Capstone Proposal

Eduardo Castro Bermúdez

Machine Learning Engineer
Nanodegree

February 24, 2021

Hate Speech Detection

Twitter Mexico - AMLO

Proposal

Domain Background

Social media has overtaken our understanding of human communication. To clarify, 500 million tweets are sent every day by humans… and bots[1]: a massive real-time interaction. This reality has allowed small hate groups to connect with people that think similarly to them, in appearance making their ideas bigger than they really are.

Therefore, social media is being exploited as platforms for bigotry[2]. Political and social figures have exploited this, encouraging hate groups with incendiary rhetoric that stigmatizes and dehumanizes minorities, migrants, refugees, women and any so-called “other”.

Problem Statement

Hate speech represents a challenge that must be treated from both a technical and a social point of view. In order to analyze it and detect it on social media must be measurable. Therefore, one path to follow is to detect offensive language and negative sentiment in messages, like tweets. Once it is detected, it can be measured in ones and zeros (binary): data that could work to train a model.

Datasets and Inputs

The input dataset to train the model consists of 13,586 tweets (in spanish) that mentioned Andres Manuel Lopez Obrador (AMLO), President of Mexico, through hashtags. This dataset is related to the hate speech issue due to the polarization crisis in Mexico since the 2018 Presidential Election[3]. Therefore, the speech in the “political Twitter Mexico” is continuously a confrontation, without debate, between supporters and opponents: a value sample to detect hate speech.

The tweets were collected through the Twitter API using Tweepy, a Python library. To build the dataset, every tweet was cleaned and tokenized. Then, using an offensive words dataset (in

[1] Smith, Kit. *60 Incredible and Interesting Twitter Stats and Statistics*. https://www.brandwatch.com/blog/twitter-stats-and-statistics/. February 2020.

[2] Guterres, António. *Strategy and Plan of Action on Hate Speech, United Nations (UN)*. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf. May 2019.

[3] Domínguez González, Gerardo. *Elección, polarización y hartazgo social en México.* https://nuso.org/articulo/eleccion-polarizacion-y-hartazgo-social-en-mexico/. June 2018.

spanish) from Hatebase[4], all tweets were labeled in order to know if they contained offensive language or not. So, the input data is a text file (.txt) that contains the features: user, publication date (date_time), likes, tweet, tokenized tweet, offensive language.

	user	date_time	likes	tweet	tokenize_tweet	offensive
0	MontoyaDian	2021-02-24 05:25:32	27541	De verdad creo que me hace daño hablar, ver, l...	verdad creo hace daño hablar ver leer escuchar...	0
1	ESilmarillion	2021-02-24 04:44:46	11862	#ElPeorPresidenteDeLaHistoria\nUn #burro\n@lop...	peor presidente historia presidente rompa pacto	0
2	Lor_Tanatologa	2021-02-24 03:59:52	29782	#NiUnVotoAMORENA #niunamas #VotoUtil2021 #YaB...	voto voto util basta presidente rompa pacto lo...	0
3	JanetteGongora	2021-02-24 03:29:02	3772	#UnVioladorNoSeraGobernador #PresidenteRompaEl...	violador sera gobernador presidente rompa pact...	0
4	Rparra13	2021-02-24 03:11:45	7441	Y mientras tanto Morena con un candidato presu...	mientras morena candidato presunto violacion l...	0
5	LaGabyOr	2021-02-24 02:06:00	1154	@lopezobrador_ En México, país de feminicidios...	méxico país feminicidios acusado violación deb...	0
6	BlancaJimenezMx	2021-02-24 01:59:02	10707	Al poner de candidato a Felix Salgado, #Morena...	poner candidato felix salgado morena iría prop...	0
7	gabybravo29	2021-02-24 01:52:39	2487	Platicando con papá-\nYo: Pa y tú votarias por...	platicando papá pa votarias guerrero papá noo ...	1
8	josejuandenis	2021-02-24 01:49:17	369	@Tonicanto1 @PabloIglesias tan inteligente que...	tan inteligente robar robando dado cuenta puto...	1
9	NiUnaMenos_Mex	2021-02-24 01:39:11	18	Excelente columna de @olabuenaga #PresidenteRo...	excelente columna presidente rompa pacto	0

In this way, it has been identified that 1569 tweets are offensive, an 11.5 % of the total. Therefore, to add value to the data, this sample can be used to measure the sentimental analysis of every tweet before the model is trained. A tweet that has offensive language and is negative is considered hate speech.

Solution Statement

Detecting hate speech can be solved through a Natural Language Processing (NLP) trained model that classifies messages. The input dataset must be a sample of tweets labeled in 1 (hate speech: negative sentiment and bad/offensive language) or 0 (no hate speech), and contain the message as a string. Once the data is splitted in train and test data, the model can be trained and the classifying results can be measured by comparing it with test data.

Benchmark Model

The research paper **Hate Speech Detection Using Natural Language Processing Techniques**[5] allows to compare the presented solution. In this research, the author applied a Convolutional Neural Network model to classify tweets in: *hate*, *offensive language*, and *neither*.

The model was tested using the accuracy, precision, recall and F-score metrics. Finally, the model results obtain an accuracy of 91%, precision of 91%, recall of 90% and a F-measure of 90%.

Evaluation Metrics

The accuracy, precision and recall metrics will be used to evaluate the model performance. Since the proposal is a classification model, these metrics can measure the false positives and false negatives in the predicted results compared to the test data.

[4] Hatebase Company. https://hatebase.org/. The world's largest structured repository of regionalized, multilingual hate speech.

[5] Shanita Biere. *Hate Speech Detection Using Natural Language Processing Techniques.* https://beta.vu.nl/nl/Images/werkstuk-biere_tcm235-893877.pdf. Vrije Universiteit Amsterdam. August 2018.

The accuracy metric will evaluate the number of correct predictions made by the model, which could represent a global overview of the performance. The precision measures what proportion of tweets that are classified as hate speech, actually are hate speech. While, the recall provides an indication of missed positive predictions: tweets that really are hate speech, but were misclassified.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textbf{Precision} = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Positives}}$$

$$\textbf{Recall} = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}}$$

Project Design

Modern human communication is evolving faster than we are analyzing it, which has allowed small hate groups to connect with people that think similarly to them, making their ideas bigger than they really are, and resulting in polarization and hate speech. In Mexico, these problems are growing since the 2018 Presidential Election, there is a low-level social debate and there are very few midpoints. In this case, Twitter is a rational picture of the problem in social media and a logical field of study.

This issue, even if it is clearly social, is a consequence of the development of the tech industry: the social media. So the problem must be analyzed from both points of view. The first step is to measure it and detect it, a job that can be done through Machine Learning. Then, the data output could be a great resource for social and psychological investigators.

Thus, the path to follow is to detect offensive language and negative sentiment in tweets. The dataset that will be used contains 13,586 tweets that mentioned AMLO. To validate and label which tweets are offensive, the dataset is compared to the most used offensive words in spanish. Then, the tokenized text can be evaluated through a sentimental analysis, so a tweet that has offensive language and is negative will be considered hate speech.

Once the data is ready, it will be used to train a Natural Language Processing (NLP) model that classifies the tweets. The data will be splitted in train and test data, so the model can be trained and the classifying results can be measured by comparing it with test data.

Moreover, the results will be compared to the Benchmark Model, a hate speech detection research that applied a Convolutional Neural Network model to classify tweets. The comparison will be performed through the accuracy, precision and recall metrics, which were utilized in both models. Since the proposal is a classification model, these metrics can measure the false positives and false negatives in the predicted results compared to the test data.