasks like sentence similarity, textual entailment, and named entity recognition. In this work, BERTimbau was specifically used to develop a model for detecting offensive language in sentences.

3.3. SpaCy

SpaCy is an open-source NLP library for Python, written in Cython [Honnibal et al. 2020], that facilitates tasks like part-of-speech tagging, named entity recognition (NER), and dependency parsing. It provides pre-trained models and allows users to train their own models for NER, where sentences are segmented into words, each categorized (e.g., nouns, adverbs, or specific problem-related categories like offensive and non-offensive). In this work, SpaCy was used in the second stage to detect which words in an offensive sentence are offensive, following the methodology of OLID-BR, which also used SpaCy for tasks like detecting offensive spans in text.

[1] https://www.perspectiveapi.com/

4. Methodology

This study focused on developing NLP models to detect offensive language in extended texts. Both the acquisition of training data and the actual censorship of words detected as offensive were carried out simplified due to them not being the primary focus. The development process consisted of four stages: data collection, cleaning/tokenization, model training, and result evaluation explained in the following subsection.

4.1. Data Collection

For data collection, we employ the HateBR [Vargas et al. 2022] and OLID-BR [Trajano et al. 2023] datasets to train the offensive content detection models, the former to classify a text segment as potentially offensive or not and the latter to identify words or expressions that contain offensiveness. Additionally, to evaluate qualitatively our solution, we selected news articles from the G1 portal [Monteiro 2023], Catraca Livre [Leray 2023], and two blog posts from Senso Incomum [Trielli 2021, Martins 2022].

4.2. Data Cleaning and Feature Extraction

We used tools from the Transformers and spaCy libraries for data cleaning and feature extraction. Specifically, for the model trained with the Transformers library, we employed a tokenizer ready to transform sentences into numerical representations used by the model for calculations. For the model trained with spaCy, an embedded tokenization functionality was available. In summary, in this step, the trained models were capable of cleaning the data and processing it without needing external code. For the training and evaluation of the first model (developed with the Transformers library), we compared it to the ML models presented by HateBR, which had the best results, using the same feature extraction method they did: TF-IDF. TF-IDF (Term Frequency–Inverse Document Frequency) calculates how relevant a word in a corpus is to a text, obtained by the ratio between the number of times the term in question appears in one of the corpus texts (Term Frequency) and the frequency of appearances of this same term in the entire corpus (Inverse Document Frequency). The frequency of the word in a text refers to the ratio between the number of times it appears in the text and the number of words in the text, while the frequency of the word in the corpus is the count of how many times it appears in all texts of the dataset. The reason for using such feature extraction is due to the empirical results demonstrated by HateBR [Vargas et al. 2022], which show that, in general, models trained with features extracted using TF-IDF outperformed other methods.

4.3. Model Training

The model training process was split into two parts: the first was responsible for detecting offensiveness in a text (in this case, in a selected paragraph), and the second was responsible for identifying words or expressions that contain offensive language in segments classified as offensive by the first model.

We employed the BERTimbau model from the HuggingFace platform for the first model. BERTimbau is a Brazilian language model developed by NeuralMind [Souza et al. 2020] through a technique called fine-tuning. In the case of this study, this process involved adding an extra layer of neurons at the end of the model, which was responsible for classifying an input text as offensive or non-offensive.

For the second model, the training process used by the OLID-BR study [Trajano et al. 2023] was utilized to detect offensive spans (i.e., sequences of characters containing offensive language, not limited to isolated words but also including expressions and punctuation).

4.4. Model Demonstration

We evaluate the offensive language detection models using Precision, Recall, and F1-Score metrics. Precision measures how much we can trust a model when it predicts that an example belongs to a particular class by calculating the number of examples the model correctly predicted as belonging to that class divided by the total number of examples it predicted as belonging to that class. Recall is the number of samples the model correctly identified as belonging to a class divided by the total number of samples that belong to that class in the data. F1-Score is the harmonic mean between precision and recall, i.e., it is the average of both precision and recall values, giving more importance to low values, as a much lower precision or recall value indicates that the model is not balancing these two metrics well when we want to give equal importance to both.

To demonstrate the models' effectiveness, we conducted qualitative tests on selected texts, as long as they were at least one page long or had more than one paragraph.

For the application process of the trained models, Figure 1 presents the following steps graphically: (1) a large text is collected; (2) the text is divided into fragments based on the characters of periods, commas, semicolons, exclamations, questions, and new lines; (3) for each fragment, the BERTimbau-based model is used to check if it is offensive; (4) if not offensive, the fragment is returned as usual, but if it is, it proceeds to the next step; (5) the spaCy-trained model is used to identify the offensive spans; (6) returning of the offensive censored spans, and the censored version of the fragment; (7) finally, all fragments, censored or not, are reassembled using the same separators as before, thus returning the entire text now with the appropriate censorship.

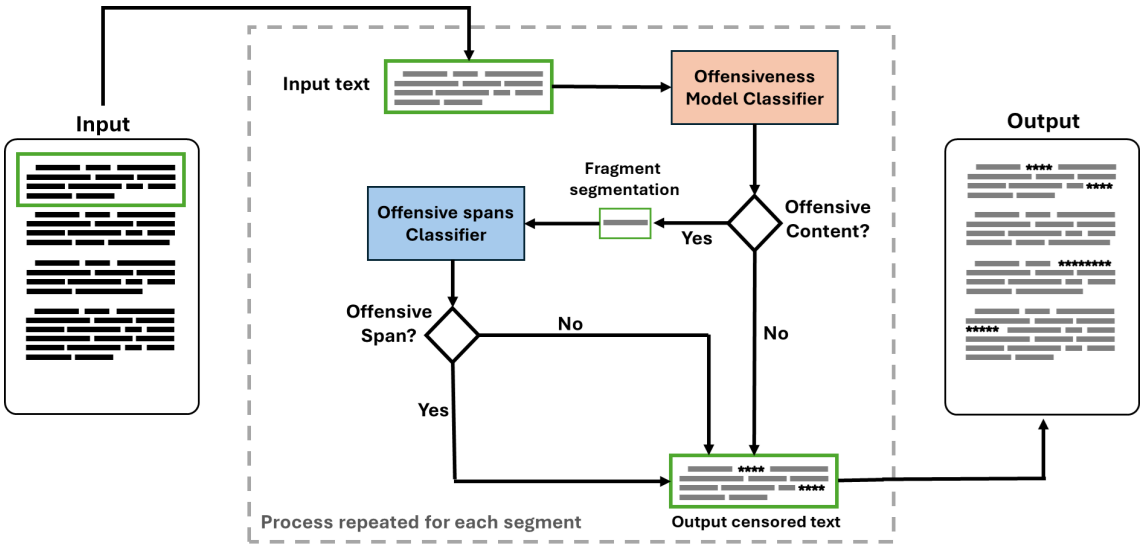

Figure 1. Censorship process of our proposal.

5. Results

The results obtained from the training sessions, in terms of Precision, Recall, and F1-score evaluated across the four different models tested for toxicity detection in sentences, are presented in Table 2. The parameters for conducting the training and testing followed the proposal in the HateBR study [Vargas et al. 2022]. Since HateBR does not provide the original pre-trained models and also the train/validation/test split configuration, it was necessary to retrain each of the four models on this dataset to guarantee consistency between the results: SVM, Naive Bayes, Logistic Regression, and MLP models. We also evaluated our fine-tuned version of the BERTimbau model [Souza et al. 2020], following the TF-IDF method for feature extraction. For training the BERTimbau-based model, we utilized the tokenizer that comes pre-built with the model to perform feature extraction, as this is how BERT-based models operate.

Model	Precision	Recall	F1-score
SVM	0.87	0.84	0.86
NaiveBayes	0.85	**0.87**	0.86
LogisticRegression	0.86	0.85	0.85
MLP	0.86	0.82	0.84
BERTimbau + *fine-tuning*	**0.92**	**0.87**	**0.90**

Table 2. Comparison of trained models on the HateBR dataset [Vargas et al. 2022].

The data analysis shows that the BERTimbau + *fine-tuning* model employed in this study outperforms previous results in all comparison parameters, establishing this work as the state-of-the-art for the HateBR dataset. This result underscores the importance of using DL models in the NLP context, as they can yield significant benefits in text recognition and classification processes.

Having completed the test of the BERTimbau model and trained the spaCy model following the same methodology as OLID-BR [Trajano et al. 2023], we conducted the final evaluation of the complete pipeline. For this purpose, we executed the pipeline proposed in Figure 1, i.e., if a sentence is toxic, the offensive parts are detected and censored. For presentation and qualitative evaluation purposes, censorship is performed by simply replacing the characters that constitute the offensive word or expression with asterisks (*).

We selected some news articles from specialized portals, such as G1, Estadão, and Catraca Livre, to perform a preliminary test, but no offensive language was detected. Subsequently, an opinion article about Chico Buarque from the Senso Incomum blog [Martins 2022] was used, where the presence of swear words, which would be appropriately detected, was evident.

Table 3 presents examples for qualitative analysis of the results of applying the proposed pipeline in this work. The Table shows the comparison between the original text and the censored text. The demonstrated text is the only post from a collected blog, as the three obtained news articles contained no offensive language. These results are limited to demonstrating the parts of the original text that our algorithm censored.

One consideration is that the model censored the word *"censura"* (in english censorship), which is usually not considered offensive, as well as the expression *"Rock das*

Original Text Part	Censored Part
observador de **bonobos**	observador de *******
A auto**censura** é o pior tipo de **censura** que existe	A auto******* é o pior tipo de ******* que existe
Joga pedra na Geni/Joga **bosta** na Geni/Ela é feita pra **apanhar/Ela é boa de cuspir/**Ela dá pra qualquer um/Maldita Geni!	Joga pedra na Geni/Joga ***** na Geni/Ela é feita pra **********Ela dá pra qualquer um/Maldita Geni
seu lesbofóbico **"Rock das Aranhas"!**	seu lesbofóbico **********
Paga-pau dos porcos estadunidenses, com toda certeza!	**********, com toda certeza!
Espumando de **ódio**	Espumando de ****

Table 3. Examples of original text excerpts (on the left) and censored excerpts (on the right).

Aranhas" (Rock of the Spiders) instead of the previous word *"lesbofóbico"* (lesbophobic), which we judged to be far more offensive than an expression about rock and spiders. The remaining censorships we considered appropriate.

5.1. Source Code of the Experiments

The source code developed has been made publicly available in the form of Notebooks, available on the GitHub Repository: `https://github.com/ICDI/censorship-offensive-language`:

- The notebook for training and testing the model based on the BERTimbau model;
- The notebook for training the spam detection model with spaCy, reusing the code from OLID-BR
- The notebook with the tests using both models for text detection and censorship.

6. Conclusion

This work developed an ML/DL-based program for detecting and censoring offensive language in extended Portuguese texts extracted from the web. To this end, we present a pipeline comprising two parts: one that detects the offensiveness in a portion of text (precisely a sentence) and another that detects the offensive parts (spam) within the sentence. This allows for censoring a large text part by part and returning the censored text.

A positive aspect of this work is providing a program specifically focused on detecting and censoring large texts, presenting satisfactory results in quantitative and qualitative analyses. As an evolution of this work, developing a specific dataset for large text censorship can be envisaged, which, unlike the datasets used [Vargas et al. 2022] and [Trajano et al. 2023], would represent a significant advancement.

The results for detecting offensiveness in texts were superior to our original reference (the HateBR article), representing a new benchmark concerning the state-of-the-art, surpassing the techniques used as a reference in the HateBR dataset [Vargas et al. 2022]. The censorship part was performed simply by replacing the characters in the offensive parts with asterisks, which can undoubtedly be improved by future work. The analysis could have been more robust due to the lack of a specific dataset for large texts, even when discussing qualitative data. This fact demonstrates the need for creating a specific dataset built from expert judgments on offensive language or by calculable metrics, which can evolve in future works.

References

Cook, S. (2024). Cyberbullying statistics and facts for 2024 | comparitech. https://www.comparitech.com/internet-providers/cyberbullying-statistics/. (Accessed on 10/10/2024).

Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Economist (2019). Increasing numbers of children have internet addiction – how worried should parents really be? https://inews.co.uk/news/long-reads/internet-addiction-children-increase-parents-guide-242434. (Accessed on 10/10/2024).

Hajibabaee, P., Malekzadeh, M., Ahmadi, M., Heidari, M., Esmaeilzadeh, A., Abdolazimi, R., and Jones, J. H. (2022). Offensive language detection on social media based on text classification. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0092–0098.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Husain, F. and Uzuner, O. (2021). A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Leite, J. A., Silva, D., Bontcheva, K., and Scarton, C. (2020). Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924.

Leray, W. (2023). Série de harry potter? envolvimento de j.k. rowling divide fãs. https://catracalivre.com.br/entretenimento/nova-serie-de-harry-potter-polemica-envolvendo-j-k-rowling-divide-fas/. (Accessed on 09/02/2024).

Martins, T. (2022). Chico buarque dá comida aos censores - senso incomum. https://sensoincomum.org/2022/01/28/chico-buarque-da-comida-aos-censores/. (Accessed on 09/02/2024).

Monteiro, E. (2023). Caso bruno e dom: justiça decide levar amarildo e outros dois réus a júri popular | amazonas | g1. https://g1.globo.com/am/amazonas/noticia/2023/10/03/caso-bruno-e-dom-justica-decide-levar-amarildo-e-outros-dois-reus-a-juri-popular.ghtml. (Accessed on 09/02/2024).

Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Trajano, D., Bordini, R. H., and Vieira, R. (2023). Olid-br: offensive language identification dataset for brazilian portuguese. *Language Resources and Evaluation*, pages 1–27.

Trielli, L. (2021). Escócia: estupradores que se declararem mulher serão colocados em prisões femininas - senso incomum. `https://sensoincomum.org/2021/12/14/escocia-estupradores-que-se-declararem-mulher-serao-colocados-em-prisoes-femininas/`. (Accessed on 09/02/2024).

Vargas, F., Carvalho, I., Rodrigues de Góes, F., Pardo, T., and Benevenuto, F. (2022). HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183, Marseille, France. European Language Resources Association.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.