Using Multilingual Approach in Cross-Lingual Transfer Learning to Improve Hate Speech Detection

Aillkeen Bezerra de Oliveira[1][a], Cláudio de Souza Baptista[1][b], Anderson Almeida Firmino[1][c]
and Anselmo Cardoso de Paiva[2][d]

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

Keywords: Hate Speech Detection, Natural Language Processing, Cross-Lingual Learning.

Abstract: In the Internet age people are increasingly connected. They have complete freedom of speech, being able to share their opinions with the society on social media. However, freedom of speech is often used to spread hate speech. This type of behavior can lead to criminality and may result in negative psychological effects. Therefore, the use of computer technology is very useful for detecting and consequently mitigating this kind of cyber attacks. Thus, this paper proposes the use of a state-of-the-art model for detecting political-related hate speech on social media. We used three datasets with a significant lexical distance between them. The datasets are in English, Italian, and Filipino languages. To detect hate speech, we propose the use of a Pre-Trained Language Model (PTLM) with Cross-Lingual Learning (CLL) along with techniques such as Zero-Shot (ZST), Joint Learning (JL), Cascade Learning (CL), and CL/JL+. We achieved 94.3% in the F-Score metric using CL/JL+ strategy with the Italian and Filipino datasets as the source language and the English dataset as the target language.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Over the years, humanity has made significant advances in communication technology, such as radio, television, and the Internet. The Internet, combined with mobile devices such as tablets, cell phones and smartphones, allowed the transmission of information in real time.

According to the Datareportal[1], most of these devices are currently dedicated to social activities. People's interest in these activities and the availability of real-time communicability encouraged companies to create large social networks, facilitating the sharing of opinions among people. Social media is a structure composed of people or organizations that are connected by interests in which they share common opinions and objectives. The number of people interested in expressing their opinions on social platforms has become increasingly large (Mathew et al., 2019).

By using these networks (Facebook, Instagram, Twitter, YouTube, TikTok, etc.) the population has complete freedom of speech, being able to share their ideologies, opinions, dissatisfactions, happiness, unhappiness, etc. This kind of sharing occurs most often through texts open to the public and/or directed to someone so that anybody can see it and discuss it.

However, this freedom of speech is also used to spread aggressiveness on social media, because people produce attacks such as cybernetic aggression (Fortuna and Nunes, 2018; Mladenovic et al., 2021; Whittaker and Kowalski, 2015). Such behavior produces what is called Hate Speech. (Fortuna and Nunes, 2018) defined hate speech as a language that encourages the increase of violence and it leads to attacks on certain people groups. Most attacks target people who fit certain aspects, such as physical appearance, religion, descent, nationality, ethnic origin, gender, etc.

The term hate speech became the subject of great interest and research in computing (Mladenovic et al., 2021). The dispersion of hate speech on social networks can negatively affect psychologically people who are targets of this kind of attack. It can lead victims to more serious psychological problems such as

[a] https://orcid.org/0000-0002-0736-4945
[b] https://orcid.org/0000-0002-2200-1405
[c] https://orcid.org/0000-0003-2199-8191
[d] https://orcid.org/0000-0003-4921-0626
[1] https://datareportal.com/social-media-users

de Oliveira, A., Baptista, C., Firmino, A. and de Paiva, A.
Using Multilingual Approach in Cross-Lingual Transfer Learning to Improve Hate Speech Detection.
DOI: 10.5220/0011851800003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 1*, pages 374-384
ISBN: 978-989-758-648-4; ISSN: 2184-4992
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

depression, anxiety disorders and even suicide. These aggressive acts practiced by people can also encourage other individuals to do the same behavior. This behavior can generate a negative propagation of these acts, which can reach more victims in society. Besides that, we can also correlate behaviors like these with criminal practices (Mondal et al., 2018). Therefore, this is a subject of great interest and study, because reducing hate speech cases means a decrease in cases of violence and criminality. If we reduce the spread of hate speech, it also increases the well-being of people who suffer from this type of attack.

However, mitigating this problem is not easy because the amount of messages sent daily is very large. Twitter, for instance, is one of the social networks with the most problems with messages related to hate speech (Hewitt et al., 2016). According to the Internet Live Stats[2], this social network published 500 million messages daily, which corresponds to 350,000 posts per minute. However, Twitter relies on the collaboration of the platform users to identify and remove these aggressive comments (Hewitt et al., 2016). The task of watching, deleting, or manually restricting messages from social media like that is extremely exhausting and expensive for companies that own this type of social platform.

Given the lack of control and the infeasibility of monitoring hate speech by humans, computational techniques can be used to speed up, reduce costs and automate the detection of such problems. Natural Language Processing (NLP) and state-of-the-art techniques in Machine Learning can be useful to detect and control hate speech in social media (Agrawal and Awekar, 2018; López-Vizcaíno et al., 2021; Chang et al., 2021). Therefore, considering the great importance of computing in this social problem, this paper aims to use state-of-the-art computational tools in NLP, Cross-Lingual Learning (CLL), and Pre-Trained Language Model (PTLM) to detect cases of hate speech in social media texts.

We structured the remainder of this article as follows. In Section 2, discusses related works on hate speech detection. In Section 3, we highlight our method for hate speech detection. We present the corpora used in our experiment in Section 4. Section 5, focuses on the experiments that we performed. We present an analysis of the results from the experiments in Section 6. Concluding remarks and limitations about the proposed work are addressed in Section 7.

2 RELATED WORK

(Waseem and Hovy, 2016) built a dataset with 3,383 tweets related to sexist content and 1,972 tweets related to racist content. They used logistic regression to classify the texts. They obtained 73.93% in the F-Score metric. The authors only performed experiments and analyzed the results using a single model (RL), thus not making comparisons with other computational models. Moreover, there is no comparison with other datasets with more than one language.

(Davidson et al., 2017) mention that there is a problem in distinguishing and classifying Hate Speech sentences from other common offenses. The authors used the Hatebase.org website that contains a set of terms considered offensive. They used these terms to gather the tweets. Once they collected the data, the authors randomly selected 25,000 tweets from this dataset. The authors submitted the tweets to CrowdFlower so that people could label the collected data. They used traditional models for classification. The best metric that the authors obtained was 91% in the F-score. They did not perform any experiments using datasets with more than one language.

(Fortuna and Nunes, 2018) did an important survey and developed a definition of the concept of Hate Speech based on the code of conduct of the European Union Commission, the terms and conditions of social networks such as Facebook and Twitter, and scientific articles in the area. They categorized the articles according to the tools used to detect hate speech (n-grams, TF-IDF, etc.), the model used for classification, and also the domain of hate speech, such as racism, sexism, etc. They pointed out some challenges and opportunities related to this area of research. Despite providing a wealth of information on hate speech, the authors have explored very little about this area when related to multiple languages. Most of the works presented by the authors are related to the English language, so there is a lack of research related to other languages.

(Frenda et al., 2019) focused on hate speech related to women. The authors investigated analogies and differences between sexism and misogyny from a computational point of view. They used data on misogyny called IberEval 2018 and Evalita 2018 and they also used data from (Waseem and Hovy, 2016) that contains data on sexism. The authors performed n-grams and TF-IDF on data and used SVM as a classifier to detect hate speech in texts. The authors did not perform experiments with more than one language. They only used data in the English language. The authors used only accuracy to evaluate the performance of the model, they did not used other met-

[2] https://www.internetlivestats.com/twitter-statistics

rics such as precision, recall, and f-score. In addition, they only used a single model (SVM) to perform the experiment, with no comparisons to other models.

(Pamungkas and Patti, 2019) used the concept of transfer learning. The authors used datasets in the following languages: English, Spanish, Italian, and German. The authors performed experiments using the models: Linear Support Vector Classifier (LSVC) and Long Short Term Memory (LSTM). They achieved the best result using the Joint Learning strategy, along with Hurtlex (Bassignana et al., 2018). They used datasets with more than one language, however, instead of using the source datasets, they used the Google Translation API to translate the non-English datasets. In addition, the authors did not take any approach to mitigate translation errors made automatically by the API.

(Stappen et al., 2020) used an approach to detect hate speech in more than one language, in which they added part of the classification target language in the model training step. The authors used an approach called Attention-Maximum-Average Pooling (AXEL) and a FastText-generated embedding and the extractor feature (BERT or XLM) to accomplish this task. They used datasets with more than one language, however, instead of using the source datasets, the authors used an Amazon tool to translate the data into English. They did not use any approach to mitigate translation errors by the tool.

(Corazza et al., 2020) used a technique to detect hate speech regardless of the language being used. They developed a modular neural architecture that contains a hidden layer of 100 neurons. They performed some experiments in which they achieved the best results for the corpus in English using LSTM. Besides the English corpus, they performed experiments with an Italian dataset in which they obtained the best values with the LSTM model combined with character embeddings, unigrams, and emoji transcripts. The authors also used a German dataset and they obtained the best result for that language using character embeddings and a GRU network. They also used resources such as embeddings, emoji embeddings, n-grams, emotional lexicon, and social network resources (hashtags, mentions, links, etc). The authors reported that they transcribed the emojis into english text using a python library and then they used the Google translation tool to translate English into Italian and German, but they did not mention the level of accuracy of this translation.

(del Arco et al., 2021) addressed the detection of Hate Speech on social networks in the Spanish language. The authors compared the performance of Deep Learning models with more recent pre-trained Transfer Learning models and with traditional machine learning models. They used two datasets. The first one is a dataset with 6,000 tweets that were collected using HaterNet: an intelligent system used by the Spanish National Office against Hate Crimes of the Secretary of State for Security in Spain. The second dataset was provided by SemEval 2019, called HatEval 2019, with 1,600 tweets. They compared some Deep Learning models and traditional models. Deep Learning models outperformed traditional models according to the authors. Besides that, the authors concluded that models using Transfer Learning had great results. The authors performed the experiments with an unbalanced dataset. They did not perform the same experiment with balanced data to compare the results. In addition, the authors used one language (Spanish) in the experiment, with no comparison of models tested in other languages.

(Bigoulaeva et al., 2021) presents the problem of Hate Speech, showing that it is difficult to detect because social networks are vast. The authors used data in German and English for the training and classification of the models. This approach was used to train the model in one language and test it in another language. In the experiment, the authors used the English as source language and German as the target language. They used multiple models of Deep Neural Networks using zero-shot and joint learning strategies. The authors built the models using transfer learning based on Cross-Lingual Learning (CLL) and Bilingual Word Embedding(BWE). However, the authors did not perform experiments with balanced data. They used two languages whose lexical distance between them is close. Therefore, they did not perform experiments between languages with a greater lexical distance to compare the results.

(Pamungkas et al., 2021) used CLL to detect hate speech in texts. The authors used seven languages: English, Portuguese, French, Spanish, German, Indonesian, and Italian. They used English as the source language and the other six languages as the target language. The authors used traditional machine learning models and Deep Learning models such as BERT. They used one-shot and joint-learning as learning strategies in their experiments. They concluded that the best model tested by them was an LSTM neural network along with multilingual embeddings provided by Facebook (MUSE - (Lample et al., 2018)). The authors proposed a model called Joint Learning, which comprises using two models to classify the data separately and, at the end of the model, connecting the results obtained by these two models through a dense layer. The results presented by this model were promising when compared to other models or comparison techniques presented in the article (LASER Embeddings, Facebook Muse, and BERT Multilingual). However, in the experiments, the authors used Google's translation API to translate the texts into English to train the models with the translated data. The authors did not take any approach to mitigate translation errors made by the API. In addition, the authors did not perform isolated experiments using a pre-trained model for each language.

(Karim et al., 2021) used Machine Learning, Deep Learning, and PTLMs to detect hate speech in Bengali texts. The authors used multiple techniques, such as Naive Bayes, Logistic Regression, CNN, and PTLMs. They achieved the best results using a combination of PTLMs (BERT Bengali, XLM-R, and BERT Multilingual). They achieved state-of-the-art detection in Bengali language texts containing hate speech, scoring around 88% on the F1-Score. The data used in the experiment are unbalanced in their categories, with the most predominant being the personal attacks category and the least predominant being the political category. However, they did not perform experiments with the data balanced for each category.

(Soto et al., 2022) used a CNN network together with different embeddings to perform text classification. They used embeddings for the corpus called HSD and embeddings obtained from NILC (Hartmann et al., 2017). The authors used and tested the Wang2Vec, Word2Vec, FastText, and Glove embeddings. The authors achieved the best result for the HSD corpus using the Glove with 300 dimensions combined with the NILC embeddings.

Most related works used only a single language as a source and works that use more than one language usually used translation tools. In this work, we use more than one source language to detect hate speech in a target language, without translation tools. Most of the related works use languages with a close lexical distance. In this work, we will address the classification of hate speech using languages whose lexical distance is not close.

3 METHODOLOGY

In this work, we used CLL to detect hate speech in texts published by users in social media. The first step to detect hate speech in social media is to collect the data necessary to train the model. In our experiment, we used data from three different languages.

In the second step, we used a Pre-Trained Language Model (PLTM) to carry out the data classification. In the third step, we submitted the model to four training strategies (Pikuliak et al., 2021) to detect hate speech in a target language (T_L). In order to achieve this goal we used two different languages as source language (S_L).

Finally, in the last step, we performed the model evaluation. We evaluated the model using the results obtained in the classification of the target language (T_L). We used Precision, Recall, and F-measure to get model evaluation metrics. Figure 1 demonstrates an overview of the methodology that we used. In the next subsections, we give more details on the methodology.

3.1 Corpora Acquisition and PTLM

Usually, we need a dataset so we can use it to train the model. In CLL, at least two corpora are required: one in the source language and the other one in the target language. These corpora can be obtained by crawlers to collect texts on websites. For instance, (de Pelle and Moreira, 2017) collected politics and sports data by comments from news pages. Another way to collect data is by using an API. This way we can pass search parameters to the API[3] and it returns the texts that were found based on the informed parameters. For instance, Twitter provides an API to search for tweets published by its users. We can use data that other works have already collected, and that is publicly available. In most cases, these data are available by the authors themselves, or at conferences, events, workshops, etc.

Among the presented options, in this work, we chose a publicly available corpora. We used the corpora described in the works by (Vigna et al., 2017; Grimminger and Klinger, 2021; Cabasag et al., 2019). The next step is to find a PTLM to train it with the collected corpora. In this work, we chose BERT as the PTLM option. The corpora that we used are from three languages: Italian, English, and Filipino. In Section 4, we describe why we chose these three languages. Because of the choice of these three languages, we used two BERT distributions: Italian BERT (Schweter, 2020), and English BERT (Devlin et al., 2019). We did not use any BERT Filipino language distribution because we did not find a pre-trained BERT model for that language.

Another thing necessary for this step is data pre-processing. Usually, most of the pre-processing techniques in Deep Learning use vector representation of data. We can achieve this using approaches such as Word2Vec (Mikolov et al., 2013), FastText (Grave et al., 2018), ELMO (Peters et al., 2018), etc. Usually, pre-trained models already have some functions to perform data pre-processing. The model we chose

[3] https://developer.twitter.com/en/docs/twitter-api

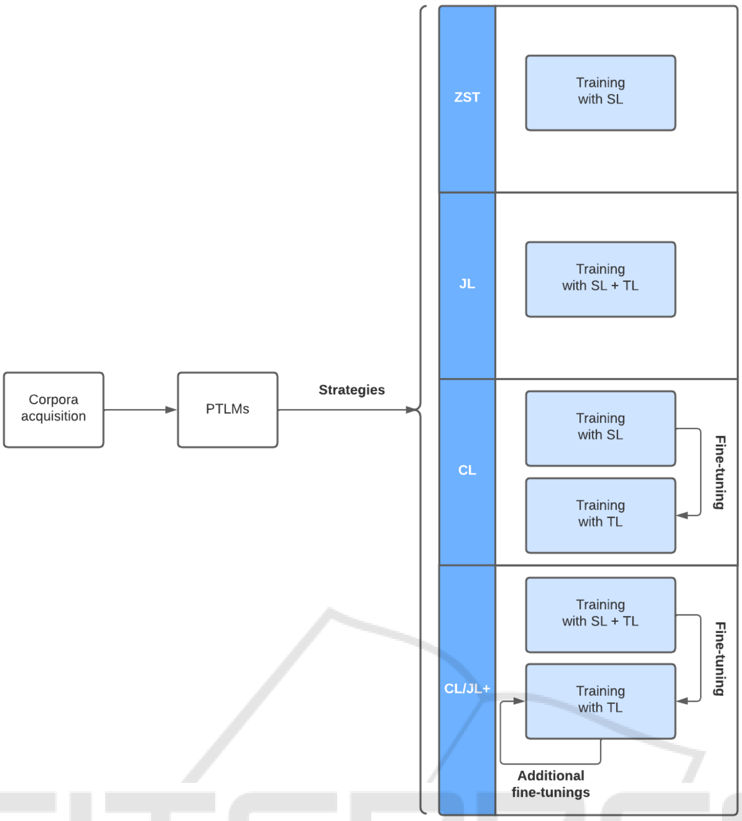

Figure 1: Methodology overview.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

already has a module responsible for carrying out this task.

3.2 Training Steps

To carry out model training, we used five approaches:

- Zero-shot transfer (ZST): in this strategy, we use only the source language(S_L) in the first fine-tuning of the PTLM. The target language(T_L) is used to evaluate the model;
- Joint Learning (JL): in this strategy, we use both S_L and T_L simultaneously in the first fine-tuning of the PTLM;
- Cascade Learning (CL): in this strategy, we use the S_L data in the first fine-tuning, then we use the T_L data in the second fine-tuning;
- CL/JL+: in this approach, we use both CL and JL strategies. Therefore, we use a percentage of the T_L data in the first fine-tuning, and the remaining percentage of the T_L data is used for the test and evaluation stages of the PTLM. Besides that, in the fine-tuning stage, we perform multiple fine-tunings using T_L data.

In section 5, we show details of how we used each one of these strategies in our experiment.

3.3 Evaluation

This is the last step. We used metrics to evaluate the model after the previous steps. Hence, we used Precision, Recall, and F1-measure (Zhang et al., 2009). It is worth mentioning that we used weighted F1-measure for model evaluation in our experiment.

4 CORPORA USED IN OUR EXPERIMENT

Some related works on hate speech detection used datasets that contain texts related to several domains at the same time, such as religion, sexism, racism, etc. In our work, we did a different approach. We used only a single domain related to hate speech to verify if we could obtain good results. Therefore, in our experiments, we used three datasets related to politics, a domain present in most of the world.

Many works related to CLL and hate speech detection have datasets whose lexical distance between

Table 1: Corpora summary.

Corpus	Texts without hate speech	Texts with hate speech	Total	Context
Italian	1,941 (48.5%)	2059 (51.5%)	4,000	Politics
English	2,640 (88%)	360 (12%)	3,000	Politics
Filipino	9,864 (53.42%)	8,600 (46.58%)	18,464	Politics

their languages is close. For instance, Portuguese and Spanish or Italian and French, etc. For that reason, in this work, we used three datasets with a significant lexical distance between them. Our goal is to verify if there are any significant results in the model results, even in languages with a significant lexical distance. The datasets that we used are in English, Italian, and Filipino languages. Table 1 summarizes the data.

4.1 The Evalita 2018 Corpus

The Evalita 2018 corpus has 17,000 texts collected from Facebook users' comments in Italian. The Istituto di Informatica e Telematica, CNR, Pisa (Vigna et al., 2017) created the corpus. The corpus contains 1,941 texts labeled as non-hateful, and 2,059 texts labeled as hateful.

4.2 The Grimminger Klinger WASSA 2021 Corpus

(Grimminger and Klinger, 2021) collected 3,000 texts in English from the 2020 election between Biden and Trump. They collected it from a social media during the election campaign. They labeled 360 texts as hateful and 2,640 texts as non-hateful.

4.3 The Filipino Corpus

(Cabasag et al., 2019) collected Filipino tweets and labeled 8,600 texts as hate speech and 9,864 as non-hate speech. The authors collected the data during the 2016 Philippine Presidential Elections.

5 EXPERIMENTS

In this section, we describe the settings and some information about our experiments.

5.1 Experiment Description

We used several strategies during the experiment to achieve significant results. For each experiment, we used the strategies described in section 3.2: ZST, JL, CL, and CL/JL+.

We used two datasets as S_L and one dataset as T_L in each experiment. In the first experiment, we used the English and Filipino datasets as S_L, the Italian dataset as T_L, and the Italian PTLM BERT. In the second experiment, we used the Italian and Filipino datasets as S_L, the English dataset as T_L, and the English PTLM BERT. Most experiments from other works use only two datasets for CLL tasks. These datasets usually are from two different languages. They use one of these datasets as S_L and the other one as T_L. In our experiment, we used three datasets with different languages. We used two datasets as S_L and one dataset as T_L. Our goal is to verify if there is a significant increase in the results when adding another language as S_L.

We used the same settings for all PTLMs in this work. Our experiments used a learning rate of 1×10^5, and a number of epochs equal to three (Devlin et al., 2019; Conneau et al., 2021). 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 ran all experiments on Google Colab with an Nvidia Tesla K80 GPU and the PyTorch library.

It was necessary to pre process the data. This process comprises removing noise from the text, as well as undesirable characters, such as URLs, emojis, special characters, blank lines or spaces, etc. This pre processing is fundamental, as it will help the PTLM to better understand the data improving the results.

After pre processing the data, we used the data to train and test the model. We used 128 as text size tokenization because, after we analysed data, most of the texts was less than 128 words. We did that to improve the experiment performance and to reduce computational costs. Therefore, the model mapped the entries into a vector with a dimension equal to 128. When the text length is smaller than this value, the vector is filled with 0's. When the text is greater than this value, it was truncated.

In the ZST strategy, we used 90% of the corpus data S_L when training the PTLMs, and the remaining 10% for testing. In the test step, we used only the corpus T_L. Regarding the JL strategy, we used the same proportion of the data presented in the ZST strategy. However, we used a subset of 30% of the test corpus T_L in training, and the remaining 70% left we used only for the final test of the model.

In the CL strategy, we used 70% of the S_L corpus for training, 10% for the validation step, and 20% for the test step. After performing this process, we adjusted the PTLMs. We then used the T_L corpus, following the same previous pattern, with 70% of the T_L corpus for training, 10% for validation, and 20% for testing.

In the CL/JL+ strategy, we used 70% of the data from the corpus S_L and 30% of the data from the corpus T_L for training. For validation, we used 20% of the data from the corpus S_L, and 10% of the data from the corpus T_L. We used the remaining data from the corpus S_L to perform the test.

We used the remaining data from the T_L corpus to perform the fine-tuning. Therefore, in this step, we divided the corpus of the remaining T_L according to the number of fine adjustments performed. We used k-fold cross-validation (k=5) to guarantee the proportionality of the classes over the iterations. For each cross-validation iteration, we kept the same proportion as in the previous steps, 70% for training, 20% for validation, and 10% for testing.

5.2 Monolingual Baseline Experiment

We did an experiment using only one corpus based on the PTLM language. Our objective is to create a baseline experiment to understand if using CLL brings any significant improvement to the proposed method. To accomplish this, we did not use a T_L dataset. Therefore, we did not use an auxiliary corpus in the method.

In the first experiment, we used PTLM BERT English. We used 80% of the English dataset to train the PTLM and 20% to test it. In the second experiment, we used the PTLM BERT Italian and 80% of the Italian dataset to train it and 20% to test the PTLM. We did not use the Filipino dataset because none of the two PTLMs belongs to that language. Table 2 presents the results.

Table 2: Monolingual baseline results.

PTLM	Precision	Recall	F1-Score
BERT English	80%	90%	85%
BERT Italian	81%	84%	82%

5.3 Experiment Results

In the first experiment, we used the English and Filipino datasets as S_L, the Italian dataset as T_L, and the Italian PTLM BERT. In the second experiment, we used the Italian and Filipino datasets as S_L, the English dataset as T_L, and the English PTLM BERT. We used these settings for all strategies. In Table 3 we can see the settings.

Table 3: Experiment settings.

PTLM	S_L	T_L
BERT English	Italian and Filipino	English
BERT Italian	English and Filipino	Italian

Table 4 shows the ZST strategy results using the settings from Table 3. The results to PTLM BERT were really poor, we only obtained 34% in the F1-Score. In contrast, we achieved 84% on the F1-Score for PTLM English, which is a good value.

Table 4: ZST strategy results.

PTLM	Precision	Recall	F1-Score
BERT English	86%	82%	84%
BERT Italian	59%	51%	34%

Regarding the JL training strategy, Table 5 presents the results using this strategy for the PTLMs. The results were really improved using this strategy, especially on PTLM Italian. We reached 83% in the F1-Score, which is a 49% of improvement regarding the previous results using the ZST strategy. Regarding the PTLM English, we obtained 86% in the F1-Score, which is an improvement of 2%.

Table 5: JL strategy results.

PTLM	Precision	Recall	F1-Score
BERT English	84%	88%	86%
BERT Italian	79%	87%	83%

Table 6 shows the CL strategy results. We obtained 84% in the F1-Score for PTLM Italian, which is a 1% of improvement regarding the previous results using the JL strategy. Regarding the PTLM English, we achieved 84.31% in the F1-Score, which is an improvement of 0.31%, but smaller improvement compared to the Italian model.

Table 6: CL strategy results.

PTLM	Precision	Recall	F1-Score
BERT English	85%	89%	84.31%
BERT Italian	85%	84%	84%

Table 7 shows the CL/JL+ strategy results. We used k-fold cross-validation (k=5) in this strategy, then we did the average of scores to get the results. For both PTLMs we obtained good values using this strategy. We achieved 92.4% in the F1-Score for PTLM Italian and regarding the PTLM English we obtained 94.3% in the F1-Score, which is a good value too.

Table 7: CL/JL+ strategy results.

PTLM	Precision	Recall	F1-Score
BERT English	94.1%	94.5%	94.3%
BERT Italian	93%	92%	92.4%

6 DISCUSSION

In this section, we will comment on the results accomplished by the CLL and the strategies that we used. Initially, we performed an experiment with just a single language. Our goal was to create a base experiment to verify if the CLL can improve the results obtained when adding more than one source language. Table 2 shows the results. We achieved 85% in the F1-Score metric using BERT English and 82% on that same metric using BERT Italian.

After that, we performed several other experiments using the strategies mentioned in section 3.2. In these strategies, we used two languages as the source and one language as the target, as shown in Table 3. We have a big challenge here because we need to get a good result using two languages in the training (S_L) with a lexical difference between them. In addition, the training languages have a significant lexical difference from the target language (T_L) as well.

We used ZST as the first strategy. Table 4 shows the results. The results were worse in this strategy compared to the baseline experiment. We obtained 85% in the F1-Score metric for the BERT English model and 34% in this same metric for the BERT Italian model. Both results were lower than the base experiment, especially the Italian model, which performed poorly than expected. We believe this happened because we added two languages (English and Filipino) as S_L with a very large lexical distance compared to the T_L.

Table 5 shows the results for the second strategy (JL). The results were better than the ZST strategy. We were able to obtain values of 86% in the F1-Score for the English model and 83% for the Italian model. These results were better than those of the baseline experiment too. We think the results were better because in this strategy we added part of the T_L data in the first fine-tuning of the PTLMs.

Table 6 presents the results for the third strategy (CL). The result was better for the Italian model, showing an improvement to the result obtained for this model compared to the results that we achieved in the JL strategy. However, the English model had a worse result compared to the previous strategy.

Finally, Table 7 shows the results for the last strategy (CL/JL+). We obtained the best results with it. We achieved 94.3% in the F1-Score using the BERT English model and 92.4% using the BERT Italian model. Both values are significant, especially if we consider that the languages used in both models have a large lexical distance. We believe the results were better than the previous strategies because in this strategy we performed multiple fine-tunings in conjunction with cross-validation.

7 CONCLUSION

In this work, we developed strategies for detecting hate speech in texts. In most works involving hate speech detection, the authors usually use only a single language as S_L to perform this task. Most of the time, this happens because of the lack of data, making the works restricted to only one language. In addition, many works only use data in English to carry out the experiments because it is a widely used language globally, making most hate speech detection restricted to that language. Therefore, there is a lack of experiments aimed at other languages.

In this work, we used more than one language to perform hate speech detection, because this way we can reduce the problem of lack of data and expand the detection of hate speech using multiple languages, even if these are not in the same language as the target texts used to perform detection.

To achieve this goal, we used three datasets from three different languages: English, Italian, and Filipino. Our objective was to verify if using CLL can help us get results similar to experiments that use only a single language or if it is possible to reach better results, even using different languages that have a considerable lexical distance.

We did a baseline experiment to compare with our experiments using CLL. We did this to verify if using CLL we could really get significant results when compared to an experiment that uses only one language. Besides the CLL, we used some strategies to improve the results obtained by the model. The strategies are Zero-shot transfer, Joint Learning, Cascade Learning (CL), and CL/JL+. For each of them, we used two datasets as source(S_L) and one dataset as target (T_L), as shown in Table 3.

The best strategy was the CL/JL+. We achieved 94.3% in the F1-Score using the BERT English model and 92.4% using the BERT Italian model. Both values are significant, especially if we consider that the languages used in both models have a large lexical distance. Thus, we can conclude that using more than one language in conjunction with the CLL is indeed promising to accomplish good results in the detection of hate speech.

7.1 Limitations

A limitation of this work is that we did not perform any experiments with balanced data. Therefore, we do not know if the results could be better or worse with balanced data.

Another limitation is that we only used a corpus on politics. We do not know the model behavior when performing an experiment on a dataset containing texts on various domains at the same time, such as sexism, racism, etc.

Another limitation of this work is the subjectivity of data labeling. Different people annotated the data, so there may be subjectivity in the text labeling. For instance, there are texts that one person could identify as hate speech, but someone else who is labeling the data might not identify as hate speech in that same text.

7.2 Future Work

We suggest as future work to use datasets from other languages with a smaller lexical distance. It would also be interesting to carry out the experiment with other PTLM besides BERT to compare results. In addition, the experiment could be carried out with balanced data, perhaps the results could be better.

In this work, we used data related to politics. We suggest also running the experiment using hate speech data related to other domains, such as sexism, racism, etc. Besides that, we used three datasets in this work. It would also be interesting to carry out the experiment with more source languages or with more target languages to verify if there are significant improvements to the model.

ACKNOWLEDGEMENTS

The authors would like to thank the Brazilian National Council for Scientific and Technological Development (CNPq) for partially funding this work.

REFERENCES

Agrawal, S. and Awekar, A. (2018). Deep learning for detecting cyberbullying across multiple social media platforms. In *In Proceedings of the European Conference in Information Retrieval (ECIR), pages 141-153, Grenoble, France.*

Bassignana, E., Basile, V., and Patti, V. (2018). Hurtlex: A multilingual lexicon of words to hurt. In *in E. Cabrio, A. Mazzei, F. Tamburini (Eds.), Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018, volume 2253 of CEUR Workshop Proceedings, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2253/paper49.pdf.*

Bigoulaeva, I., Hangya, V., and Fraser, A. (2021). Cross-lingual transfer learning for hate speech detection. In *in Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, LT-EDI@EACL 2021, Online, April 19, 2021, Association for Computational Linguistics, 2021, pp. 15–25. URL: https://www.aclweb.org/anthology/2021.ltedi-1.3/.*

Cabasag, N. V., Chan, V. R., Lim, S. C., Gonzales, M. E., and Cheng, C. (2019). Hate speech in philippine election-related tweets: Automatic detection and classification using natural language processing. In *In Philippine Computing Journal, XIV No. 1 August*.

Chang, V., Gobinathan, B., Pinagapan, A., and Kannan, S. (2021). Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. In *Computers and Electrical Engineering, Vol.92, pp. 1-17*.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In *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, Brno, Czechia, 30 August - 3 September 2021, ISCA, 2021, pp. 2426–2430. URL: https://doi.org/10.21437/Interspeech.2021-329. doi: 10.21437/Interspeech.2021-329.*

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2020). A multilingual evaluation for online hate speech detection. In *volume 20, 2020, pp. 10:110:22. URL: https://doi.org/10.1145/3377323.*

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *in: Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017, AAAI Press, 2017, pp. 512–515. U.*

de Pelle, R. P. and Moreira, V. P. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *in: Proceedings of the VI Brazilian Workshop on Social Network Analysis and Mining, SBC https://doi.org/10.5753/brasnam.2017.3260. doi:10.5753/brasnam.2017.3260.*

del Arco, F. M. P., Molina-González, M. D., López, L. A. U., and Valdivia, M. T. M. (2021). Comparing pre-trained language models for spanish hate speech detection. In *volume 166, p. 114120. URL: https://doi.org/10.1016/j.eswa.2020.114120 . doi: 10.1016/j.eswa.2020.114120.*

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional trans-

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

formers for language understanding. In *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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423 . doi: 10.18653/v1/N19-1423*.

Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. In *ACM Computing Surveys (CSUR) volume 51, 2018, pp. 85:1–85:30. URL: https://doi.org/10.1145/3232676 . doi: 10.1145/3232676*.

Frenda, S., Ghanem, B., y Gómez, M. M., and Rosso, P. (2019). Online hate speech against women: Automatic identification of misogyny and sexism on twitter. In *volume 36, 2019, pp. 4743–4752. URL: https://doi.org/10.3233/JIFS-179023 . doi: 10.3233/JIFS-179023*.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, European Language Resources Association (ELRA), 2018. URL:http://www.lrec-conf.org/proceedings/lrec2018/summaries/627.html*.

Grimminger, L. and Klinger, R. (2021). Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. In *In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 171–180, Online. Association for Computational Linguistics*.

Hartmann, N., Fonseca, E. R., Shulby, C., Treviso, M. V., Rodrigues, J. S., and Aluísio, S. M. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *in: G. H. Paetzold, V. Pinheiro (Eds.), Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology, STIL 2017, Uberlândia, Brazil, October 2-5, 2017, Sociedade Brasileira de Computação, 2017, pp. 122–131. URL: https://aclanthology.org/W17-6615/*.

Hewitt, S., Tiropanis, T., and Bokhove, C. (2016). The problem of identifying misogynist language on twitter (and other online social spaces). In *in: Proceedings of the 8th ACM Conference on Web Science, pp. 333–335*.

Karim, M. R., Dey, S. K., Islam, T., Sarker, S., Menon, M. H., Hossain, K., Hossain, M. A., and Decker, S. (2021). Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *in: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2021, pp. 1–10. URL: https://doi.org/10.1109/DSAA53316.2021.9564230. doi: 10.1109/DSAA53316.2021.9564230*.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In *in: International Conference on Learning Representations, 2018. URL: https://openreview.net/forum?id=H196sainb*.

López-Vizcaíno, M. F., Nóvoa, F. J., Carneiro, V., and Cacheda, F. (2021). Early detection of cyberbullying on social media networks. In *Future Generation Computer Systems (118), pp. 219-229*.

Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. (2019). Spread of hate speech in online social media. In *in Proceedings of the 10th ACM conference on web science, pp. 173–182*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013. URL: http://arxiv.org/abs/1301.3781*.

Mladenovic, M., Osmjanski, V., and Stankovic, S. V. (2021). Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. In *ACM Computing Surveys 54 (2021) 1:1–1:42*.

Mondal, M., Silva, L., Correa, D., and Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. In *New Review of Hypermedia and Multimedia 24, 110–130*.

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

Pamungkas, E. W. and Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *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, Association for Computational Linguistics, 2019, pp. 363–370. URL: https://doi.org/10.18653/v1/p19-2051*.

Peters, M. E., Neumann, M., Zettlemoyer, L., and Yih, W. (2018). Dissecting contextual word embeddings: Architecture and representation. In *in E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, Association for Computational Linguistics, 2018, pp. 1499–1509. URL: https://doi.org/10.18653/v1/d18-1179 . doi: 10.18653/v1/d18-1179*.

Pikuliak, M., Simko, M., and Bieliková, M. (2021). Cross-lingual learning for text processing: A survey. In *volume 165, 2021, p. 113765. URL: https://doi.org/10.1016/j.eswa.2020.113765. doi:10.1016/j.eswa.2020.113765*.

Schweter, S. (2020). Italian bert and electra models. In *2020. URL: https://doi.org/10.5281/zenodo.4263142. doi: 10.5281/zenodo.4263142*.

Soto, C. P., Nunes, G. M. S., Gomes, J. G. R. C., and Nedjah, N. (2022). Application specific word embeddings for hate and offensive language detection. In *volume 81, 2022, pp. 27111–27136. URL: https://doi.org/10.1007/s11042-021-11880-2. doi: 10.1007/s11042-021-11880-2*.

Stappen, L., Brunn, F., and Schuller, B. W. (2020). Cross-lingual zero and few-shot hate speech detection utilising frozen transformer language models and axel. In *volume abs/2004.13850, 2020. URL: https://arxiv.org/abs/2004.13850.arXiv:2004.13850*.

Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *in: A. Armando, R. Baldoni, R. Focardi (Eds.), Proceedings of the First Italian Conference on Cybersecurity (ITASEC17), Venice, Italy, January 17-20, 2017, volume 1816 of CEUR Workshop Proceedings, CEUR-WS.org, 2017, pp. 86–95. URL: http://ceur-ws.org/Vol-1816/paper-09.pdf*.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *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, San Diego California, USA, June 12-17, 2016, The Association for Computational Linguistics, 2016, pp. 88–93. URL: https://doi.org/10.18653/v1/n16-2013*.

Whittaker, E. and Kowalski, R. M. (2015). Cyberbullying via social media. In *Journal of School Violence, 14(1):11–29, 2015*.

Zhang, E., Zhang, Y., and F-Measure (2009). Springer us, boston, ma, 2009, pp. 1147–1147. url: https://doi.org/10.1007/978-0-387-39940-9_483. doi: 10.1007/978-0-387-39940-9_483.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS