# Hate Speech Detection using A Multilingual transformer Model

Saravanamuthu Muthusamy
Department of Computer Science
Saint Louis Univeristy
20N, Lindell Blvd, Saint Louis, Missouri,USA

***Abstract*— Social media today has made it possible for the spread of harmful and unfriendly content to expand exponentially, increasing the number of individuals who are exposed to it. As a solution to this issue, a lot of people in the natural language processing community have recently shown an interest in the automatic detection of harmful content like hate speech, provocative language, and abusive language. This work employs machine learning and multilingual transformer models to categorize comments spoken in Spanish as offensive or not offensive.**

***Index Terms*— Transformers, Multilingual model**

## I. INTRODUCTION

Social media has become one of the most important modes of expressing ideas. Over the past 15 years, the use of social media has increased dramatically. Policymakers, educators, parents, and doctors are interested in and concerned about the effects of social media on our lives and psychological well-being to the increase in usage of Facebook, Instagram, Twitter, Snapchat, and other social media platforms, as well as the time spent on them.

Since the last few years, hate speech has dramatically escalated. In fact, users of various ages, cultures, and hobbies are rapidly using social media sites. These social media platforms offer a free forum for users to express their opinions and share or transmit their ideas throughout the globe, but the sheer volume of posts and communications exchanged makes maintaining content control nearly difficult. .

In the past, social networking organizations developed a variety of tools to identify content that would be harmful or deceptive by their standards. It can contain nudity, pornography, hate speech, and misinformation or it can be just plain text. However, it is complicated and impossible to train a machine learning model to automatically distinguish between objectionable content and humor, nudity-themed art, and political satire. Because new topics and jargon are frequently added to our issue of concern—hate speech—the process gets more difficult and calls for ongoing training and the addition of new research.

In this paper, the focus of the work will be on Spanish tweets. These days there is an increase in the spread of hate speeches through tweets and these tweets should be screened and removed which needs a model to detect hate speech tweets. Detecting hate speech in a different language is a challenging task because every day there is a new annotation created by social media users to put down some groups or an individual which may not be offensive but they must combine these two modalities to understand the actual message conveyed in the tweet. Other reasons are text is ambiguous which can mislead the model

## II. LITERATURE SURVEY

### A. Background

Natural Language Processing is a branch of artificial intelligence that gives the computers ability to understand the text and spoken words in a way humans can. Natural Language Processing blends statistical, machine learning, and deep learning models with computational linguistics—rule-based modeling of human language. With the use of these technologies, computers are now able to process human language in the form of text or audio data and fully "understand" what is being said or written, including the speaker's or writer's intentions and sentiments.[1]

Most opinion mining and sentiment analysis methods in the English Language rely on sentiment lexicons due to their large size and accuracy over corpus-based data sets. Corpus-based methods involve supervised learning by applying known classifiers.[2] The problem we took can be considered a sentiment analysis problem on Spanish lexicons. We use a multilingual transformer, which can be used to for other monolingual dialects.

### B. Literature Survey

It has been studied for many different sorts of texts, including product reviews, movie reviews, blogs, and social media. Classification and sentiment analysis of texts are well-known issues. The polarity lexicon-based approach, which collects information on the text sentiment based on the polarity of its words, is possibly the simplest and most intuitive method of sentiment analysis. The issue with this strategy is often minimal coverage, as some texts may only contain a few or no terms from the polarity lexicon, which is problematic for brief messages in particular. Additionally, techniques based on polarity lexicons fail to account for the context in which words are employed.

Because they make use of the contextual data from the target corpora, word embedding-based techniques offer higher coverage. [3]

Some work has been done on the Multilingual task. The author used a deep learning architecture to predict sentiments that were made up of several layers of LSTMs. K. Yadav and their team [4] proposed an ensemble approach based on combining SVM, Linear Regression, However, many researchers have concluded that they cannot predict hate speech and abusive speech easily because social media users keep on creating new terms to abuse. Using natural language processing (NLP), the negative and positive polarities in tweets were found to be insufficient. Due to the introduction of new terms, we need to train the model periodically. Because of these challenges it's better to use a universal model like transformers which can perform one or more task[5]

Our problem can be considered as the sub-task of sentiment analysis task. Sentiment analysis is the process of predicting sentiment information of the specific input sequence. For this task, the labels are mostly created by the human annotator. To do edge-cutting research we need a large number of dataset.

[7]

## III. PROPOSED MODEL

### A. Transformers

NLP's Transformer is a new architecture that aims to solve sequence-to-sequence tasks, while easily handling long-distance dependencies. It relies entirely on self-attention, computing the input and output representations without using sequence-aligned RNNs or convolutions.

Bidirectional Encoder Representations from Transformers (BERT) is the acronym. It is aimed to pre-train deep bidirectional representations from the unlabeled text by conditioning on both the left and right context. When applied to NLP, the BERT model can be refined with just one additional output layer, resulting in models that are among the most advanced in the industry. BERT is built on Transformers, a deep learning model in which each output element is connected to each input element, and the weightings between them are dynamically computed based on their connection. BERT is unique in that it can retweets negative and positive polarities any other device. The term "bi-directionality" refers to this new ability made possible by the invention of Transformers.

Classification tasks such as sentiment analysis are done similarly to Next Sentence classification, by adding a classification layer on top of the Transformer output for the [CLS] token.

### B. Model implementation

I use the Bert model to train the Spanish tweets data sets. Data cleaning is done by removing unwanted characters, removing punctuation. The data set consists of positive and negative labels. All the unwanted characters and emoji are removed by using the regular expression. The auto tokenizer from the transformer library is responsible for all the preprocessing. The pre-trained model can be called directly on a singletring or a list. After data cleaning the data set is divided in two, train data and test data. The Multilingual BERT model is trained on train data set and test data is used for validation.

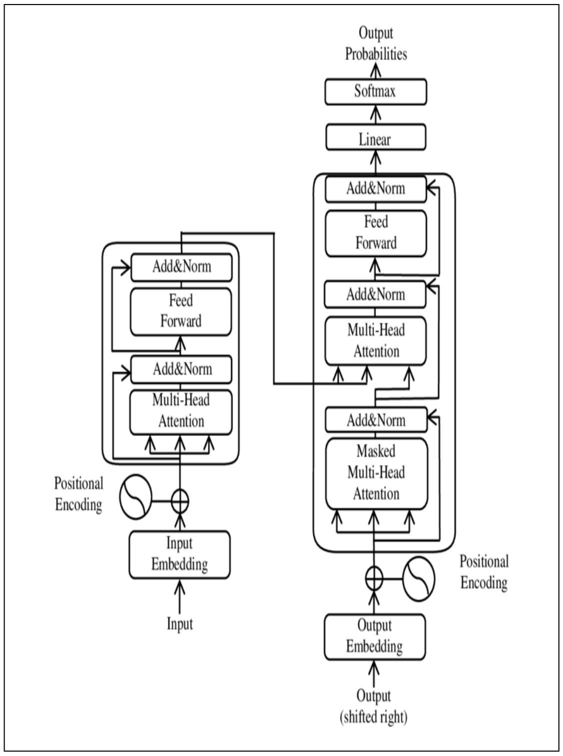

Fig. 1. Transfomer Architecture[6]

In the tokenization process, the input text is divided into a list of tokens that are present in the vocabulary.BERT employs a method known as BPE based WordPiece tokenization to cope with terms that are not included in the lexicon. With this method, a term that is not in the vocabulary is gradually broken up into smaller words, which are subsequently used to represent the word as a whole. The context of the word is just the combination of the context of the subwords. because the subwords are a part of the lexicon and we have learned representations for them.

## IV. WORK DONE AND RESULT ANALYSIS

To train the model we have used only 5000 lines of tweets, 1000 lines of tweets for validation. Though our data set is very small, still, the model has produced 81 percent accuracy on just 5 epochs. On a task like this, it is very important to have a big data set. Attached is the confusion matrix of the result(fig 2). If we train the model with more epoch, the model becomes overfit. So in order to achieve good results, it's better to increase the size of the data set.

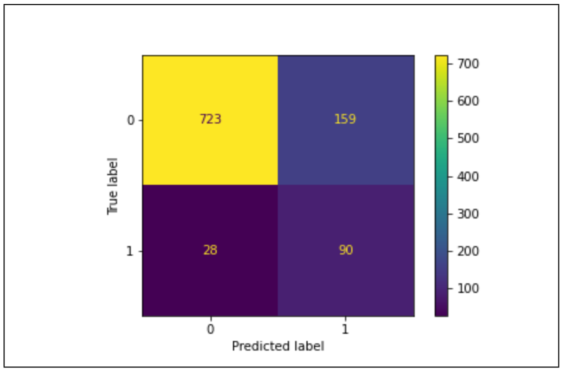

Fig. 2. Confusion Matrix Analysis

## V. CONCLUSIONS

On the topic of hate speech in social media, this study presented experimental work and the results of the task to detect hate speech content in a code-mixed data set of Spanish in order to address the issue of offensive language in social media. We have also presented some of the related work done in the past. Due to the corpus limitation, we have not produced extensive final output. An interesting direction of work is to train multilingual models on data where script representation is more balanced.[7]To perform classification accurately we need more data sets based on our topic. In the future, we have to perform Data extraction from social media to have fully defined. The future of Natural Language Processing is Multilingual models, so it is important to keep on implementing research and ideas in these type of models.

## ACKNOWLEDGMENT

I would like to thank Dr. Kevin Scannell for his valuable lectures on Natural Language Processing concepts. His teachings made develop a keen interest on Natural Language processing.

## REFERENCES

[1] Natural Language Processing [online]. Available: https://www.ibm.com/cloud/learn/natural-language-processing. [Accessed 7 Jan. 2018].

[2] K. S. Sabra, R. N. Zantout, M. A. E. Abed and L. Hamandi, "Sentiment analysis: Arabic sentiment lexicons," 2017 Sensors Networks Smart and Emerging Technologies (SENSET), 2017, pp. 1-4, doi: 10.1109/SENSET.2017.8125054.

[3] J. Vankka, A. Vesselkov, H. Myllykoski and O. Kosomaa, "Framework for Analyzing and Visualizing Topics and Sentiments on Social Media: the Case of MH 17 Tweets," 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), 2021, pp. 257-266, doi: 10.1109/ICBDA51983.2021.9403069.

[4] K. Yadav, A. Lamba, D. Gupta, A. Gupta, P. Karmakar, and S. Saini, "Bi-LSTM and Ensemble-based Bilingual Sentiment Analysis for a Code-mixed Hindi-English Social Media Text," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-6, DOI: 10.1109/INDICON49873.2020.9342241.

[5] J. A. Cerón-Guzmán and E. León-Guzmán, "A Sentiment Analysis System of Spanish Tweets and Its Application in Colombia 2014 Presidential Election," 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom), 2016, pp. 250-257, doi: 10.1109/BDCloud-SocialCom-SustainCom.2016.47.

[6] https://en.wikipedia.or/wiki/Transformer_%28machine_learning_model%

[7] Specializing Multilingual Language Models: An Empirical Study Ethan C. Chau, Noah A. Smith

[8] Z. Ke, J. Sheng, Z. Li, W. Silamu and Q. Guo, "Knowledge-Guided Sentiment Analysis Via Learning From Natural Language Explanations," in IEEE Access, vol. 9, pp. 3570-3578, 2021, doi: 10.1109/ACCESS.2020.3048088.