Transformer based Offensive Language Identification in Spanish[*]

Sreelakshmi K[1], Premjith B[1], and K. P. Soman[1]

Center for Computational Engineering and Networking (CEN),
Amrita School of Engineering, Coimbatore,
Amrita Vishwa Vidyapeetham, India
k_sreelakshmi@cb.students.amrita.edu, b_premjith@cb.amrita.edu,
kp_soman@amrita.edu

Abstract. This paper presents the work done for the shared task on MeOffendEs@IberLEF 2021 Non-contextual binary classification for Mexican Spanish. We implemented two deep neural network architectures such as a network containing a Bi-LSTM, LSTM, fully connected layer and another with a Bi-LSTM and LSTM stack. In addition to that we also implemented a BERT classifier. Among the three models the BERT exhibited better training performance, and we submitted the predictions based on the same. BERT performed well compared to other languages as it has pretrained embeddings that are trained on huge corpus of multiple languages.

Keywords: Long short-term memory · Bidirectional Long Short-Term Memory · Bidirectional Encoder Representations from Transformers

1 Introduction

Social media platforms such as Facebook, Twitter and YouTube hearten interpersonal communication by keeping these platforms open and free of cost. This has made people to interact, post messages and comments and express their views online. Unfortunately, it is often used as a means to attack or offend people leading to unfolding of hateful and offensive content resulting in cyber violence. Offensive content is non-verbal or oral communication expressing disparity against a group or person based on their religion, age, sexual orientation, race, gender, nationality, and ethnicity [5], [2] [4].

Putting an end to usage and spread of offensive content is peremptory for all the social media platforms and content moderation establishments. At present most of the moderations are limited to the community platforms that reckon

[*] *IberLEF 2021, September 2021, Málaga, Spain.*

on repeatedly used words and block-lists or human moderators. Furthermore, these are not reliable options for all the platforms due to their sheer cost and long-drawn-out process [11]. Considerable amount of work was done to identify offensive content in English,but little research has been done Non-English languages like Spanish [3], [14], [15], [13], [16, ?] [18].

MeOffendEs@IberLEF 2021 focuses on offensive language analysis in social networks for Spanish. MeOffendEs@IberLEF 2021 has four subtasks.

- Subtask 1: Non-contextual multiclass classification for generic Spanish
- Subtask 2: Contextual multiclass
- Subtask 3: Non-contextual binary classification for Mexican Spanish
- Subtask 4: Contextual binary classification for Mexican Spanish

We participated in the Subtask 3, were we employed three different models, one using BERT and the other two deep learning models like a Hybrid model using a Bi-LSTM, LSTM stack and a model with Bi-LSTM, LSTM stack followed by a fully connected layer. The BERT based model gave the highest precision on the test data.

2 Literature Review

For the past few years, one of the major concern for social media platforms and users is offensive messages that tarnish an individual or a group. Various abusive and offensive language identification problems and shared tasks have been explored in the literature ranging from aggression to cyberbullying, hate speech, toxic comments, and offensive language but there is very few work done in Spanish language though it being the fourth most spoken language [6].

Due to the high amount of offensive content that spread through social media, oodles of academic events and shared tasks on offensive and hate speech detection have taken place. Few of them are the first, second and third Workshop on Abusive Language [12], SemEval 2020 [21] on offensive language identification from multilingual languages (OffensEval) like English, Arabic, Danish, Greek, Turkish, FIRE 2019 on offensive language identification from Indo-European languages [7], various editions of GermEval Shared Task on the Identification of Offensive Language [17]. This shared task deals with the classification of German tweets to binary classes (OFFENSE and OTHER), fine-grained multi-classes (OFFENSE, OTHER, PROFANITY, INSULT) and classification of offensive tweets as explicit or implicit.

SemEval 2019 [20] conducted a task on offensive and non-offensive comments detection from English tweets. The dataset (OLID) consists of 13240 tweets for training and 860 tweets for testing. Deep learning models like Convolutional Neural Networks, Bidirectional Encoder Representations from Transformers (BERT), Long Short Term Memory (LSTM), LSTM with attention, Embeddings from Language Models (ELMo), basic machine learning models were a part of the assorted models used.

IberLEF 2019, had a task on aggressiveness detection from Mexican Spanish tweets. Teams used various features such as Bag of Words with TF-IDF weights, hierarchical features obtained with CNN, Statistical descriptors, Document frequency, mutual information, and lexical Availability, Linguistic features and different types of n-grams and classifiers such as CNN, LSTM and Multilayer Perceptron, GRU and machine learning models like SVM, Naïve Bayes, logistic regression [1].

Study of various Deep Learning methods with recently pre-trained language models based on Transfer Learning and basic machine learning models have been done in this line of research. BERT, XLM and BETO have given promising results in Spanish hate speech detection [10].

3 Dataset Description

Dataset Description

The shared task [9], [8] has 4 subtasks namely,

- Non-contextual multiclass classification for generic Spanish
- Contextual multiclass classification for generic Spanish
- Non-contextual binary classification for Mexican Spanish
- Contextual binary classification for Mexican Spanish

Among the four we participated in Non-contextual binary classification for Mexican Spanish. The dataset statistics and annotation is as given in Table 1.

Table 1. Details of the class labels available in the dataset and the data statistics.

Dataset Name	Labels	Train set	Valid set	Test set
Non-contextual binary classification for Mexican Spanish	0 - Non-offensive and 1 - offensive	5060	76	2183

4 System Description and Results

This section gives the details of the models used to experiment on the data. In this work we have experimented using a BERT model and deep learning models to identify the offensive texts.

4.1 Preprocessing

The dataset comprises social media texts and hence includes user names, hashtags, and URLs. Since these entities do not contribute much to the classification task, we employed a preprocessing step to replace the user names with word 'USERNAME', hashtags with the word 'HASH' and URLs with the word 'URL' and also removed the punctuations and symbol like ' % & ∗ − < > : /) (' from the text.

4.2 Models

We experimented with deep learning models and a BERT model for classifying the social media text into different categories. The BERT model gave the highest result on the test set.

Deep Learning We conducted the experiments to classify the social media text into offensive and non-offensive using two deep learning models. The two models are:

- Model 1: Hybrid BiLSTM, LSTM stack followed by a fullyconnected layer dedicated for classification
- Model 2: Hybrid BiLSTM, LSTM, dense layer stack followed by a fullyconnected layer dedicated for classification

The preprocessed text undergoes few more steps before the classification.

- Tokenization: An "<OOV>" token is used to mark the Out-of-Vocabulary words
- Padding: To maintain equal length sentences are padded with zeros.

The hyperparameters used for both the models are given in Table 2

Table 2. Set of hyperparameters used in building the model.

Hyperparameter	Value
No. of neurons in the LSTM	128
dropout in LSTM	0.2
recurrent dropout in LSTM	0.2
Activation function at LSTM layer	ReLU
No. of neurons in the Bi-LSTM	128
dropout in Bi-LSTM	0.2
recurrent dropout in Bi-LSTM	0.2
Activation function at Bi-LSTM layer	ReLU
Number of neurons in dense layer	128
Activation function at theoutput layer	Sigmoid
Loss	Binary crossentropy
Optimizer	Adam
Learning rate batch size	128
epochs	10

The results obtained on the test data for both the models are as given in Table 3

Table 3. Performance of the model over the test data.

Model	Precision	Recall	F1 Score
Model 1	0.783333	0.339350	0.473552
Model 2	0.810000	0.348887	0.487707

BERT We made use of 12-layer BERT model (”bert-base-multilingual-cased”) for classification [19]. The ”bert-base-multilingual-cased” pretrained model was used and fine tuned over our data. The major steps involved in the experiment are as follows:

- Tokenise the sentences
- Add special tokens [CLS] and [SEP]
- Map the tokens to their IDs
- Pad and truncate the sentences to the same length
- Creating the attention masks

The finetuned hyperparameters are given in Table 4

Table 4. Set of hyperparameters used in building the model.

Hyperparameter	Value
tokenizer	bert-base-multilingual-cased
max truncate length	200
batch size	32
learning rate	2e-5
epochs	10

The performance of the model is validated using precision, recall and F1 Score. Table 5 gives the details of the model perfromance on the test data.

Table 5. Performance of the model over the test data.

Dataset	Precision	Recall	F1 Score
Non-contextual binary classification for Mexican Spanish	0.918333	0.314318	0.468338

The BERT based model gave the highest precision among the three models experimented and hence the prediction from the BERT model was submitted.

5 Conclusion

This paper presents the submission to the shared task MeOffendEs@IberLEF 2021 on Offensive Language Identification from Mexican Spanish text. Two Deep

Learning models, such as a hybrid network with a LSTM layer, a Bi-LSTM layer, a network consisting of a LSTM layer, a Bi-LSTM layer and a fully connected network were implemented. A transformer based BERT model was also implemented. The BERT gave the highest result of 91% precision for Non-contextual binary classification from Mexican Spanish text.

References

1. Mario Ezra Aragón, Miguel Angel Alvarez Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor Pineda, and Daniela Moctezuma. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *IberLEF@ SEPLN*, pages 478–494, 2019.
2. Andrew Arsht and Daniel Etcovitch. The human cost of online content moderation. *Harvard Law Review Online, Harvard University, Cambridge, MA, USA. Retrieved from https://jolt. law. harvard. edu/digest/the-human-cost-ofonline-content-moderation*, 2018.
3. Bharathi Raja Chakravarthi, Mihael Arcan, and John Philip McCrae. Improving wordnets for under-resourced languages using machine translation. In *Proceedings of the 9th Global WordNet Conference*, pages 77–86, 2018.
4. Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019.
5. Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33, 2018.
6. Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis. Understanding the predictive power of social media. *Internet Research*, 2013.
7. Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17, 2019.
8. Manuel Montes, Paolo Rosso, Julio Gonzalo, Ezra Aragón, Rodrigo Agerri, Miguel Ángel Álvarez-Carmona, Elena Álvarez Mellado, Jorge Carrillo-de Albornoz, Luis Chiruzzo, Larissa Freitas, Helena Gómez Adorno, Yoan Gutiérrez, Salud María Jiménez-Zafra, Salvador Lima, Flor Miriam Plaza-de Arco, and Mariona Taulé, editors. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021).* 2021.
9. Flor Miriam Plaza-del-Arco, Marco Casavantes, Hugo Jair Escalante, M. Teresa Martin-Valdivia, Arturo Montejo-Ráez, Manuel Montes-y-Gómez, Horacio Jarquín-Vásquez, and Luis Villaseñor-Pineda. Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. *Procesamiento del Lenguaje Natural*, 67(0), 2021.
10. Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120, 2021.
11. Manikandan Ravikiran and Subbiah Annamalai. Dosa: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–17, 2021.

12. Sarah T Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem. Proceedings of the third workshop on abusive language online. In *Proceedings of the Third Workshop on Abusive Language Online*, 2019.
13. T Tulasi Sasidhar, B Premjith, and KP Soman. Emotion detection in hinglish (hindi+ english) code-mixed social media text. *Procedia Computer Science*, 171:1346–1352, 2020.
14. K Sreelakshmi, B Premjith, and Soman Kp. Amrita_cen_nlp@ dravidianlangtech-eacl2021: Deep learning-based offensive language identification in malayalam, tamil and kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 249–254, 2021.
15. K Sreelakshmi, B Premjith, and KP Soman. Detection of hate speech text in hindi-english code-mixed data. *Procedia Computer Science*, 171:737–744, 2020.
16. Soman K P Sreelakshmi K, Premjith B. Amrita cen at hasoc 2019: Hate speech detection in roman and devanagiri scripted text. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17, 2019.
17. Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, Manfred Klenner, et al. Overview of germeval task 2, 2019 shared task on the identification of offensive language. 2019.
18. Sreelakshmi K Soman K.P. Tulasi Sasidhar T, Premjith B. Sentiment analysis on hindi–english code-mixed social media text. volume 171, 2017.
19. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
20. Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*, 2019.
21. Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*, 2020.