www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(2), May-Aug 2025, 17-24. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2025.v16i2.534

NLP with Transformers for toxicity detection: corpus construction and evaluation for MisProfesores.com platform

María Lucía Barrón Estrada[1], Ramón Zatarain Cabada[1], Ramón Alberto Camacho Sapien[1], Víctor Manuel Bátiz Beltrán,[1] and Néstor Leyva López[1]
[1] TecNM-Instituto Tecnológico de Culiacán, Posgrado e Investigación, Culiacán, Sinaloa, México
{lucia.be, ramon.zc, ramon.cs, victor.bb, nestor.ll}@culiacan.tecnm.mx

Abstract. The growth of social networks as mass media has enabled faster and closer interaction between users, but it also presents challenges, such as the risk of spreading hate speech. Early detection of such harmful posts is critical. This article presents a methodology to create a unique corpus of Spanish-language comments collected from MisProfesores.com platform, covering all states in Mexico. This process resulted in a dataset of 18,000 unlabeled samples and 853 manually labeled samples. In addition to describing the corpus construction process, the results of the evaluation of different models trained with these data are presented, as well as their comparison with previous works for toxicity detection, highlighting the relevance of the Spanish corpus development for specific tasks. As a result, our Transformer-based model performed better than the state-of-the-art models in the binary toxicity classification, reaching a value of 0.9649 in accuracy and 0.9645 in F1 score.

Keywords: Sentiment analysis, Deep learning, BERT, Corpus, Toxicity, Transformers.

Article Info
Received October 19, 2024
Accepted March 12, 2025

1 Introduction

This paper is an extended version of a paper presented originally at the XVI Mexican Conference on Artificial Intelligence (XVI Congreso Mexicano de Inteligencia Artificial) (COMIA, 2024). Sentiment analysis is a branch of Natural Language Processing (NLP) which focuses on the automatic identification and classification of emotions and sentiments expressed in text (Tan et al., 2023). Today, sentiment analysis is applied to a wide variety of sectors in society. For example, digital media such as social networks, where sharing and exchanging ideas is a recurring activity, requires continuous monitoring to assure the privacy and integrity of users. This has turned these media into an important area to perform sentiment analysis. As a result, sentiment analysis has become increasingly popular among research communities in recent years (Wankhade et al., 2022).

In platforms such as MisProfesores (MisProfesores.com, n.d.), where Spanish is the prevalent language, the use of learning models for sentiment analysis represents a challenging task due to the complexity and diversity of the language. The development of effective machine learning models for sentiment analysis in Spanish texts is restricted by the existing Spanish corpora. In this paper, we will review the process of building a Spanish language corpus from the extraction of comments submitted by Mexican users on MisProfesores platform.

This article is structured as follows. In section 2, related works that refer to corpus construction in Spanish are reviewed. Section 3 describes the process of data collection and processing. Section 4 shows the methodology employed, including the data construction process, as well as the algorithms and machine learning models used for testing. The test results are presented in Section 5. Finally, in Section 6, we discuss our findings and conclusions.

2 Related Works

Toxicity detection on Internet comments has become a growing area of interest within the field of NLP. Different platforms that allow rating and reviewing teachers have become significant sources for sentiment analysis, given that they gather a wide variety of comments. A notable example of research in this area is the study presented in (Arceo-Gomez et al., 2019) where a comprehensive statistical analysis was conducted on approximately 600,000 evaluations. This study not only provides important perspectives on the interactions on such platforms, but also highlights the perception of gender stereotypes, underlining the importance of NLP techniques to identify and mitigate the presence of implicit biases.

On the other hand, the work presented in (Kolhatkar et al., 2020) describes the development of a substantial corpus of English-language comments collected from news websites, including nearly 500,000 samples. This study has particular relevance because they not only developed a large corpus, but also labeled a subset of approximately 1,000 samples, focusing on the toxicity classification of comments. This approach provides a solid foundation for future research on automated content moderation tools.

In the work reported in (Hartvigsen et al., 2022), the development of the TOXIGEN corpus is described, defined as a large-scale machine-generated dataset. The corpus consists of 274,000 toxic and nontoxic statements about 13 minority populations. The authors developed a proof-based prompting framework and an adaptive adversarial classifier decoding method to generate subtly toxic and benign texts with a pre-trained massive language model. The authors concluded that their experiments provide evidence that fine-tuning existing pre-trained hate classifiers using TOXIGEN can improve their performance on three popular human-generated toxicity datasets.

Regarding the use and effectiveness of artificial intelligence (AI) models, the research in (Nabiilah et al., 2023) provides a valuable side-by-side comparison between previously trained models on different corpora. The results showed that models trained with language-specific corpora performed better on toxicity classification tasks in the same language. This highlights the importance of considering cultural and language differences when designing and training AI models for content moderation.

Similarly, in (Fan et al., 2021), the creation of an effective model for detecting and classifying social media toxicity from user-generated content (tweets) using BERT is presented. The authors conducted the fine-tuning of the pre-trained BERT model and three of its variants (Multilingual BERT, RoBERTa and DistilBERT), using a publicly accessible corpus available on Kaggle platform (Toxic Comment Classification Challenge). The results revealed that the BERT-based model outperformed all compared models, achieving superior results, and confirmed that the model can effectively classify and analyze toxic tweets.

In the study published in (Roy et al., 2021) the authors use Transformer-based models to identify whether a comment on Twitter (now called X) has hateful and offensive content. With a previously trained Transformer-based multilingual text encoder as a basis, the authors report they were able to successfully identify and classify hate speech multilingually. In the test corpora used during the HASOC 2020 competition (Mandl et al., 2020), they obtained Macro F1 scores of 90.29, 81.87 and 75.40 for English, German and Hindi, respectively, on hate speech detection.

Contrasting with the previously mentioned works, this study presents an approach that merges the procedures and results reviewed above, by introducing a well-rounded methodology that covers areas such as opinion mining, dataset construction, and model training.

3 Methodology

A comprehensive process is detailed in the methodology section, from the initial definition of the desired features of the corpus to the training process of the different models using a manually labeled corpus developed in this work. A diagram illustrating each stage of the methodology is depicted in **Fig. 1**.

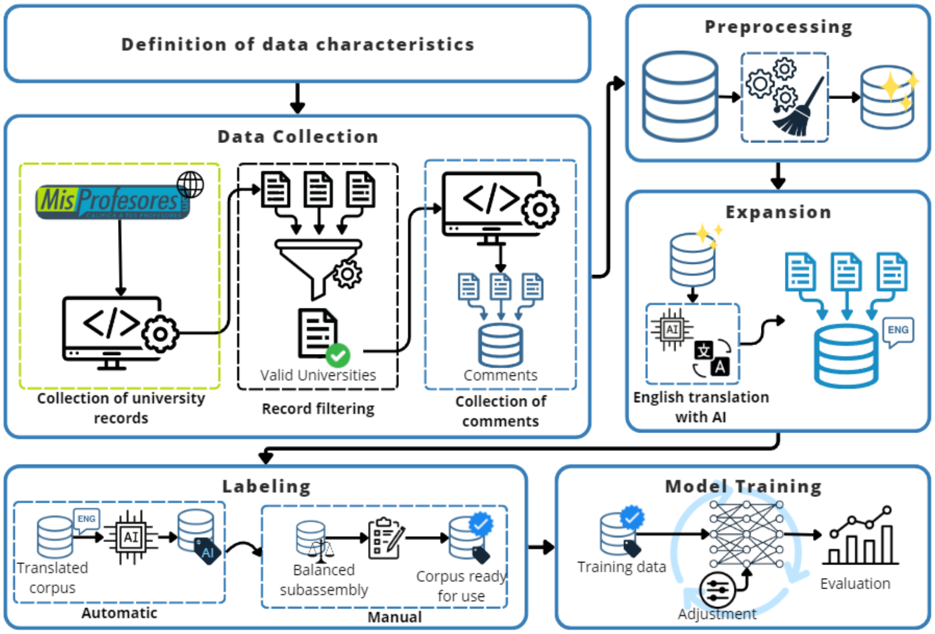
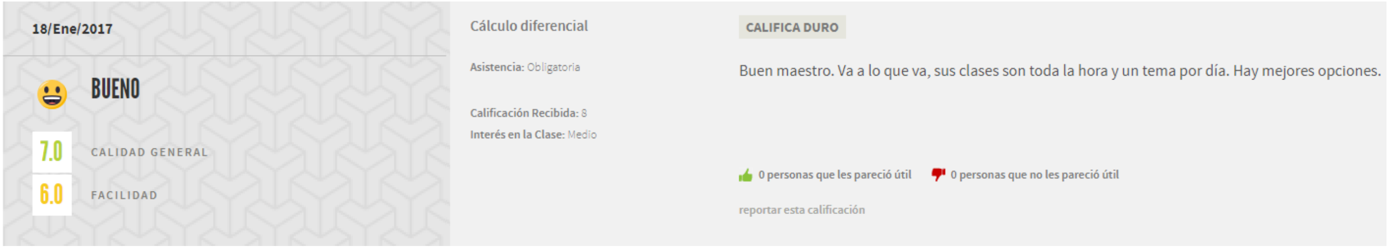


Fig. 1. Representative diagram for the methodology.

3.1 Definition of data characteristics

As can be observed in Fig. 1, the first step in the construction of the corpus is to define the data features to be extracted. A geographical scale at national level was established as the most appropriate, given that it is bounded by the scope of MisProfesores platform. Each of the country's states were selected as focus points for data extraction from that platform. This assured that Mexico's linguistic diversity was represented.

Subsequently, the characteristics of the academic institutes from which the data were extracted were defined. These selected institutions met the following criteria: they must be public, school-based universities with the highest statewide relevance according to the number of registered students enrolled. This was based on the data gathered from ANUEIS 2023 (ANUEIS, 2024). To obtain the data of interest, the reviews submitted by the students to the professors of the previously selected institutions were used. Each review on the platform is composed of a comment/opinion of the student, regarding the subject, the grade achieved with the teacher in the course, and the date, amongst others (see **Fig. 2**).


18/Ene/2017

😃 BUENO

7.0 CALIDAD GENERAL

6.0 FACILIDAD

Cálculo diferencial

Asistencia: Obligatoria

Calificación Recibida: 8

Interés en la Clase: Medio

CALIFICA DURO

Buen maestro. Va a lo que va, sus clases son toda la hora y un tema por día. Hay mejores opciones.

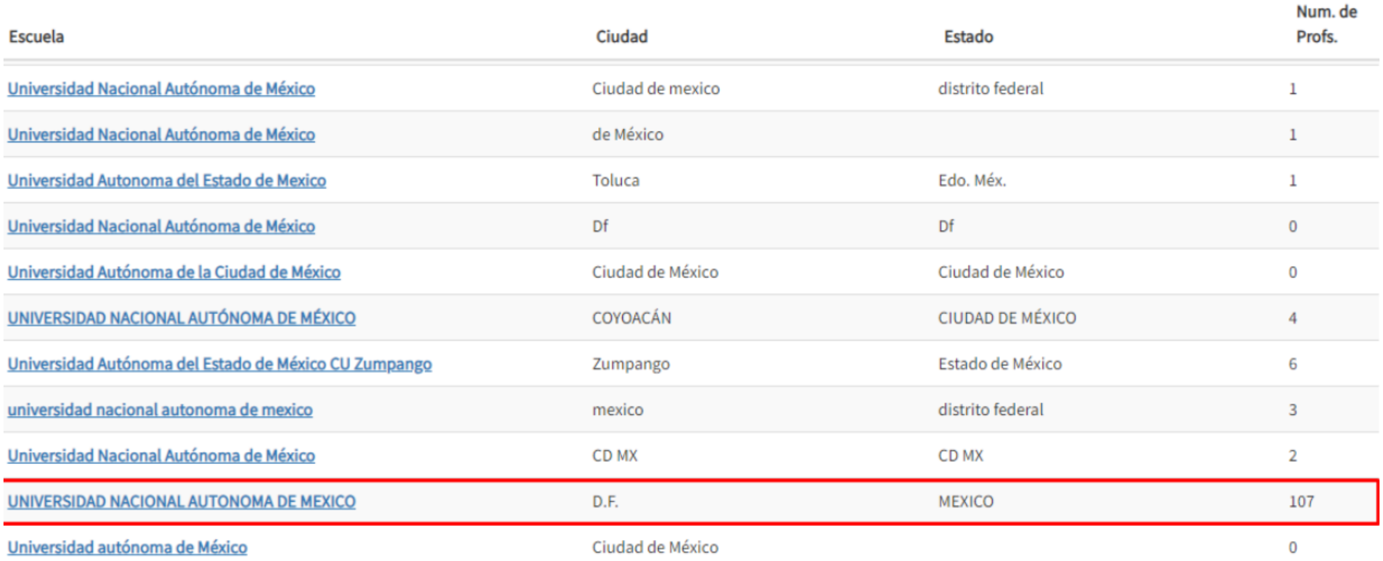
0 personas que les pareció útil   0 personas que no les pareció útil

reportar esta calificación

Fig. 2. Review extracted from *MisProfesores* platform.

3.2 Data Collection

For collecting the data comprising the dataset, we used techniques for extracting text from websites (commonly known as Web Scrapping) and extracted the data from MisProfesores website. For this task, tools such as Selenium and BeautifulSoup were used, both Python libraries used to extract dynamically and statically generated data respectively. Because of the flexibility offered by MisProfesores platform to capture teachers and educational institutions, it was necessary to include a filter in the extraction algorithm. This filter only retrieved records from the most relevant universities in the results of the platform, as shown in **Fig. 3**. The filter selection was based on the number of teacher records associated with each educational institute.



Escuela	Ciudad	Estado	Num. de Profs.
Universidad Nacional Autónoma de México	Ciudad de mexico	distrito federal	1
Universidad Nacional Autónoma de México	de México		1
Universidad Autonoma del Estado de Mexico	Toluca	Edo. Méx.	1
Universidad Nacional Autónoma de México	Df	Df	0
Universidad Autónoma de la Ciudad de México	Ciudad de México	Ciudad de México	0
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO	COYOACÁN	CIUDAD DE MÉXICO	4
Universidad Autónoma del Estado de México CU Zumpango	Zumpango	Estado de México	6
universidad nacional autonoma de mexico	mexico	distrito federal	3
Universidad Nacional Autónoma de México	CD MX	CD MX	2
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO	D.F.	MEXICO	107
Universidad autónoma de México	Ciudad de México		0

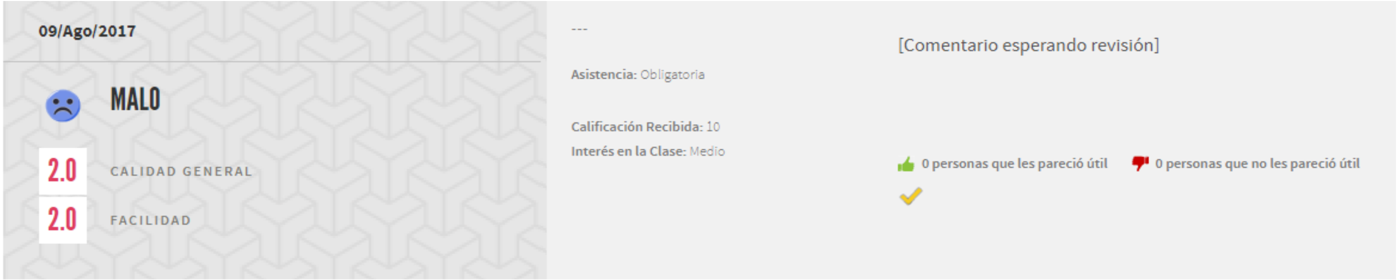
Fig. 3. Search results for "Universidad Nacional Autónoma de México" highlighted in red the most relevant record based on the number of registered professors.

3.3 Data Preprocessing

For each review commentary, the retrieved text was run through a process where unnecessary punctuation marks were removed. For example, exclamation marks or question marks repeated at the beginning or end of a sentence. With this approach, the noise found within the corpus was significantly reduced. Reviews with a comment that met any of the following criteria were also discarded:

- Comments with only blanks or completely empty spaces.
- Comments composed only of special characters and/or numbers.
- Comments pending for revision or blocked by MisProfesores platform itself.

An example of a review with an invalid comment is given in **Fig. 4**, where the comment is awaiting review from the platform. After this data filtering procedure, a corpus consisting of 18,000 samples was obtained, where each sample contains the data previously mentioned in section 3.1.



09/Ago/2017

MALO

2.0 CALIDAD GENERAL

2.0 FACILIDAD

Asistencia: Obligatoria

Calificación Recibida: 10

Interés en la Clase: Medio

[Comentario esperando revisión]

0 personas que les pareció útil
0 personas que no les pareció útil

Fig. 4. Sample review with an invalid comment.

3.4 Expansion of the data set

For the enrichment of the dataset, a translation from Spanish to English was applied for each comment included in the corpus. Given the number of samples in the corpus, the translation was performed automatically using Transformers-based models. As a result, an English version of the corpus was created, which expanded the corpus use cases. This English version served the purpose of automatically labelling the corpus.

3.5 Labeling of the data set

The process of labeling the dataset, given the resulting size of the collected corpus, required the application of automatic labeling methods to the English version of the comments, using a variety of classifier transformer-based models trained to analyze sentiment and to classify toxicity in plain text. Given the nature of this approach, it was feasible to get preliminary results and to assess the state of the balance of the data. Based on this first automatic labeling process, a balanced subset containing toxic and non-toxic samples was extracted. After this subset was collected, a manual labeling process was performed by a group formed by graduate students and professors of the *Instituto Tecnológico de Culiacán*.

We set up a series of criteria to make the manual labeling process. For example, the team was instructed to carefully read each comment to identify possible sarcastic phrases. Furthermore, the importance of not basing the labeling of the comment exclusively on swear words, but to analyze and evaluate the context in which these words were used, was stressed. As a result, we obtained a manually labeled corpus with 853 samples. Each sample contains a text in Spanish in the commentary of the original review and a binary label. In this binary label, the number “1” represents that the sample text is toxic and the number “0” indicates the absence of toxicity in the text.

3.6 Model training for text classification

For the process involving the development of a model able to classify comments on MisProfesores platform as toxic or non-toxic, a subset of 853 manually labeled samples were used for training. Given the focus of the study in the Spanish language, it was opted to use machine learning (ML) models such as support vector machines (SVM) and deep learning (DL) such as neural networks models with LSTM architecture and Transformer-based models.

All these models were trained establishing 10 epochs, a batch size of 16, and a weight decay of 0.01. By defining such hyper-parameters, it ensured fair training between models. The training and evaluation processes were performed in cloud-based environments, using the same hardware device for every tested model.

4 Results and discussion

Due to the complexity of the analysis, this section is presented in two parts. In the first one, state-of-the-art models for the task of binary classification of toxicity in texts are presented and compared, providing a detailed overview of their performance. The second part focusses on the comparison among different model architectures using MisProfesores corpus, evaluating their performance to identify toxicity patterns. In both sections, a brief discussion of the results of each comparison is included, highlighting the advantages and constraints of the approaches addressed.

4.1 Comparison with State-of-the-Art Models in Binary Toxicity Classification

After the model selection, the model was compared with other state-of-the-art models in binary toxicity classification. The first, an XLM-RoBERTa model trained with a corpus presented in (TextDetox, 2024) for binary toxicity classification task, is a compilation of different corpora in several languages, including the Spanish language. The second one, referred to as dehateBERT (Aluru et al., 2020), is a multilanguage model based on BERT that was trained with a Spanish toxicity corpus. The models were evaluated with 20% of the corpus that had not been used as part of the training. The results of the evaluation of each model are presented in **Table 1**.

Table 1. Results from different text toxicity classification models

Model	Accuracy	Recall	F1 score
BETO-MP	**0.9649**	**0.9649**	**0.9645**
TextDetox XLMR	0.7894	0.7894	0.7942
dehateBERT	0.6783	0.6783	0.6031

In the comparison of the models used for classifying toxic comments in MisProfesores platform, BETO-MP stands out for its outstanding performance. This model managed to reach a value of 0.9649 in accuracy and recall metrics, and a value of 0.9645 in F1 score, consolidating itself as the most effective option for this task. These high-performance results prove its capability to handle the Spanish language in an efficient manner, which makes it the best option among the assessed models.

In contrast, the TextDetox XLMR model achieved an accuracy of 0.7894, a recall of 0.7894, and an F1 score of 0.7942. Although its performance is poor compared to BETO-MP, it is still a competitive model and may be considered a feasible alternative in given scenarios. However, it fails to capture the complexities of the Spanish language in the same way as BETO-MP.

At last, the dehateBERT model had the lowest performance, with an accuracy of 0.6783, a recall of 0.6783 and an F1 score of 0.6031. This could be related to its multilingual approach which, although it includes Spanish, does not capture the language-specific features as effectively as BETO-MP does. In summary, dehateBERT ranked behind the other models, especially in comparison to the superior performance of BETO-MP.

Overall, the results show that BETO-MP is the most suitable model for the classification of toxic comments in Spanish on MisProfesores.com platform, outperforming both traditional machine learning models and other advanced deep learning models. Its outstanding performance is driven by its dedicated pre-training in Spanish, which allows it to better capture the peculiarities of the language compared to multilingual approaches such as TextDetox XLMR and dehateBERT.

4.2 Comparison between different model architectures using MisProfesores Corpus

Using the previously discussed subset of comments and their corresponding labels, different ML and DL models were trained. The ML models used were bag-of-words (BoW) and a combination of BoW with text representation, both models were built using the EvoMSA library (Graff et al., 2020). Concerning the DL models, a basic LSTM network, known for its performance in natural language processing tasks, was trained. Also, three models based on Transformers were trained. The first one was mBERT, which is a model based on BERT (Devlin et al., 2019), pre-trained with a large corpus in different languages. The second model was XLM-RoBERTa, a variation model of RoBERTa (Conneau et al., 2020) pretrained on a multilingual corpus. The third model used was BETO (Cañete et al., 2020), which, while it shares its architecture with BERT, was pre-trained on an exclusively Spanish corpus.

As **Table 2** shows, the best model for this task was the BETO-based model, achieving outstanding results in each metric, so this model was taken as a reference for the following comparisons. Therefore, for practical purposes, we will refer to the selected BETO model as BETO-MP.

When comparing different models, significant variations in their performance were observed. The EvoMSA BoW model reached an accuracy of 0.8479, a recall of 0.8212, and an F1 score of 0.8067. Although its results are good, they are below EvoMSA BoW + Text Representation and DL-based models. When a supplementary text representation was added to the EvoMSA BoW model, the metrics were slightly improved, achieving an accuracy of 0.8596, a recall of 0.8522, and an F1 score of 0.8262, proving that the additional text representation provided a significant improvement in the metric values.

Table 2. Comparison between different model architectures using the MisProfesores Corpus

Model	Accuracy	Recall	F1 score
EvoMSA BoW	0.8479	0.8212	0.8067
EvoMSA BoW + Text Representation	0.8596	0.8522	0.8262
LSTM	0.7134	0.6714	0.6761
mBERT base	0.8011	0.8011	0.8027
XLM RoBERTa base	0.8245	0.8245	0.8233
BETO-MP	**0.9649**	**0.9649**	**0.9645**

On the other hand, the basic LSTM model performed lower, with an accuracy of 0.7134, a recall of 0.6714, and an F1 score of 0.6761. This can be attributed to the complexity of the comments and the limitation of the model to capture more extensive contextual information, which affected its performance on this specific task. Also, the mBERT-based model, trained on a multilingual corpus, achieved an accuracy of 0.8011, a recall of 0.8011, and an F1 score of 0.8027. Despite being a robust model, its performance was inferior compared to more specialized models in the target language. In comparison, the base RoBERTa XLM model outperformed mBERT with an accuracy of 0.8245, a recall of 0.8245, and an F1 score of 0.8233, although it was still outperformed by other better optimized models for Spanish comment analysis. Finally, the BETO-MP model demonstrated superior performance in terms of accuracy, recall, and F1 score (0.9649, 0.9649 and 0.9645, respectively). This established it as the best model for this task, presumably because of its pre-training on a Spanish corpus, which made it suitable for the fine-grained analysis of Spanish-language comments.

In summary, the BETO-MP model proved to be the most effective for the task, outperforming both deep learning-based models and those pre-trained in multilingual corpora, thanks to its focus on the Spanish language. The results obtained are even superior to the one presented in (Roy et al., 2021) where it is mentioned that a value of 90.29 is obtained in the Macro F1 score metric in the task of hate speech detection in the English language.

5 Conclusions

This study represents a significant advance in the construction of Spanish corpora for sentiment analysis, especially in the context of comments on MisProfesores platform. Through a detailed process that included data collection, preprocessing, cleaning, and labeling, a unique corpus was developed that describes the diversity and complexity of the Spanish spoken in Mexico. This corpus is relevant not only for its size, consisting of 18,000 unlabeled samples and 853 manually labeled samples (both versions of the corpus are available at https://catalabs.mx/datasets/misprofesores/), but also for its focus on capturing the linguistic and cultural richness specific to this particular context.

The results obtained with our model highlight the importance of building a corpus for the specific task of Spanish toxicity detection to improve accuracy in detecting toxic comments on academic social platforms. Our Transformer-based model obtained values of 0.9649 and 0.9645 in the accuracy and F1 score metrics, outperforming other state-of-the-art models built with traditional machine and deep learning techniques. These results not only contribute to the academic field of sentiment analysis and AI, but also offer practical applications for online educational platforms, aiding in the development of safer and more positive learning environments. By proactively detecting and managing hate speech, a more respectful and constructive exchange of ideas, crucial for educational and social development, can be encouraged. Finally, this study highlights the ongoing need to continue to expand and refine the non-English corpora, adapting them to specific contexts to improve the effectiveness of sentiment analysis models.

As future work, it is proposed to increase the number of manually labeled samples and explore other platforms and contexts, with the target of developing more robust and versatile models. Building these resources not only benefits academic research, but also has a direct impact on society, promoting more inclusive and respectful digital environments.

References

Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2021). A Deep Dive into Multilingual Hate Speech Classification. *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020*, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V, 423–439.

ANUIES. (2024). Asociación Nacional de Universidades e Instituciones de Educación Superior. Retrieved March 1, 2024, from http://www.anuies.mx/informacion-y-servicios/informacion-estadistica-de-educacion-superior/anuario-estadistico-de-educacion-superior

Arceo-Gomez, E. O., & Campos-Vazquez, R. M. (2019). Gender stereotypes: The case of MisProfesores. com in Mexico. *Economics of Education Review*, 72, 55-65

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish Pre-trained BERT Model and Evaluation Data. *PML4DC at ICLR 2020*. http://arxiv.org/abs/2308.02976

COMIA (2024). XVI Congreso Mexicano de Inteligencia Artificial. Retrieved June 7, 2024, from https://smia.mx/comia/2024

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Dementieva, D., Moskovskiy, D., Babakov, N., Ayele, A. A., Rizwan, N., Schneider, F., Wang, X., Yimam, S. M., Ustalov, D., Stakovskii, E., Smirnova, A., Elnagar, A., Mukherjee, A., & Panchenko, A. (2024). Overview of the Multilingual Text Detoxification Task at PAN 2024. https://dardem.github.io

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Fan, H., Du, W., Dahou, A., Ewees, A. A., Yousri, D., Elaziz, M. A., Elsheikh, A. H., Abualigah, L., & Al-qaness, M. A. A. (2021). Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. *Electronics, 10(11)*. https://doi.org/10.3390/electronics10111332

Graff, M., Miranda-Jimenez, S., Tellez, E. S., & Moctezuma, D. (2020). EvoMSA: A Mul-tilingual Evolutionary Approach for Sentiment Analysis [Application Notes]. *IEEE Computational Intelligence Magazine, 15(1), 76–88*. https://doi.org/10.1109/MCI.2019.2954668.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. *3309–3326.* https://doi.org/10.18653/v1/2022.acl-long.234

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2020). The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments. *Corpus Pragmatics, 4.* https://doi.org/10.1007/s41701-019-00065-w

Mandl, T., Modha, S., Shahi, G. K., Kumar Jaiswal, A., Nandini, D., Patel, D., & Schäfer, J. (2020). Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages under Creative Commons License Attribution 4.0 International (CC BY 4.0). http://ceur-ws.org

MisProfesores.com. (n.d.). Retrieved March 1, 2024, from https://www.misprofesores.com

Nabiilah, G. Z., Prasetyo, S. Y., Izdihar, Z. N., & Girsang, A. S. (2023). BERT base model for toxic comment analysis on Indonesian social media. *Procedia Computer Science, 216, 714–721.* https://doi.org/10.1016/j.procs.2022.12.188

Roy, S. G., Narayan, U., Raha, T., Abid, Z., & Varma, V. (2021). Leveraging multilingual transformers for hate speech detection. arXiv preprint arXiv:2101.03207.

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *In Applied Sciences (Switzerland) (Vol. 13, Issue 7). MDPI.* https://doi.org/10.3390/app13074550

Wankhade, M., Rao, A., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55, 1–50.* https://doi.org/10.1007/s10462-022-10144-1