Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Improving hate speech detection using Cross-Lingual Learning

Anderson Almeida Firmino [a,*], Cláudio de Souza Baptista [a], Anselmo Cardoso de Paiva [b]

[a] *Federal University of Campina Grande, Rua Aprigio Veloso, 882 - Universitário, Campina Grande, Paraiba, Brazil*
[b] *Federal University of Maranhão, Av. dos Portugueses, 1966 - Vila Bacanga, São Luís, Maranhão, Brazil*

ARTICLE INFO

Keywords:
Hate speech detection
Natural language processing
Social media
Cross-Lingual Learning
Deep learning

ABSTRACT

The growth of social media worldwide has brought social benefits and challenges. One problem we highlight is the proliferation of hate speech on social media. We propose a novel method for detecting hate speech in texts using Cross-Lingual Learning. Our approach uses transfer learning from Pre-Trained Language Models (PTLM) with large corpora available to solve problems in languages with fewer resources for the specific task. The proposed methodology comprises four stages: corpora acquisition, the PTLM definition, training strategies, and evaluation. We carried out experiments using Pre-Trained Language Models in English, Italian, and Portuguese (BERT and XLM-R) to verify which best suited the proposed method. We used corpora in English (WH) and Italian (Evalita 2018) as the source language and the OffComBr-2 corpus in Portuguese (the target language). The results of the experiments showed that the proposed methodology is promising: for the OffComBr-2 corpus, the best state-of-the-art result was obtained (F1-measure = 92%).

1. Introduction

The growth of mobile technology and social media has brought significant changes regarding news and information reading habits. According to a recent survey (Shearer & Mitchell, 2021), people prefer to use their smartphone applications and social media to get news instead of printed newspapers and television. In particular, people prefer platforms such as Facebook and Twitter to gather information and news.

A report available at datareportal.com surveyed countries around the world and concluded that the most common use of mobile devices has been for social activities (Kemp, 2021). Indeed, increasingly people prefer to express their opinions on social media (Mathew et al., 2019). The same research also showed that most interviewed people often used to share videos and photos on their social networks. In addition, more than half of mobile phone and social media users in the countries surveyed have already posted their opinions on important issues.

Despite the benefits that social media provides to society, potential anonymity can be harmful. Some people behave more aggressively when using their social networks (Fortuna & Nunes, 2018). In the last decade, a growing proliferation of hate speech on the Internet occurred. Consequently, the European Union Commission has conducted and financed several initiatives to tackle this behavior. In 2016, there was pressure on several social media platforms and organizations, such as Facebook, Microsoft, Twitter, and YouTube, to sign a Hate Speech Code, reviewing valid notifications for removing hate speech within 24 h.

Fortuna and Nunes (2018) define hate speech as a language that attacks and incites violence against people groups based on their specific characteristics, such as physical appearance, religion, nationality, or ethnic origin, and gender (Nobata et al. (2016), Bourgonje et al. (2017)).

Hate speech detection on the Internet has become an arduous task nowadays. According to Bloomberg (Wagner, 2020), Facebook removed 22.5 million hate speech posts between April and June 2020, twice the amount withdrawn in the first quarter of the same year. In addition, Facebook announced that its terms were updated to ban what they coined as “implicit hate speech” (Wagner, 2020).

Hate speech detection on the Internet is highly relevant to society since many studies correlate hate speech with crimes (Burnap & Williams, 2016; Mondal et al., 2018; Schmidt & Wiegand, 2017). An increased number of hateful messages in a short time can manifest suspicious behavior in a community. Identifying and monitoring users that spread hateful comments may prevent these kinds of attacks against society. This information can circumvent incidents such as racial violence, terrorist attacks, or other crimes before they happen, thus providing steps toward anticipatory governance (Schmidt & Wiegand, 2017).

The United Nations Rabat Plan of Action (U.N. Human Rights Council, 2013) out instructions for differentiating between freedom of

[*] Corresponding author.
E-mail addresses: andersonalmeida@copin.ufcg.edu.br (A.A. Firmino), baptista@computacao.ufcg.edu.br (C. de Souza Baptista), paiva@nca.ufma.br (A.C. de Paiva).

https://doi.org/10.1016/j.eswa.2023.121115
Received 31 December 2021; Received in revised form 1 July 2023; Accepted 2 August 2023
Available online 10 August 2023

expression and hate speech. This plan also recommends discriminating between three types of expression: “an expression that makes up a crime; an expression that is not punishable but can justify civil action or administrative sanctions; and an expression that does not give rise to criminal, civil or administrative sanctions, but that causes apprehension in terms of tolerance, civility, and esteem for the rights of others”.

Among all social networks, Twitter is the one that contains the most messages with hateful comments (Fortuna & Nunes, 2018; Hewitt et al., 2016). Since Twitter relies on users to report hate tweets, removing hate speech tweets is a complex task (Hewitt et al., 2016). Therefore, the European Union Commission accused Twitter of not dealing with removing hate speech on its platform (Kottasová, 2017).

Hate speech detection involves several challenges, among which we can highlight: the low agreement on the labeling of hate speech by humans, showing that this classification is even more difficult for computational models; the difficulty in identifying all insults against groups because of social phenomena and the evolution of language, which grows at high speed, especially among young people who access social networks frequently. Thus, the hate speech detection task requires knowledge of the local culture and the community’s social organization (Fortuna & Nunes, 2018; Poletto et al., 2021; Schmidt & Wiegand, 2017).

Annotating hate speech text is a time-consuming task. There are far more positive or neutral texts than hateful texts in any random data sample. Despite the offensive nature of hate speech, abusive language can be very fluent and correct, and figures of speech – such as sarcasm – can be used. Another point worth highlighting is biased data sets. Some studies have already identified incorrectly labeled data because of the use of terms belonging to a specific dialect (Sap et al. 2019). Also, the predominance of the English language is explored in related work. Several non-English-spoken languages have limited contributions regarding hate speech detection.

Therefore, the ability to build intelligent systems for these languages is restricted. The importance of Natural Language Processing (NLP) tasks for low-resource languages has emerged recently during several crises in regions of the world where people speak languages that are not commonly addressed in the NLP community, such as the Ebola outbreaks in West Africa, such as Niger-Congo languages. As discussed by Pikuliak et al. (2021), most languages do not have enough data to create suitable models.

According to Pikuliak et al. (2021), English is the most researched language and is the only language considered in over 60% of published articles (Pikuliak et al., 2021). Lack of data is a pertinent problem for hate speech detection, and there are proposals for exploiting resources from English to improve in languages with fewer hate speech data (Stappen et al. (2020), Bigoulaeva et al. (2021)). A potential solution to this problem is using cross-lingual learning — CLL. Thus, we can reuse learning models in languages with few resources.

As discussed above, most work on hate speech detection uses English corpora (Mozafari et al., 2022). We use corpora in other languages, mainly Portuguese, as it is our mother tongue and because few papers use it. This article intends to contribute to the hate speech detection task in the Portuguese language, still proposing a novel method for other languages with few annotated corpora for this task. Furthermore, we aim to investigate whether Cross-Lingual Learning may achieve better results for the hate speech detection task. We also want to explore the impact of using different languages as target languages (e.g., English and Italian).

Data available in Portuguese for the hate speech detection problem is scarce. For example, the site www.lingua *teca.pt* lists resources in Portuguese, and we can see only a few annotated corpora available in this language. On the other hand, many hate speech corpora are available in English, with more than 50 corpora listed at hatespeechdata.com.

This article addresses the hate speech detection problem in texts using CLL. In particular, we use Italian, English, and Portuguese. Italian and Portuguese originate from the same mother tongue — Latin; whereas English has an Anglo-Saxon origin. The Portuguese corpus was published by de Pelle and Moreira (2017), the Italian corpus comes from the Evalita 2018 in the Hate Speech Detection (HaSpeeDe) task (Bosco et al. 2018), and Waseem and Hovy published the English corpus (Waseem & Hovy, 2016). Furthermore, to the best of our knowledge, this research work is the first to perform a Cross-Lingual Learning task for hate speech detection using a Portuguese corpus.

The remainder of this article is structured as follows. In Section 2, we discuss related research on hate speech detection. Section 3 focuses on our proposed method for detecting hate speech using CLL. In Section 4, we present the corpora used, and in Section 5, we address the experimental setup and the experiments performed. The results obtained are detailed and discussed in Section 6. Concluding remarks and a summary of the proposed work are highlighted in Section 7.

2. Related work

This section presents related work on hate speech detection using CLL. Initially, we discuss articles on hate speech detection using Machine Learning. Then, papers that used Deep Learning are addressed. Finally, we highlight the works that used CLL in their approaches.

2.1. Hate speech detection using machine learning

Schmidt and Wiegand (2017) wrote the first survey on hate speech. This work presents a brief overview of 22 studies in the area of hate speech. It is important to note that the authors considered other types of offensive languages, such as cyberbullying, to be hate speech — unlike the approach by Fortuna and Nunes (2018), which differentiated hate speech from other types of offense.

Fortuna and Nunes (2018) published a survey on hate speech. They developed a definition of the concept, based on the code of conduct of the European Union Commission, on the terms and conditions of social networks such as Facebook and Twitter, and scientific articles in the area. The authors classified the researched articles according to the type of hate speech addressed (racism, sexism, etc.); regarding the characteristics used in the hate speech detection process (n-grams, tf-idf, etc.), and regarding the classifier used (SVM, neural networks, among others). In addition, the authors point out some challenges and opportunities for research in the area.

Waseem and Hovy (2016) presented a corpus of about 16,000 tweets, of which 3383 were labeled as sexist content and 1972 as racist content. Experiments were performed using logistic regression to classify whether the tweet’s content contained hate speech, and the NLP technique used was n-grams of characters ($n = 4$). The best result (F1-measure of 73.93%) was obtained by combining n-grams of characters with the gender of the user who posted the tweet.

Bourgonje et al. (2017) were the first to use hate speech data in more than one language (English and German). Three public text corpora were used: two from Twitter (one in English and one in German) and one from Wikipedia (in English). The work compared several techniques for detecting hate speech, such as Logistic Regression and Maximum Entropy. For all corpora, only bag of words was used, which the authors considered to be the same as unigrams. Five different classifiers were tested, among which Naive-Bayes and Logistic Regression stood out.

Frenda et al. (2019) dealt with hate speech against women. The authors sought to investigate analogies and differences between sexism and misogyny from a computational point of view. They claimed to be the first to detect both concepts using the same approach. Using data from IberEval 2018 and Evalita 2018 (both about misogyny) and Waseem and Hovy (2016) about sexism, the authors found that tweets that were labeled as sexist or misogynistic had more female pronouns than the tweets labeled as neutral. To classify the texts, the authors performed a pre-processing (removal of emoticons, URLs and symbols, and lemmatization), used n-grams of characters and words and tf-idf, and used SVM as a classifier.

2.2. Hate speech detection using deep learning

Badjatiya et al. (2017) addressed the detection of hate speech – specifically, racism and sexism – in micro texts. The authors compared several techniques, and the best performance (F1-measure of 93%) was obtained when using LSTM together with Gradient Boosting and bag of words. They used a corpus with 16,000 tweets, of which about 3000 were labeled as sexist and about 1900 as racist. According to the authors, this was the first work to use Deep Machine Learning to detect hate speech.

Del Vigna et al. (2017) addressed hate speech detection in social media texts. About 17,000 Facebook comments from 99 posts were collected. Five undergraduate students labeled the comments as either hate speech or not. The authors used two classifiers in the experiment: an SVM and an LSTM neural network. They used lexicons to identify the polarity of feelings in the texts. Del Vigna et al. used features such as the number of tokens, character and word *n*-grams, and lemmas in the SVM classifier. In the LSTM classifier, they used the polarity of each term and lexicon. The authors performed experiments using two approaches: a binary and a multi-class with levels of hate speech in texts. The best results were obtained with the binary classification, with 80% of F1-measure for the SVM and 79% for the LSTM.

Silva et al. (2019) developed a novel approach to detecting hate speech in Portuguese, which comprises a model that uses CNN and a psycholinguistic dictionary – the Linguistic Inquiry and Word Count (LIWC) – with Logistic Regression (LR + LIWC). They used three Portuguese corpora: Off ComBr-2 and Off ComBr-3 (de Pelle & Moreira, 2017) and HSD (Fortuna et al., 2019) and compared their results with the baseline corpora. Silva et al. obtained the best results using a CNN and a 300-dimensional word embedding.

Fortuna et al. (2021) created an approach for detecting hate speech using different corpora in English. They standardized the classes present in the corpora to be able to use the other corpora interchangeably. They used ALBERT and BERT PTLMs, and a classifier based on fastText and SVM. The results showed how a cross-corpora approach could be used to detect similarities between corpora and categories and help to create a uniform categorization of corpora.

Plaza del Arco et al. (2021) used corpora in Spanish. They compared several models regarding hate speech detection. The corpora used include texts extracted from Twitter, one containing 6000 tweets obtained by HaterNet — an intelligent system used by the Spanish National Office against Hate Crimes of the Secretary of State for Security in Spain. The other corpus contained 1600 tweets obtained from the HatEval 2019 challenge. The authors concluded that Deep Learning models achieved better results than traditional machine learning. In addition, models that use Transfer Learning had an excellent performance.

Karim et al. (2021) developed an approach to detect hate speech in Bengali using PTLMs. They used several techniques, from Naive Bayes and Logistic Regression to CNN and PTLMs. The authors obtained the best results with a combination of PTLMs (BERT in Bengali, XLM-R, and multilingual BERT). In addition to performing binary classification, the approach developed by the authors was also able to perform multi-class classification, achieving the best results in the state-of-the-art for the Bengali language.

Soto et al. (2022) performed experiments using different embeddings with different dimensions and a CNN network to classify the texts. The authors used specific embeddings generated for the HSD and embeddings obtained from NILC (Hartmann et al., 2017). The representations of embeddings wang2vec, word2vec, fastText, and Glove were tested. The best result for the HSD corpus was achieved using the Glove representation with 300 dimensions, with the embeddings NILC.

2.3. Hate speech detection using cross-lingual learning

Corazza et al. (2020) created an approach for detecting language-independent hate speech. They argued that several works in the literature use domain-specific embeddings and other features that only apply to a particular corpus. They built a modular neural architecture that uses a hidden layer of 100 neurons. The best configuration found by the authors for the English corpus was the use of LSTM with character embeddings. For Italian, the best results were obtained when using LSTM combined with character embeddings, unigrams, and emoji transcriptions. The best result for the German language was obtained when using character embeddings as features and a GRU network. They used word embeddings, emoji embeddings, *n*-grams, emotion lexica, and social network features (number of hashtags, mentions, and others).

Pamungkas and Patti (2019) developed a multilingual hate speech detection approach in which the concept of transfer learning from a language richer in resources to another language with fewer resources is used. They performed experiments with English, Spanish, Italian, and German languages and obtained the best results using Hurtlex (Bassignana et al., 2018) and multilingual embeddings as features and an LSTM architecture. The best approach used Joint Learning and Multilingual embedding.

Ranasinghe and Zampieri (2021) used multilingual word embeddings to detect hate speech. Besides performing experiments with different languages, the authors also tested distinct domains. They collected data in English, Spanish, Hindi, and Bengali and used the XLM-R framework to perform the classification. The idea of using a multilingual approach is to train the model in a richer language (with more data) and test it in another language with fewer resources (i.e., fewer data). The authors trained the model in English and experimented with it in the other three languages researched. The results surpassed the state-of-the-art of each corpus and language.

Stappen et al. (2020) developed an approach to detect multilingual hate speech, adding some target language samples to the training cycle. The proposed architecture uses FastText-generated embeddings, a feature extractor (BERT or XLM), and an approach called Attention-Maximum-Average Pooling (AXEL) to perform the classification. This approach improves the features generated by XLM, using the maximum and average pooling of the XLM output.

Pamungkas et al. (2021) developed an approach to hate speech detection using CLL. They used English as the source language and six languages as the target: Portuguese, French, Spanish, German, Indonesian, and Italian. The authors experimented with different models: models that use traditional machine learning, such as logistic regression, and deep learning models, such as BERT. They used zero-shot transfer and joint learning in the experiments. The model that achieved the best performance was an LSTM neural network using multilingual embeddings provided by Facebook (MUSE - (Lample et al., 2018)).

Nozza (2021) argued on the problems of using only zero-shot approaches in the hate speech detection task using CLL. According to the author, a zero-shot approach cannot deal with expressions intrinsic to a language. Also, she argues it cannot transfer knowledge about hate speech when different domains are used. Although the author claims that detecting hate speech is language-specific, other works claim the opposite (Firmino et al., 2021; Pamungkas et al., 2021; Stappen et al., 2020). In addition, Nozza used corpora from Twitter and Facebook, which have different text representation patterns. She should have specified how much data from each platform she used in training and testing cycles, which may have needed to be more balanced and lead to consistent results.

Bigoulaeva et al. (2021) also performed hate speech detection using CLL. They used German as the source language and English as the target language. Several models of deep neural networks were tested using the zero-shot and joint learning strategies. The authors thus developed a transfer learning approach using CLL based on embeddings of bilingual words (BWEs).

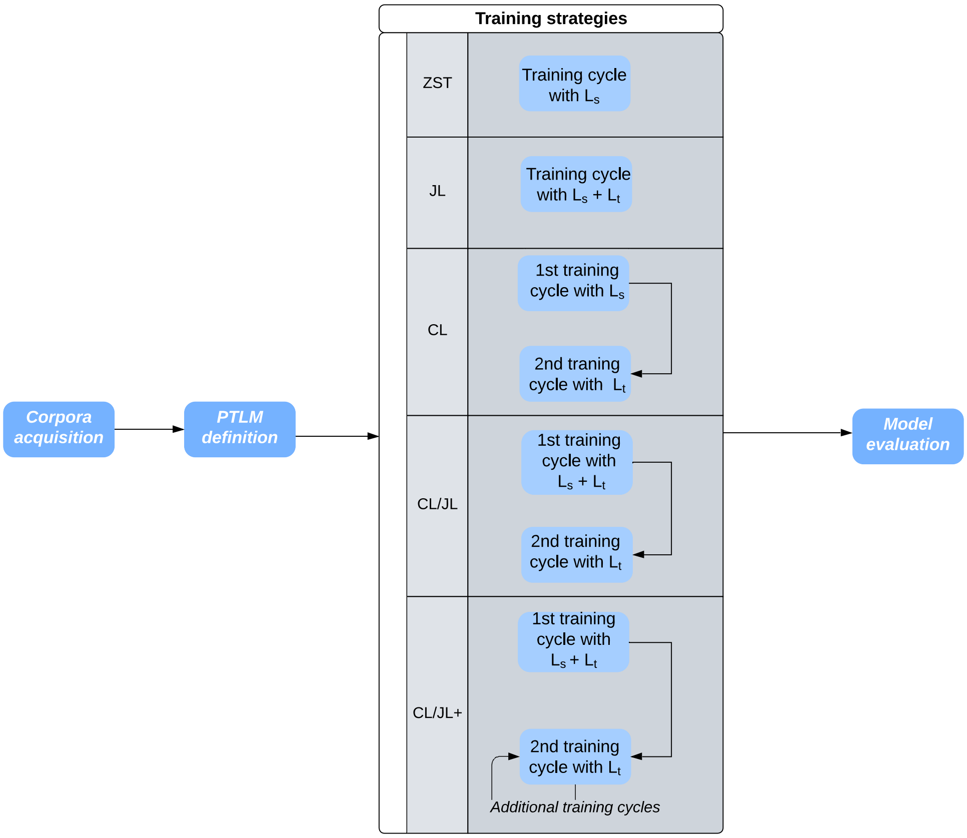
Corpora acquisition
PTLM definition
Training strategies
ZST
Training cycle with L_s
JL
Training cycle with $L_s + L_t$
CL
1st training cycle With L_s
2nd traning cycle with L_t
CL/JL
1st training cycle with $L_s + L_t$
2nd training cycle with L_t
CL/JL+
1st training cycle with $L_s + L_t$
2nd training cycle with L_t
Additional training cycles
Model evaluation

Fig. 1. Overview of the proposed method.

Among the works presented, we did not find any that sought to investigate the impact of using Cross-Lingual Learning. We also did not find articles with ablation studies with CLL — showing its use improves performance in any approach. Moreover, we did not find any papers investigating the impact of using languages from different families as target languages (e.g., when using German as the target language, would using English as the source language have more impact than using Dutch? Or would the result be similar?). In this sense, this research aims to bring these investigations and fill these gaps in the state-of-the-art.

3. Method

This work proposes the idea of applying CLL to detect hate speech. Fig. 1 presents an overview of the methodology, which includes four steps. The first step is corpora acquisition since this work uses an approach that requires more than one language. Therefore, using at least two corpora in different languages is required.

In the second step, we defined the Pre-Trained Language Model(s) (PTLM). The next step is to determine the training strategy. Based on the work of Pikuliak et al. (2021), we use five training strategies to detect hate speech in Portuguese — our target language (L_t). Thus, we perform the model induction with the information obtained through the data annotated in a source language L_s. Finally, the last step is the model evaluation.

For all approaches, we evaluate the final model using data from L_t. We used precision, recall, and F1-measure as evaluation metrics. In the next step, we detail each step of the methodology.

3.1. Corpora acquisition

Initially, it is necessary to build or to obtain corpora. We emphasize that when using CLL, at least two corpora are required: one in the source language and another in the target language. Two widespread methods to build corpora are web crawlers to collect texts on web pages (de Pelle & Moreira, 2017) or web APIs (del Arco et al., 2021; Waseem & Hovy, 2016). In this case, we must define the scope of the texts to be collected and choose which keywords will be used in the search or which web pages will be accessed. For example, de Pelle and Moreira (2017) built a corpus by obtaining comments from news pages about politics and sports. Waseem and Hovy (2016) produced a corpus with tweets about the Australian television show My Kitchen Rules. For this purpose, the authors obtained tweets containing the hashtag #MKR (indicating the TV show initials).

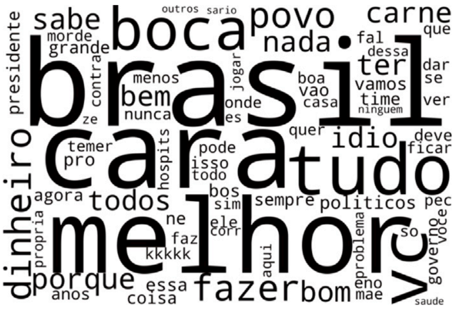

Fig. 2. Corpus OffComBr-2: Word cloud.

We can also use corpora already built and publicly available (Chung et al., 2019; Davidson et al., 2017; Frenda et al., 2019; Waseem & Hovy, 2016). There is also the possibility of using corpora available from conferences, events, or workshops, such as Evalita 2020 (Basile et al., 2020), and Evalita 2018 (Bosco et al., 2018), among others. In this work, we chose to use publicly available corpora. We used corpora described by Waseem and Hovy (2016), de Pelle and Moreira (2017), and Bosco et al. (2018).

After obtaining the corpora, the next step is to pre-process the texts. In approaches that use Deep Learning, the most common pre-processing comprises computing the vector representation of texts. This can be done using approaches such as Word2Vec (Mikolov et al., 2013), FastText (Grave et al., 2018), ELMO (Peters et al., 2018), among others. It is important to emphasize that when a Pre-Trained Language Model is used, in general, the model contains a module responsible for the text vectorizations (Bhaskaran & Bhallamudi, 2019; van der Heijden et al., 2021) - this includes those used in our work.

3.2. Pre-trained language model (PTLM) definition

In this work, we use two different PTLMs: XLM and BERT. Concerning XLM, we use the XLM-RoBERTa distribution — called XLM-R (Conneau et al., 2021); and regarding BERT, we use three distributions: the Portuguese BERT (BERTimbau) (Souza et al., 2020) Italian BERT (Schweter, 2020), and the English BERT (Devlin et al., 2019).

Although BERT has been trained in over 100 languages, it has yet to be optimized for multilingual models — most vocabulary is not shared across languages, so shared knowledge is limited. To deal with this, Conneau and Lample (2019) proposed XLM, making two significant modifications in the BERT architecture. First, instead of using words or characters as input to the model, they used Byte Pair Encoding (BPE), which splits the information into the most common sub-words across languages, thus increasing the vocabulary shared across languages.

XLM-R (Conneau et al., 2021) showed excellent results in Cross-Lingual Learning tasks, reaching an accuracy of 23% greater than that of the PTLM BERT in low-resource languages. XLM-R was trained in 104 languages with 2.5 terabytes of data. XLM-R is also compatible with monolingual benchmarks while achieving the best results in cross-lingual benchmarks (Conneau et al., 2021).

3.3. Training strategies

The first training strategy is Zero-shot transfer (ZST). In this case, no data from the target language L_t is used in the PTLM first training cycle. The training cycle in this strategy is performed using only the source language L_s corpus. The target language L_t corpus is used only in the test step.

The second strategy is named Joint Learning (JL). Here, both L_t and L_s corpora are used simultaneously in the first training cycle of the PTLM. Thus, the data used in the training cycle consists of all the source language corpus L_s plus a part of the target language corpus L_t. Finally, we use the remaining subset of the L_s corpus not used in the training cycle to test the model.

Cascade Learning (CL) is the third training strategy and consists of using only the L_s corpus in the first training cycle. Then, we perform an additional training cycle in which only part of the L_t corpus is used. In the end, we use the remaining L_t corpus to test the model.

In the first training cycle of the fourth strategy (CL/JL), a subset of the L_t corpus is used, along with L_t corpus. After the first training cycle, the weights are saved and then loaded for a second training cycle. Finally, the remaining subset of the L_t corpus, not used in the first training cycle, is used in both the second training and the test cycles.

The last training strategy is CL/JL+, where we made additional training cycles using the L_t corpus. In this case, the flow is similar to CL/JL, where the L_s corpus and part of the L_t corpus are simultaneously used in the first training cycle. The remainder of the L_t corpus is divided as follows: a small subset of it is kept to test the model at the and, and the subset is used to perform additional training cycles.

3.4. Evaluation

The last step of the proposed methodology consists of evaluating the induced model. Evaluation metrics are used to monitor and measure the model performance during training and testing cycles. In the proposed methodology, the metrics used were Precision, Recall, and F1-measure (Zhang & Zhang, 2009). We emphasize that we used weighted F1-measure in all experiments in this work.

4. Corpora used

In this research, we used four different corpora. The first is composed of data in Portuguese (OffComBr-2 - (de Pelle & Moreira, 2017)), with comments containing hate speech gathered at the Brazilian news site g1.globo.com. The authors labeled the data with two classes, indicating whether or not there was the presence of hate speech. From the OffComBr-2 corpus, we generated the second corpus that contains text translated into English. The translation was performed using the Google Neural Translation Machine (Wu et al. 2016).

The third corpus comprises Facebook posts in Italian, made publicly available at the Evalita 2018 conference, in the Hate Speech Detection task (Bosco et al., 2018). This corpus contains two classes: hate speech and neutral.

The fourth corpus is also composed of tweets in English. Waseem and Hovy (2016) collected tweets over two months and annotated the data into three classes: sexism, racism, and neutral.

4.1. The OffComBr-2 corpus

De Pelle and Moreira (2017) divided the corpus into two parts: Off-ComBr-2, which contains 1250 comments identified as offensive or neutral by at least two judges; Off-ComBr-3, which comprises 1033 comments annotated (out of 1250) by at least three judges. Regarding the Off ComBr-2, at least two judges labeled 419 (out of 1250) comments as offensive, representing 32.5% of the total. In the Off ComBr-3, at least three judges noted 202 comments as offensive (out of 1033), corresponding to 19.5% of the cases. The judges also said whether the offense in the comment was about racism, sexism, homophobia, xenophobia, religious intolerance, or a curse. Table 1 shows a sample of comments from this corpus. The authors did not provide information on whether there was any pre-processing in the corpus, but we noticed they removed punctuation marks and accents from the texts.

Fig. 2 depicts a word cloud with the most frequent words in the corpus. The most frequent word is Brazil, followed by terms such as ‘best’, ‘people’, and ‘money’, among others. The explanation is that De Pelle and Moreira (2017) generated the corpus from comments on politics and sports.

4.2. The evalita 2018 corpus

This corpus was created at the Istituto di Informatica e Telematica, CNR, Pisa, in 2016 (Del Vigna et al. (2017)). It comprises around 17,000 Facebook comments drawn from 99 posts from selected pages. Five people wrote about 4000 comments and classified them as hateful or not. The authors did not provide information about performing pre-processing in the texts provided. We noticed that uppercase and lowercase letters were kept, as well as punctuation and accent marks.

Table 2 displays a sample of comments from the Italian corpus. We can see the most frequent words in the Italian corpus in Fig. 3. ‘Mateo’, ‘Salvini’, ‘vote’, and ‘Renzi’ are some of the most common terms in the corpus. Matteo Salvini was the head of the Northern League, while Matteo Renzi was the Democratic Party (PD) leader when the corpus was collected. The content of the corpus was mainly political.

Table 1
Sample of the OffComBr-2 corpus.

Original text (PT)	Translation (EN)	Hate speech
cuidado com a poupanca pessoal Lembram o que aconteceu na epoca do Collor ne	Beware of personal savings. Remember what happened at President Collor's time, don't you?	No
os cariocas tem o que merecem um pessoal que so sabem tomar banho de sol e pratica a violencia e nao deu outra de onde se tira e nao coloca um dia acaba	Residents of Rio de Janeiro have what they deserve. They only know how to sunbath and practice violence and there is no ill that lasts forever, nor any boon that never ends.	Yes
Porque nao corta os gastos dos politicos	Why not cutting politicians' spending?	No
Mais um pobre metido a besta ja ja fica sem dinheiro e vai sentir falta dos bons tempos	Another poor stupid almost running out of money and sooner will miss the good times	Yes
PEC DA VIDAAAA VIDA LIVRE DE MAMATA ESQUERDALHAAAAAA kkkkkkkkkk	LAW OF LIFEEE LIFE FREE FROM the STUPID LEFT LOL	No

Table 2
Sample of the Evalita corpus.

Original text (IT)	Translation (EN)	Hate speech
sono indigeste, fanno anche venire la colica intestinale.	They are indigestible, they also cause intestinal colic.	Yes
Ho cambiato canale...le 2 del pd e ncd mi fanno schifo	I changed the channel ... the 2 of the pd and ncd make me sick	Yes
Ma chi le ha votate cosa pensa di loro	But who voted for them, what do you think of them	No
Siamo arrivati a un punto di non ritorno. POVERI NOI!!!!	We have reached a point of no return. POOR US!!!!	No
La Malpezzi vuole farci credere che esiste Babbo Natale e la Castaldini appartiene a un partito fantasma	Malpezzi wants us to believe that Santa Claus exists and Castaldini belongs to a ghost party	No

4.3. The WH corpus

Waseem and Hovy (2016) presented a corpus (to be called WH from then on) with around 16,000 tweets, of which 3383 were labeled as having sexist content and 1972 as having racist content. We merged the sexist and racist tweets into one set in this work. Thus, 5355 tweets were considered hate speech, equivalent to 33% of the corpus. The authors initially searched for slurs and terms used to refer to religious, sexual, gender, and ethnic minorities. Waseem and Hovy identified frequent words referencing specific entities, such as "#MKR", the Australian TV show My Kitchen Rules hashtag, which often generates sexist tweets aimed at women. The authors annotated the tweets along with an external annotator.

Table 3 presents a sample of the WH corpus. The most frequent words in this corpus can be seen in Fig. 4. 'Kat' was the most cited term, followed by 'like', 'women', and 'sexist'. 'Kat' was a contestant on the MKR show in 2015, shown to have temper issues, and she was one of the biggest victims of sexist comments by viewers.

Table 4 summarizes the corpora. We noticed that the corpora have different sizes and that the proportion of classes differs for both. The Evalita corpus contains balanced data, while the OffComBr-2 corpus contains mainly texts without hate speech (67,5%), corresponding to real-world behavior. Furthermore, we see that the Evalita and the WH corpora were collected through social networks, while the OffComBr-2 corpus was collected through a news website. We noticed an intersection between the OffComBr-2 and Evalita corpora since both address political issues.

Table 3
Sample of the WH corpus.

Original text (EN)	Hate speech
@user Mosque that is still standing. Typical Muslim terrorist asshole.	Yes
Not sure if I'm speaking there this year or not, but they were the first conference at which I ever presented.	No
Wish these blondes were in that How To Get Away With Murder show....#MKR	Yes
@user that's why they really can't effect us. gaming industry is involved, sure. But it's a much bigger conversation.	No
#mkr deconstructed by girls that have deconstructed brains ! Nearly brought up my dinner when I saw that crap on the plate	Yes

5. Experimental setup

This section presents the experiments to evaluate the proposed method to detect hate speech in texts using CLL.

5.1. Experiments description

Five experiments were carried out with each training strategy to investigate and evaluate the research problems according to the proposed methodology. In all experiments, we used the experimental strategies mentioned in Section 3:

- ZST: Zero-Shot Transfer;
- JL: Joint Learning;
- CL: Cascade Learning;
- CL/JL;
- CL/JL+.

The first experiment, called Evalita/OffComBr-2, used four PTLMs and the aforementioned experimental strategies. This experiment used the Evalita corpus (in Italian) as L_s and the OffComBr-2 corpus (in Portuguese) as L_t. In the second experiment – baseline – we used BERTimbau and XLM-R as PTLMs, and the experimental strategies

Table 4
Summary of corpora used.

Corpus	Texts without hate speech	Texts with hate speech	Total	Source	Context
Evalita	1941 (48,5%)	2059 (51,5%)	4000	Facebook	Politics
OffComBr-2	831 (67,5%)	419 (32,5%)	1250	News website	Sports and politics
WH	5.038 (62%)	3.098 (38%)	8.136	Twitter	Sexism and racism

Fig. 3. Corpus Evalita: Word cloud.

Fig. 4. Corpus WH: Word cloud.

ZST and CL. We used the Evalita and OffComBr-2 corpora in this experiment.

We carried out the third experiment – data balancing – using XLM-R and BERTimbau PTLMs, and the corpora Evalita and OffComBr-2. The training strategies used in the third experiment were JL and CL.

The fourth experiment – WH/OffComBr-2 – used XLM-R and BERTimbau PTLMs. We used all training strategies; the WH corpus was the L_s, and the OffComBr-2 corpus was the L_t. The fifth and last experiment (WH/ OffComBr-2-EN) used the five training strategies and the BERT and the XLM-R PTLMs. We used the WH corpora and the OffComBr-2-EN corpus (originally in Portuguese and translated into English).

We adopted the same settings for all PTLMs used in this work. Our experiments used a learning rate of $1 * 10^{-5}$, and a number of epochs equal to 3 (Conneau et al., 2021; Devlin et al., 2019). We used AdamW as the optimizer with epsilon equal to $1 * 10^{-8}$. We used the binary cross entropy function as the loss function and softmax as the activation function.

We used Google Colab with an Nvidia Tesla K80 GPU in all experiments and the PyTorch library to carry out the experiments with the adopted PTLMs.

Since the chosen PTLMs accept only fixed-length inputs, we set a maximum size for the data. Then, we mapped the entries into a 128-dimensional vector. In cases where the entries were smaller than this value, a filling with 0’s was performed; in cases where the entries were larger, a truncation was performed to the appropriate size, ignoring elements with an index greater than 128. In the corpora used, only one record exceeded the limit of 128 words.

In all cases, when using the ZST strategy, 90% of the data from the L_s corpus were used in the training cycle of the PTLMs, and the remaining 10% were used for validation. In the test stage, only the L_t language corpus was used. When using the JL strategy, the L_s language corpus was used with the same distribution of the ZST approach (90% for training cycle and 10% for validation cycle). In addition, we used a subset of the L_t language corpus in training cycle and tested the model using only the remaining data from the L_t language corpus. The percentage of data from the L_t language corpus to be added to the training cycle of PTLMs ranged from 10 to 70% in the experiments performed. The best results were obtained when we added 30% of the L_t language corpus in training cycle.

When we used the CL strategy, only the L_s language corpus was used with the following division: 70% of the training cycle data, 10% for the validation step, and 20% for the test step. After that, we fine-tune the PTLMs. This time, only L_t language data was used. We followed the same pattern used in the first step, with 70% of the L_t language data being used for training cycle, 10% for validation, and 20% for testing.

When using the CL/JL strategy, we used 70% of the data from the corpus of the language L_s together with 30% of the data of the corpus of the language L_t. For validation, we used 20% of the data from the language L_s, plus 10% of the data from the language L_t, and the remaining data from the language L_s were used in the test. We used the remainder of the L_t language data to fine-tune the PTLMs. With the remaining data, the same division was used in the first step (70/20/10).

Finally, in the CL/JL+ strategy experiments, we adopted the same division of the previous strategy in the initial training cycle, in which we had 70% of the language data L_s added to 30% of the language data L_t for training cycle. The same applies to the validation and testing steps. For further training cycles, we divided the remaining L_t language corpus according to the number of training cycles performed. We used k-fold cross-validation (k=10) to ensure the proportionality of the classes over the iterations.

In all experiments, we adopted the number of epochs recommended by Devlin et al. (2019) - between 2 and 4. Some strategies (e.g., CL/JL+) performed many training cycles. In those cases, the total sum of epochs used reached 15. Thus, we performed experiments with all experimental strategies, increasing the number of epochs. However, we noticed that the increase in the models’ performance was insignificant, and the computational cost for using a more significant number of epochs became relatively high. For experiments using the JL strategy, for example, training cycle lasted more than 1 h with 15 epochs. Thus, we continued to use a few epochs in each training cycle.

According to Bosco et al. (2018), from the 4000 texts made available, 3000 were separated for the training step (75%) and 1000 for the testing step (25%). We adopted a similar percentage in our experiments using 70% of the corpus for the first training cycle.

We used a confidence level of 95% (alpha = 0.05) for all significance tests performed. In all cases, it is considered that the null hypothesis can be rejected if the minimum epsilon is less than 0.5.

The source code of the experiments is available at: https://github.com/Anderson-a-f/hatespeechcll.

5.2. Experiment 1: Evalita/OffComBr-2

The first experiment used the Evalita corpus (Bosco et al., 2018) – Italian – as the source language L_s and the OffComBr-2 corpus (de Pelle & Moreira, 2017) – Portuguese – as the target language L_t. For this experiment, the following PTLMs were used: the BERTimbau (Souza

Table 5
Experiment 1: Results of the strategy ZST.

PTLM	Precision	Recall	F1-measure
XLM-R	72.4%	71.1%	71.2%
BERTimbau	70.2%	61.1%	66.1%
Italian BERT	59.3%	62.2%	60.2%
BERT	60.5%	54.0%	58.2%

Table 6
Experiment 1: Results of the strategy JL.

PTLM	Precision	Recall	F1-measure
XLM-R	77.1%	76.0%	77.0%
BERTimbau	77.1%	75.1%	76.1%
Italian BERT	63.6%	58.2%	61.3%
BERT	65.2%	52.1%	56.1%

Table 7
Experiment 1: Results of the strategy CL.

PTLM	Precision	Recall	F1-measure
XLM-R	82.0%	79.1%	80.0%
BERTimbau	83.5%	84.3%	83.4%
Italian BERT	75.2%	74.0%	74.1%
BERT	72.2%	73.1%	72.1%

Table 8
Experiment 1: Results of the strategy CL/JL.

PTLM	Precision	Recall	F1-measure
XLM-R	83.7%	81.1%	82.4%
BERTimbau	84.3%	85.1%	84.1%
Italian BERT	79.2%	78.0%	78.1%
BERT	73.1%	75.0%	74.0%

Table 9
Experiment 1: Results of the strategy CL/JL+.

PTLM	Precision	Recall	F1-measure
XLM-R	92.1%	89.2%	90.1%
BERTimbau	92.1%	93.0%	92.0%
Italian BERT	83.1%	80.2%	81.1%
BERT	80.4%	81.0%	80.1%

et al., 2020), the Italian BERT (Schweter, 2020), the BERT (Devlin et al., 2019) and the XLM-R (Conneau et al., 2021).

Table 5 shows the ZST strategy results for the PTLMs BERTimbau, Italian BERT, BERT, and XLM-R. Considering there was a significant improvement when using the large model at the expense of the base model, the first one was chosen in each of the PTLMs.

Regarding the JL training strategy, Table 6 presents the results of using this strategy for the four previously mentioned PTLMs.

Table 7 shows the CL strategy results. Again, we used the same PTLMs. In the first training cycle, we used the Evalita corpus, and in the second training cycle, we used only the OffComBr-2 corpus.

Table 8 shows the results of using the CL/JL strategy for the four PTLMs. We used a subset of the OffComBr-2 corpus in the first training cycle of PTLMs and the Evalita corpus. In the second training cycle, only the remainder of the OffComBr-2 corpus was used.

Table 9 shows the results of using the CL/JL+ strategy for the four PTLMs with five fine adjustments.

The result presented by PTLM BERTimbau reached an F1-measure of 92%, achieving the best state-of-the-art result among the works that used the OffComBr2 corpus. Silva et al. (2019), Lima and Bianco (2019) and Pari et al. (2019) also used the OffComBr2 corpus. In Table 10, a comparison with the F1-measure of these works is presented.

Table 10
Experiment 1: State of the art comparison.

Work	F1-measure
Lima and Bianco (2019)	72%
Baseline — de Pelle and Moreira (2017)	77%
Pari et al. (2019)	86%
Silva et al. (2019)	89%
Our approach	92%

Table 11
Experiment 2: Baseline results for the ZST strategy.

PTLM	Using CLL	Precision	Recall	F1-measure
XLM-R	Yes	72.4%	71.1%	71.2%
	No	50.3%	44.2%	48.2%
BERTimbau	Yes	70.2%	61.1%	66.1%
	No	51.1%	50.1%	50.1%

Table 12
Experiment 2: Baseline results for the CL strategy.

PTLM	Using CLL	Precision	Recall	F1-measure
XLM-R	Yes	82.0%	79.1%	80.0%
	No	46.2%	68.1%	55.1%
BERTimbau	Yes	83.5%	84.3%	83.4%
	No	62.1%	59.2%	60.1%

5.3. Experiment 2: Monolingual baseline

For the second experiment, we used BERTimbau and XLM-R as PTLMs. Only the OffComBr-2 corpus was used in this experiment. One of the objectives of this work is to understand if the use of CLL brings any benefit to the proposed methodology. For this, we carried out experiments without using a L_t language. That is, we did not use an auxiliary corpus in the method. Thus, the OffComBr-2 corpus was divided according to the ZST and CL approaches.

In this monolingual experiment, we used the ZST and CL training strategies. The results of experiments using the ZST strategy can be seen in Table 11. The CL strategy used part of the data to perform a first training cycle on the PTLMs, and we saved the other amount for other training cycles and testing step. The results of this experiment can be seen in Table 12.

5.4. Experiment 3: Data balancing

This experiment verified whether data balancing would impact the performance of PTLMs in the proposed methodology. As depicted in Table 4, the corpus most used in this research (OffComBr-2) presents an imbalance between classes, with approximately 70% of the data belonging to the neutral class and the remaining 30% belonging to the hate speech class.

In this experiment, we used BERTimbau and XLM-R as PTLMs. We used JL and CL as training strategies. Again, the Evalita corpus was used as source language L_s, and the OffComBr-2 corpus was used as target language L_t.

We divided the experiment into two steps. The first consisted of maintaining the OffComBr-2 corpus in proportion to the Evalita corpus. For that, we used a subsampling technique to balance the data that removed instances of the majority class. Thus, the proportion between the classes of the OffComBr-2 corpus was equal to that between the classes of the Evalita corpus.

The second step of the experiment consisted of considering the data from the Evalita corpus in proportion to the OffComBr-2 corpus. Again, subsampling was used to remove instances from the positive class to have a balance equivalent to 70/30, in which the majority class is neutral (no hate speech).

Table 13 displays the results of the third experiment. The results of the first part of this experiment are listed at the lines where the data ratio is 50%/50%. In these cases, the data from the OffComBr-2 corpus follow the same proportion as the Evalita corpus.

The results of the second step of this experiment are represented at the lines where the proportion of the data is as 70%/30%. In this

Table 13
Results of data balancing experiments.

PTLM	Training strategy	Data proportion	Precision	Recall	F1-measure
XLM-R	JL	Original	77.1%	78.0%	77.0%
		50/50	75.2%	75.1%	75.1%
		70/30	74.3%	74.4%	74.3%
BERTimbau	JL	Original	77.1%	75.1%	76.1%
		50/50	77.1%	75.2%	76.1%
		70/30	76.2%	76.1%	76.1%
XLM-R	CL	Original	82.0%	79.1%	80.0%
		50/50	80.2%	79.3%	79.2%
		70/30	79.5%	76.4%	78.4%
BERTimbau	CL	Original	83.4%	84.1%	83.2%
		50/50	83.1%	82.2%	82.1%
		70/30	80.0%	81.2%	81.1%

Table 14
Experiment 4: Results of the strategy ZST.

PTLM	Precision	Recall	F1-measure
XLM-R	72.5%	69.3%	59.3%
BERTimbau	59.1%	66.2%	55.1%

Table 15
Experiment 4: Results of the strategy JL.

PTLM	Precision	Recall	F1-measure
XLM-R	75.2%	76.1%	74.1%
BERTimbau	70.5%	68.4%	68.5%

Table 16
Experiment 4: Results of the strategy CL.

PTLM	Precision	Recall	F1-measure
XLM-R	79.1%	80.2%	80.1%
BERTimbau	82.3%	81.2%	81.2%

Table 17
Experiment 4: Results of the strategy CL/JL.

PTLM	Precision	Recall	F1-measure
XLM-R	89.0%	89.1%	89.0%
BERTimbau	85.2%	85.1%	85.1%

Table 18
Experiment 4: Results of the strategy CL/JL+.

PTLM	Precision	Recall	F1-measure
XLM-R	92.2%	91.1%	91.0%
BERTimbau	92.7%	92.1%	92.3%

Table 19
Experiment 5: Results of the strategy ZST.

PTLM	Precision	Recall	F1-measure
XLM-R	46.1%	68.0%	55.0%
BERT	55.3%	68.1%	55.1%

Table 20
Experiment 5: Results of the strategy JL.

PTLM	Precision	Recall	F1-measure
XLM-R	74.3%	74.2%	74.2%
BERT	80.1%	75.1%	78.1%

Table 21
Experiment 5: Results of the strategy CL.

PTLM	Precision	Recall	F1-measure
XLM-R	84.6%	85.2%	84.1%
BERT	85.2%	85.3%	85.1%

Table 22
Experiment 5: Results of the strategy CL/JL.

PTLM	Precision	Recall	F1-measure
XLM-R	85.1%	86.3%	85.2%
BERT	86.2%	86.4%	86.2%

Table 23
Experiment 5: Results of the strategy CL/JL+.

PTLM	Precision	Recall	F1-measure
XLM-R	91.1%	90.2%	90.1%
BERT	91.0%	90.1%	90.1%

experiment, the Evalita corpus followed the same proportion as the OffComBr-2 corpus.

5.5. Experiment 4: WH/OffComBr-2

The fourth experiment used the WH corpus – English – as the source language L_s and the OffComBr-2 corpus as the target language L_t. Considering that this research also aimed to verify the impact of the target language on the result of hate speech detection, this experiment was based on using a language of linguistic root other than Latin to be the source language. For this experiment, we used the PTLMs BERTimbau and XLM-R.

Table 14 presents the ZST strategy's results for the PTLMs used in this experiment. Furthermore, in Table 15, the results of using JL can be seen.

The results of the CL experiments are listed in Table 16. We can see the results of the CL/JL experiments in Table 17.

In the fourth experiment, the last strategy tested was the CL/JL+. In these experiments, there are five additional training cycles in the PTLMs. Table 18 lists the results of these experiments.

5.6. Experiment 5: WH/OffComBr-2-EN

In the fifth experiment, we used the WH corpus as the auxiliary corpus and the OffComBr-2 corpus translated into English as the main corpus. In this experiment, the CLL was not used. The idea was not to have corpora of different languages in the same experiment but auxiliary corpora of the same language. These experiments can also serve as baseline studies since CLL was not used.

In this experiment, we used a translation machine to translate the OffComBr-2 corpus into English. Thus, we translated the Portuguese texts into English using Google's Neural Translation Machine (Wu et al., 2016). BERT and XLM-R were used for this experiment as PTLMs using all training strategies.

The results of the ZST strategy are listed in Table 19 for the two PTLMs used. When using the ZST strategy, the WH corpus was used to perform the training cycle, and the OffComBr-2(EN) corpus for the test stage. We can see the results of the JL strategy in Table 20. When using this strategy, a subset of the OffComBr-2(EN) corpus is added to the training cycle of PTLMs.

Table 21 lists the results of using CL. Only the WH corpus was used in the first training cycle of PTLMs. In the second training cycle, only the OffComBr-2 (EN) corpus was used. Table 22 presents the results of using CL/JL, in which a subset of the OffComBr-2 (EN) corpus was used in the first training cycle. On the other hand, in the second training cycle, the remainder of this corpus was used for training, validation, and testing steps.

Table 23 shows the CL/JL+ strategy results. In this strategy, a subset of the OffComBr-2 (EN) corpus was attached to the first training cycle of the PTLMs, as in the CL/JL experiments. In the second training cycle, the remainder of the OffComBr-2 (EN) corpus was divided according to the number of iterations.

6. Discussion of the results

In this section, we discuss the results obtained from the proposed model to detect hate speech in Portuguese using CLL.

6.1. Discussion

In the first experiment, four PTLMs were used: BERTimbau (Souza et al., 2020), Italian BERT (Schweter, 2020), BERT (Devlin et al., 2019), and XLM-R (Conneau et al., 2021). We used the XLM-R in all experiments because it was trained with a large multilingual corpus. We used BERTimbau since this model was trained using Portuguese. We used the Italian BERT because the source language data were in Italian during most of the experiments. BERT (English) was used because this model was the first PTLM to be made available, having achieved the best results in several Natural Language Processing tasks. On the other hand, the use of BERT and BERT Italian had only a purely investigative analysis.

For all cases in experiment 1, the Italian BERT and BERT performed poorly compared to the XLM-R and BERTimbau, since Portuguese was the most focused language in the first experiment. In all cases, the Portuguese corpus was used to evaluate the trained model; in some experiments, this corpus was also used in the training step. Considering that the two PTLMs mentioned were not initially trained with data in Portuguese, they presented an inferior result concerning the XLM-R and BERTimbau. The Italian BERT still showed a slightly better result than the BERT. Possibly, this is because the Italian language belongs to the same linguistic root as the Portuguese language.

Using the ZST strategy of experiment 1, we see that the PTLM XLM-R presented better performance (Table 5). This behavior was repeated in all other experiments. A possible explanation is that the XLM-R was initially trained with multilingual corpora. Thus, it manages to perform well in tasks that involve CLL.

Still, in experiment 1, in Table 6 (JL experiments), we see that the results were superior to those of the ZST strategy. We then noticed that using data from the target language L_t in the first training cycle of the PTLMs benefits the performance of the models used. From Table 7 (CL experiments), we observe that BERTimbau presented a superior result than XLM-R. In the CL experiments, the second training cycle of the PTLMs was performed solely with the OffComBr-2 corpus. In this situation, BERTimbau takes advantage of having been trained only with data in this language.

Table 8 presents the results of the CL/JL experiments. Again, we observe that adding L_t language data in the first training cycle of the model improves the performance of PTLMs.

De Pelle and Moreira (2017) provided a baseline for the OffComBr-2 corpus and used SVM to perform the classification. Pari et al. (2019) and Silva et al. (2019) used deep learning approaches. Both used convolutional neural networks; Pari et al. (2019) used word2vec with 100 dimensions together with CNN, while Silva et al. (2019) used wang2vec with 100 dimensions, obtaining the best result in state-of-the-art so far. It is worth noting that our proposed methodology is the only one that used CLL, achieving the best result among the works that used the OffComBr-2 corpus, as seen in Table 10. We performed significance tests for all strategies in the first experiment, and the results reinforce this.

Regarding the baseline studies performed (experiment 2), we observed that the use of CLL substantially improved the performance of the models used (Tables 11 and 12), and the significance tests support this result. When performing hate speech detection using only the OffComBr-2 corpus, the result was very unsatisfactory for both BERTimbau and XLM-R. Even the lowest result obtained in the experiments using CLL – 66% with BERTimbau and the ZST strategy – resulted in a superior result compared to not using CLL.

The third experiment was about balancing the data used. The OffComBr-2 corpus presents unbalanced data; then, experiments were carried out to keep them in the same proportion as the Evalita corpus and to maintain the data from the Evalita corpus in the same proportion as the OffComBr-2 corpus. The results showed that balancing did not provide a superior result; on the contrary, in some cases, there was a decrease in the evaluation metrics when balancing the corpora. There was a 0%–3% difference between the balance results and the original proportion of corpora.

The decrease in the performance of the models can be explained by the fact that the training set was reduced in the experiments. In the third experiment, the subsampling technique was used. This technique removes instances from the majority class (Table 13). The results of the significance tests for the PTLM XLMR-R demonstrated that the model performed better without balancing the data. As for the BERTimbau, there was no significant difference between the results and the balancing performed in the CL strategy.

The fourth experiment was based on the investigation of using a source language not descendant from Latin (in this case, English) to assist in detecting hate speech in Portuguese. The results of the experiments showed that the use of the English language contributed positively to the performance of the PTLMs. Once again, 92% of the F1-measure was obtained for the OffComBr-2 corpus. Regarding the PTLMs used (BERTimbau and XLM-R), we observed that the results of both were similar in almost all experiments. The difference in the CL/JL+ experiments was only 1% (Table 18).

We performed significance tests to determine whether there was a significant difference between the results of using Italian or English as the source language in the proposed approach. The results showed no significant difference; thus, both languages contribute positively as source languages, using Portuguese as the target language.

The fifth experiment addressed the WH corpus as an auxiliary corpus, and the OffComBr-2 corpus was translated into English as the main corpus. The PTLMs used were XLM-R and BERT. The results of the PTLMs were very similar, and the best result obtained here (90% with the CL/JL+ approach – Table 23) did not exceed the one obtained previously – using Italian as the source language. The results of the significance tests reaffirm the result obtained.

Regarding the computational cost in developing the proposed methodology, we used Google Colab with an Nvidia Tesla K80 GPU in the experiments. The computational cost becomes directly proportional to the number of fine adjustments performed. The experiments used about 4 GB of RAM and 40 GB of disk storage in most experimental approaches. For the CL/JL+ experiments, almost all resources available in Google Colab were used, both RAM and disk storage.

7. Conclusion

This research addressed the problem of detecting hate speech in texts. More specifically, we address the difficulty of developing solutions that detect hate speech in languages with little available data. To this end, the proposed approach uses CLL to detect hate speech in Portuguese. Corpora in Italian and English were used as source languages and corpora in Portuguese as the target language.

One of the objectives of this work was the development of a methodology to detect hate speech in Portuguese using CLL. The results showed that using CLL with Latin-based languages as source languages is promising when the target language is Portuguese. The performance of the PTLMs was improved using Italian as the source language. The same applies to Anglo-Saxon-based languages since using English as the source language also enhanced the performance of the PTLMs. The significance tests performed in both cases support these conclusions. We also emphasize that the methodology presented in this work contributed to the state-of-the-art hate speech detection in Portuguese.

Of the PTLMs used, we observed that BERT and Italian BERT could have performed better in the proposed methodology when using Portuguese as the target language. Thus, BERTimbau and XLM-R performed better than those. BERTimbau showed a better performance in all scenarios tested, with the support of the significance tests performed.

Table 24
Message of the Evalita corpus.

Original text (IT)	Translation (EN)	Hate speech
Ho cambiato canale...le 2 del pd e ncd mi fanno schifo	I changed the channel ... the 2 of the pd and ncd make me sick	Yes

We noticed that the CL/JL+ strategy performed best in all scenarios among the training strategies used. Furthermore, the baseline experiment results demonstrated that using CLL improved the performance of the PTLMs. With this strategy, we obtained the best performance for the OffComBr-2 corpus. The performed significance tests support the results of our method for the used corpora.

The results of the baseline experiments (experiment 2) showed that using CLL significantly improved the performance of the PTLMs used in the proposed methodology.

The proposed methodology obtained the best performance among the works that used the OffComBr-2 corpus. It is important to note that, in this work, the weighted F1-measure was used, while in most of the related work, the authors did not mention the average computed for the F1-measure.

One of the contributions of this work was investigating which languages would be helpful when used as source languages in the proposed method when using Portuguese as the target language. In experiment 1, using the experimental strategy CL/JL+, an F1-measure of 92% was obtained for PTLM BERTimbau, using Portuguese as the target language (corpus OffComBr-2). We then used Italian as the source language in experiments 1 and 2. The results of experiment 2 also demonstrated the contribution of using Italian as the source language.

Regarding the use of English as the source language (experiment 4), we obtained results that demonstrated that the performance of the proposed methodology was improved when using English as the source language. With PTLM BERTimbau and using the experimental strategy CL/JL+, an F1-measure of 92% was obtained — having the same result as using Italian as the source language and achieving the best result among the works used the corpus OffComBr-2.

7.1. Limitations

As described in the method, the corpora used in this research have different focuses. The two main corpora used were OffComBr-2 and Evalita. One contains texts from Facebook, while the other comprises comments extracted from news pages.

Some terms may be considered offensive in one language but not in another. In addition, some words are tied to a geopolitical context. An example is in a post in the Italian corpus, which the authors labeled as hate speech (Table 24). The targets of the offensive message in this text (PD and ncd) do not help detect hate speech in Portuguese since they are Italian political parties that do not exist in Brazil.

Another question in this example is that hate speech classification can be subjective. For the corpus annotators, the message contained hate speech. Thus, this post may not be considered hate speech for a given annotator since it does not contemplate any direct offense to a specific group of people.

As Bender et al. (2021) mentioned, language models have limitations. In addition to the computational and environmental cost of creating these models, it should be taken into account that there are inherent errors in them. Bender et al. (2021) report the biases in corpora used by PTLMs such as BERT and GPT-3. Thus, these limitations also apply to this work, considering that PTLMs are used in the proposed method.

7.2. Future work

As further work, we suggest using additional languages as a source in the proposed methodology. Other suggestions for future research are addressing other similar topics, such as offense detection. Hence, the proposed method could be validated in different domains. In addition, we suggest checking the developed method for other distinct domains, such as fake news detection.

Another future work is regarding cross-corpora exploration. Hence, corpora from different domains (i.e., not just hate speech) can be investigated. Finally, we suggest using more than one corpus as the source language and/or more than one corpus as the target language. In this way, it could be identified whether using more than one corpus of the same language or more than one corpus of different languages could benefit the proposed method.

CRediT authorship contribution statement

Anderson Almeida Firmino: Methodology, Software, Investigation, Writing – original draft, Visualization. **Cláudio de Souza Baptista:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration. **Anselmo Cardoso de Paiva:** Conceptualization, Validation, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors would like to thank the Brazilian National Council of Scientific and Technological Development (CNPq) and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 for partially funding this research.

References

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In R. Barrett, R. Cummings, E. Agichtein, & E. Gabrilovich (Eds.), *Proceedings of the 26th international conference on world wide web companion* (pp. 759–760). ACM, http://dx.doi.org/10.1145/3041021.3054223.

Basile, V., Croce, D., Maro, M. D., & Passaro, L. C. (2020). EVALITA 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for Italian. In V. Basile, D. Croce, M. D. Maro, & L. C. Passaro (Eds.), *CEUR workshop proceedings*: *vol. 2765*, *Proceedings of the seventh evaluation campaign of natural language processing and speech tools for italian. final workshop*. CEUR-WS.org, http://ceur-ws.org/Vol-2765/overview.pdf.

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In E. Cabrio, A. Mazzei, & F. Tamburini (Eds.), *CEUR workshop proceedings*: *vol. 2253*, *Proceedings of the fifth Italian conference on computational linguistics*. CEUR-WS.org, http://ceur-ws.org/Vol-2253/paper49.pdf.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In M. C. Elish, W. Isaac, & R. S. Zemel (Eds.), *FAccT '21: 2021 ACM Conference on fairness, accountability, and transparency, virtual event / Toronto* (pp. 610–623). ACM, http://dx.doi.org/10.1145/3442188.3445922.

Bhaskaran, J., & Bhallamudi, I. (2019). Good secretaries, bad truck drivers? Occupational gender stereotypes in sentiment analysis. In *Proceedings of the first workshop on gender bias in natural language processing* (pp. 62–68). Florence, Italy: Association for Computational Linguistics, https://aclanthology.org/W19-3809.

Bigoulaeva, I., Hangya, V., & Fraser, A. (2021). Cross-lingual transfer learning for hate speech detection. In B. R. Chakravarthi, J. P. McCrae, M. Zarrouk, R. K. Bali, & P. Buitelaar (Eds.), *Proceedings of the first workshop on language technology for equality, diversity and inclusion* (pp. 15–25). Association for Computational Linguistics, https://www.aclweb.org/anthology/2021.ltedi-1.3/.

Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018). Overview of the EVALITA 2018 hate speech detection task. In T. Caselli, N. Novielli, V. Patti, & P. Rosso (Eds.), *CEUR Workshop Proceedings*: *2263*, *Proceedings of the sixth evaluation campaign of natural language processing and speech tools for Italian. Final workshop (EVALITA 2018) co-located with the fifth Italian conference on computational linguistics*. CEUR-WS.org, http://ceur-ws.org/Vol-2263/paper010.pdf.

Bourgonje, P., Schneider, J. M., Srivastava, A., & Rehm, G. (2017). Automatic classification of abusive language and personal attacks in various forms of online communication. In G. Rehm, & T. Declerck (Eds.), *Lecture notes in computer science*: *vol. 10713*, *Language technologies for the challenges of the digital age - 27th International conference, GSCL 2017, Berlin, Germany, September 13-14, 2017, Proceedings* (pp. 180–191). Springer, http://dx.doi.org/10.1007/978-3-319-73706-5_15.

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. In *EPJ data sci.*: *vol. 5*, (no. 1), (p. 11). http://dx.doi.org/10.1140/epjds/s13688-016-0072-6.

Chung, Y., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019). CONAN - counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (pp. 2819–2829). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/p19-1271.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, & P. Motlícek (Eds.), *Interspeech 2021, 22nd annual conference of the international speech communication association* (pp. 2426–2430). ISCA, http://dx.doi.org/10.21437/Interspeech.2021-329.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019* (pp. 7057–7067). https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. *20*, In *ACM Transactions on Internet Technology* (2), (pp. 10:1–10:22). http://dx.doi.org/10.1145/3377323.

Davidson, T., Warmsley, D., Macy, M. W., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the eleventh international conference on web and social media* (pp. 512–515). AAAI Press, https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665.

de Pelle, R. P., & Moreira, V. P. (2017). Offensive comments in the Brazilian web: A dataset and baseline results. In *Proceedings of the VI Brazilian workshop on social network analysis and mining*. SBC, http://dx.doi.org/10.5753/brasnam.2017.3260.

del Arco, F. M. P., Molina-González, M. D., López, L. A. U., & Valdivia, M. T. M. (2021). Comparing pre-trained language models for Spanish hate speech detection. In *Expert syst. appl.*: *vol. 166*, (p. 114120). http://dx.doi.org/10.1016/j.eswa.2020.114120.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, https://aclanthology.org/N19-1423.

Firmino, A. A., de Souza Baptista, C., & de Paiva, A. C. (2021). Using cross lingual learning for detecting hate speech in Portuguese. In C. Strauss, G. Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Lecture notes in computer science*: *vol. 12924*, *Database and expert systems applications - 32nd international conference, DEXA 2021, Virtual Event, September 27-30, 2021, proceedings, Part II* (pp. 170–175). Springer, http://dx.doi.org/10.1007/978-3-030-86475-0_17.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. In *ACM Comput. Surv.*: *vol. 51*, (no. 4), (pp. 85:1–85:30). http://dx.doi.org/10.1145/3232676.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., & Nunes, S. (2019). A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online* (pp. 94–104). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-3510, https://aclanthology.org/W19-3510.

Fortuna, P., Soler Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? In *Inf. Process. Manag.*: *vol. 58*, (no. 3), (p. 102524). http://dx.doi.org/10.1016/j.ipm.2021.102524.

Frenda, S., Ghanem, B., Montes-y-Gómez, M., & Rosso, P. (2019). Online hate speech against women: Automatic identification of Misogyny and sexism on Twitter. *36*, In *Journal of Intelligent & Fuzzy Systems* (5), (pp. 4743–4752). http://dx.doi.org/10.3233/JIFS-179023.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, & T. Tokunaga (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation*. European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2018/summaries/627.html.

Hartmann, N., Fonseca, E. R., Shulby, C., Treviso, M. V., Rodrigues, J. S., & Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In G. H. Paetzold, & V. Pinheiro (Eds.), *Proceedings of the 11th Brazilian symposium in information and human language technology* (pp. 122–131). Sociedade Brasileira de Computação, https://aclanthology.org/W17-6615/.

van der Heijden, N., Yannakoudakis, H., Mishra, P., & Shutova, E. (2021). Multilingual and cross-lingual document classification: A meta-learning approach. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021* (pp. 1966–1976). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.eacl-main.168.

Hewitt, S., Tiropanis, T., & Bokhove, C. (2016). The problem of identifying misogynist language on Twitter (and other online social spaces). In W. Nejdl, W. Hall, P. Parigi, & S. Staab (Eds.), *Proceedings of the 8th ACM conference on web science* (pp. 333–335). ACM, http://dx.doi.org/10.1145/2908131.2908183.

Karim, M. R., Dey, S. K., Islam, T., Sarker, S., Menon, M. H., Hossain, K., Hossain, M. A., & Decker, S. (2021). Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th international conference on data science and advanced analytics* (pp. 1–10). IEEE, http://dx.doi.org/10.1109/DSAA53316.2021.9564230.

Kemp, S. (2021). https://datareportal.com/reports/digital-2021-global-overview-report.

Kottasová, I. (2017). https://money.cnn.com/2017/06/01/technology/twitter-facebook-hate-speech-europe/index.html.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. In *International conference on learning representations*. https://openreview.net/forum?id=H196sainb.

Lima, C., & Bianco, G. D. (2019). Feature extraction to identify hate speech in documents (in Portuguese). In *Proceedings of the XV database workshop (in Portuguese)* (pp. 61–70). Porto Alegre, RS, Brasil: SBC, http://dx.doi.org/10.5753/erbd.2019.8479, https://sol.sbc.org.br/index.php/erbd/article/view/8479.

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. In P. Boldi, B. F. Welles, K. Kinder-Kurlanda, C. Wilson, I. Peters, & W. M. Jr. (Eds.), *Proceedings of the 11th ACM conference on web science* (pp. 173–182). ACM, http://dx.doi.org/10.1145/3292522.3326034.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio, & Y. LeCun (Eds.), *1st International conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop track proceedings*. http://arxiv.org/abs/1301.3781.

Mondal, M., Silva, L. A., Correa, D., & Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *24*, In *New Review of Hypermedia and Multimedia* (2), (pp. 110–130). http://dx.doi.org/10.1080/13614568.2018.1489001.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2022). Cross-lingual few-shot hate speech and offensive language detection using meta learning. *10*, In *IEEE Access* (pp. 14880–14896). http://dx.doi.org/10.1109/ACCESS.2022.3147588.

Nobata, C., Tetreault, J. R., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, & B. Y. Zhao (Eds.), *Proceedings of the 25th international conference on world wide web* (pp. 145–153). ACM, http://dx.doi.org/10.1145/2872427.2883062.

Nozza, D. (2021). Exposing the limits of zero-shot cross-lingual hate speech detection. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 907–914). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-short.114.

Pamungkas, E. W., Basile, V., & Patti, V. (2021). A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. In *Information processing and management*: *vol. 58*, (vol. 4), (p. 102544). http://dx.doi.org/10.1016/j.ipm.2021.102544, https://www.sciencedirect.com/science/article/pii/S0306457321000510.

Pamungkas, E. W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In F. Alva-Manchego, E. Choi, & D. Khashabi (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop* (pp. 363–370). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/p19-2051.

Pari, C., Nunes, G., & Gomes, J. (2019). Evaluation of word embedding techniques in the hate speech detection task (in Portuguese). In *Proceedings of the XVI national conference of artificial and computational intelligence (in Portuguese)* (pp. 1020–1031).

Porto Alegre, RS, Brasil: SBC, https://sol.sbc.org.br/index.php/eniac/article/view/9354.

Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W. (2018). Dissecting contextual word embeddings: Architecture and representation. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1499–1509). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d18-1179.

Pikuliak, M., Simko, M., & Bieliková, M. (2021). Cross-lingual learning for text processing: A survey. *165*, In *Expert Syst. Appl.* (p. 113765). http://dx.doi.org/10.1016/j.eswa.2020.113765.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: A systematic review. *55*, In *Lang. Resour. Evaluation* (2), (pp. 477–523). http://dx.doi.org/10.1007/s10579-020-09502-8.

Ranasinghe, T., & Zampieri, M. (2021). Multilingual offensive language identification for low-resource languages. In *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*: *vol. 21*, (no. 1), New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3457610.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In A. Korhonen, D. R. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers* (pp. 1668–1678). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/p19-1163.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In L. Ku, & C. Li (Eds.), *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/w17-1101.

Schweter, S. (2020). Italian BERT and ELECTRA models. http://dx.doi.org/10.5281/zenodo.4263142.

Shearer, E., & Mitchell, A. (2021). https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/.

Silva, S. C., Serapião, A. B., & Paraboni, I. (2019). Hate-speech detection in Portuguese using CNN and psycho-linguistic dictionary. *Journal of Information Data Management*, *5*, 1–12.

Soto, C. P., Nunes, G. M. S., Gomes, J. G. R. C., & Nedjah, N. (2022). Application-specific word embeddings for hate and offensive language detection. In *Multim. tools appl.*: *vol. 81*, (no. 19), (pp. 27111–27136). http://dx.doi.org/10.1007/s11042-021-11880-2.

Souza, F., Nogueira, R., & de Alencar Lotufo, R. (2020). Bertimbau: Pretrained BERT models for Brazilian portuguese. In R. Cerri, & R. C. Prati (Eds.), *Lecture notes in computer science*: *12319*, *Intelligent systems - 9th Brazilian conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, proceedings, Part I* (pp. 403–417). Springer, http://dx.doi.org/10.1007/978-3-030-61377-8_28.

Stappen, L., Brunn, F., & Schuller, B. W. (2020). Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and AXEL. CoRR abs/2004.13850, https://arxiv.org/abs/2004.13850.

U. N. Human Rights Council (2013). https://www.refworld.org/docid/50f925cf2.html.

Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In A. Armando, R. Baldoni, & R. Focardi (Eds.), *CEUR workshop proceedings*: *vol. 1816*, *Proceedings of the first italian conference on cybersecurity* (pp. 86–95). CEUR-WS.org, http://ceur-ws.org/Vol-1816/paper-09.pdf.

Wagner, K. (2020). https://www.bloomberg.com/news/articles/2020-08-11/facebook-pulls-22-5-million-hate-speech-posts-in-second-quarter.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the student research workshop, SRW@HLT-NAACL 2016, the 2016 conference of the North American Chapter of the association for computational linguistics: Human language technologies* (pp. 88–93). The Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n16-2013.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144, http://arxiv.org/abs/1609.08144.

Zhang, E., & Zhang, Y. (2009). F-measure. In L. LIU, & M. T. ÖZSU (Eds.), *Encyclopedia of database systems* (p. 1147). Boston, MA: Springer US, http://dx.doi.org/10.1007/978-0-387-39940-9_483.