Capstone Project

Eduardo Castro Bermúdez

Machine Learning Engineer Nanodegree

February 27, 2021

Hate Speech Detection

Twitter Mexico - AMLO

I. Definition

Project Overview

Social media has overtaken our understanding of human communication. 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[1]. 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”. 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 PyTorch 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.

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 Binary Classifier model.

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.

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.

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

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

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

II. Analysis

Data Exploration

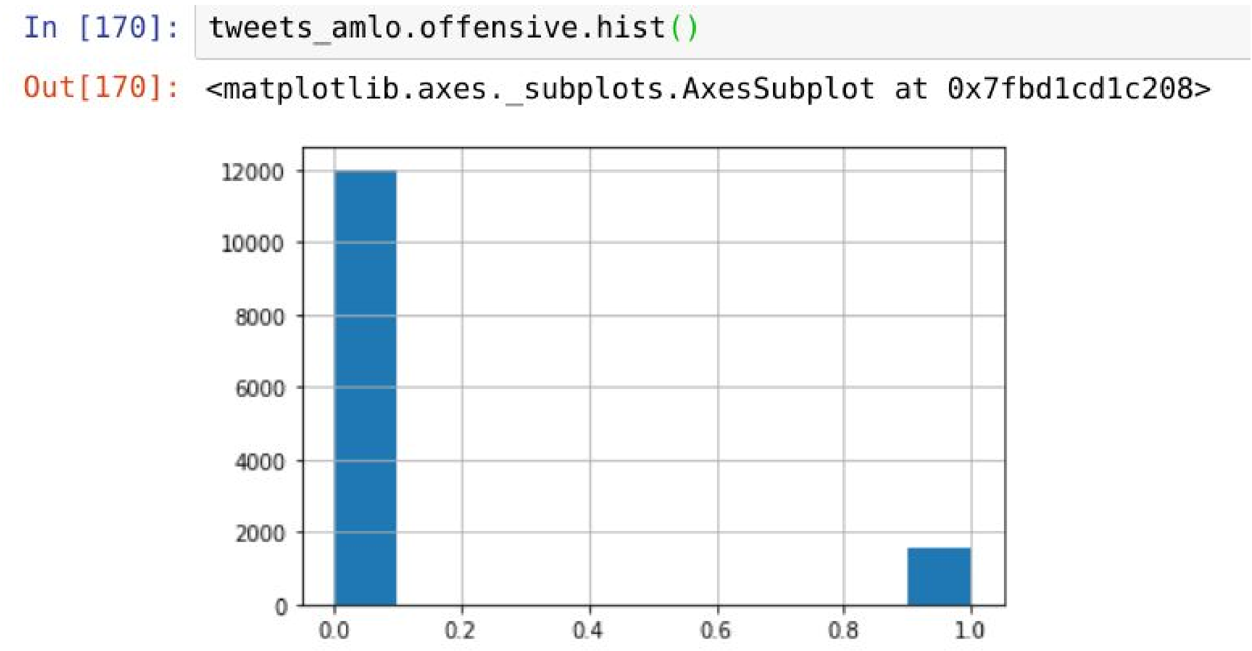
The input dataset to train the model consists of 13,586 tweets (in spanish) that mentioned Andres Manuel Lopez Obrador (AMLO), President of Mexico, through hashtags. This dataset is related to the hate speech issue due to the polarization crisis in Mexico since the 2018 Presidential Election[3]. Therefore, 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 spanish) from Hatebase[2], all tweets were labeled in order to know if they contained offensive language or not. After that, an spanish sentiment analysis model was applied to the tokenized tweets, thus So, the input data is a text file (.txt) that contains the features:

- User
- Publication date (date_time)
- Likes (integer)
- Tweet (text)
- Tokenized tweet (text)
- Offensive language (binary)
- Sentiment (float number)
- Hate Speech (binary)

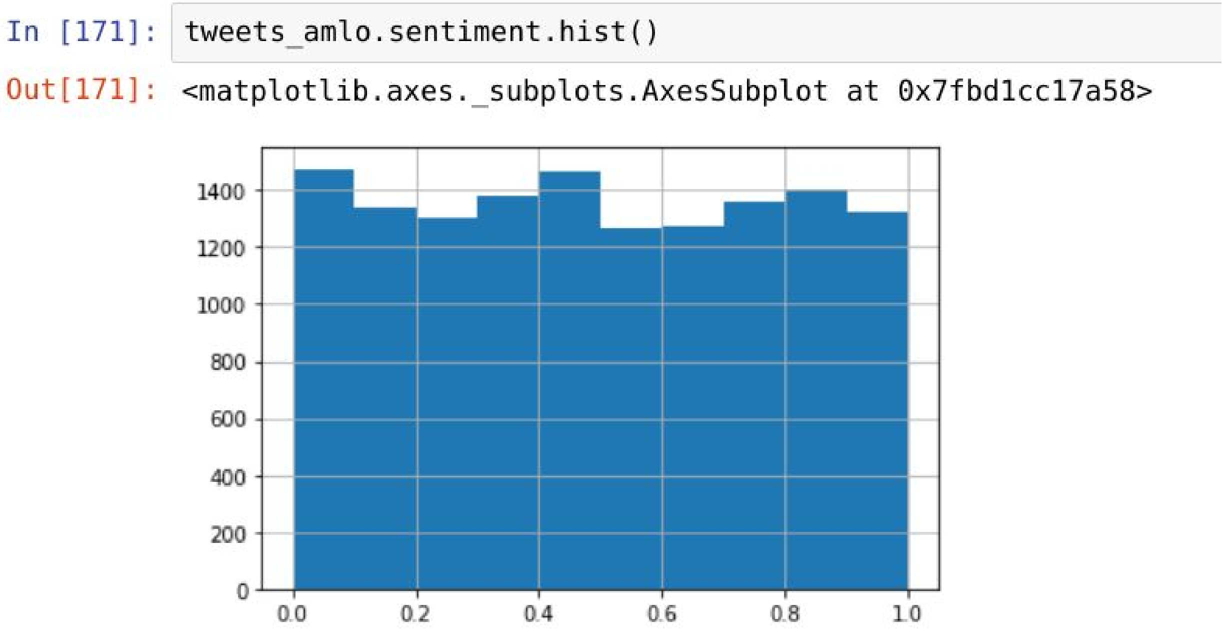
Exploratory Visualization

The features offensive language, sentiment and hate speech are significant to train the model. In this way, it has been identified that 1569 tweets are offensive, an 11.5 % of the total.



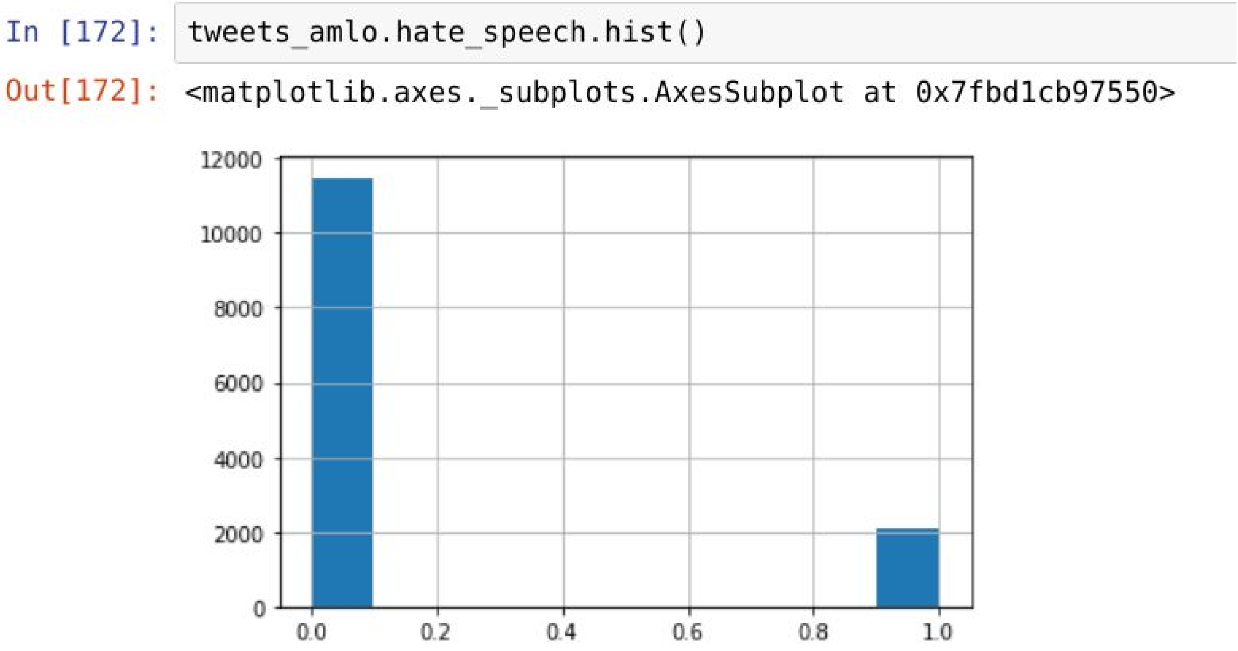
```
In [170]: tweets_amlo.offensive.hist()
Out[170]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbd1cd1c208>
```

Therefore, to add value to the data, this sample also measures the sentimental analysis of every tweet before the model is trained. A tweet that has offensive language and is negative is mainly considered hate speech.



```
In [171]: tweets_amlo.sentiment.hist()
Out[171]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbd1cc17a58>
```

Nevertheless, there are cases when a tweet is considered hate speech, even if it does not contain an offensive word. This is because there are few cases that are very negative and offend without having offensive language.



```
In [172]: tweets_amlo.hate_speech.hist()
Out[172]: <matplotlib.axes._subplots.AxesSubplot at 0x7fbd1cb97550>
```

Algorithms and Techniques

The PyTorch neural network is an algorithm that lets build a binary classification model which fits the needs to predict if a tweet is hate speech or not. Classification issues like this belong to

the category of machine learning problems where given a set of features, the task is to predict a discrete value.
This neural network accepts a number of features as input, and produces a single sigmoid value, that can be rounded to a label: 0 or 1, as output.

The parameters that were optimize allowed the model to perform better:

- Input_features
- hidden_dim
- output_dim
- epochs

Benchmark

The research paper **Hate Speech Detection Using Natural Language Processing Techniques**[3] 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%.

III. Methodology

Data Preprocessing

In order to prepare the input data to train the model, it was necessary to:

1. Split the dataset in input features and feature to predict.
2. Randomly split in train and test datasets.
3. Create a dataframe (.csv) for every dataset. Necessary format to SageMaker PyTorch estimator.

Implementation

After the data preprocessing, it was required to build the modules to the train and deploy the model with SageMaker.

First at all, the development of the model.py:

- Initializes the Binary Classifier
- Defines the layers and the sigmoid.

Then, the train.py:

- Loads the parameters to create the model
- Determines the device
- Builds the model
- Defines the loss function
- Train the model from a loop of the given configured epochs

Finally, the predict.py:

- Gets the predicted classification.

Once these modules were built, an SageMaker estimator is initialized with the given hyperparameters:

- Input_features: 2
- Hidden_dim: 2
- Output_dim: 1
- Epochs: 3000

After the model is trained, it is deployed and the endpoint is used to predict the test dataset. The predicted classification is compared with the actual hate speech classifiers and is measured through the accuracy, recall and precision metrics.

Refinement

In order to improve the model performance it was indispensable to test different hyperparameter options. The number a epochs was the most changeable parameter, given the loss results:

- 10 epochs: 0.21916286390799114
- 25 epochs: 0.08228225267934335
- 50 epochs: 0.041058589689447825
- 100 epochs: 0.02271489340677587
- 500 epochs: 0.007329866492447712
- 1000 epochs: 0.00464973892312214
- 3000 epochs: 0.0023330744952885386

IV. Results

Model Evaluation and Validation

The model clearly performs satisfactorily, even there is a possibility of overfitting. Given the input data (13586 records) and the selected epochs (3000), the predicted classification results implied the presented metrics:

	Recall	Precision
Hate Speech (1)	**1.0**	**1.0**
No Hate Speech (0)	**1.0**	**1.0**
Accuracy		**1.0**

The hyperparameter for the final trained model performed in the best way among other combinations.

Justification

In the present project the model validated results got an accuracy of 100%, precision of 100% and recall of 100%. Nevertheless, even those metrics seemed to perform satisfactorily, it is possible that overfitting was present due to the great use of epochs, resulting in over training.

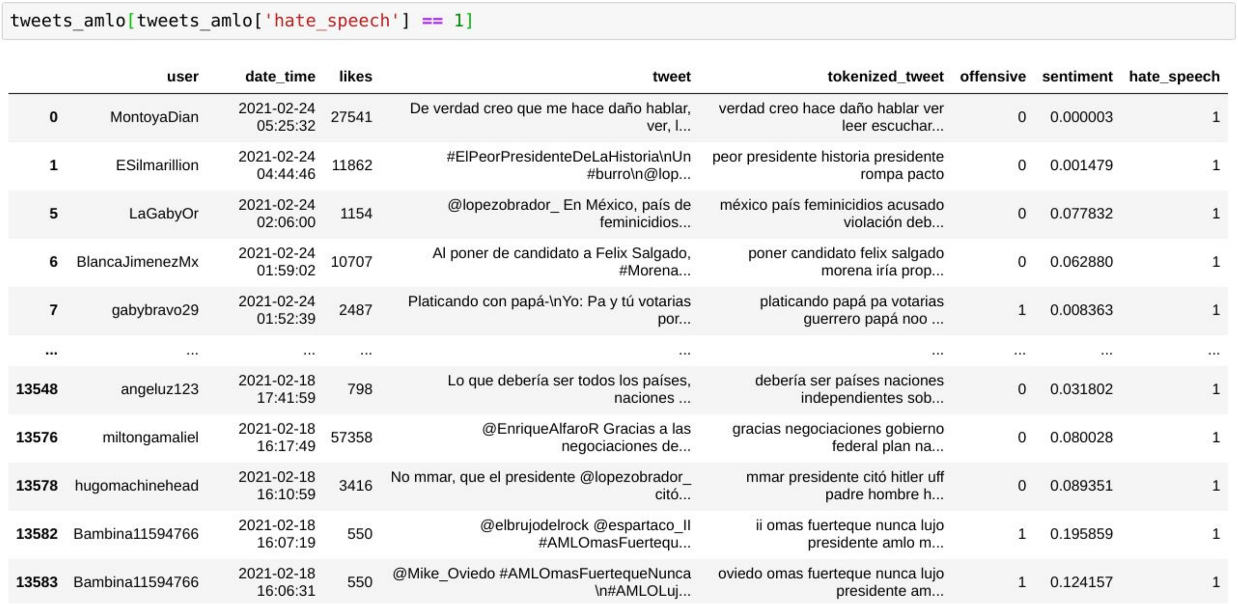
Compared to the benchmark metrics results, my model performed in a better way. However, one of the reasons for this performance project could be the unbalanced training data, thus there are considerably more tweets which are not labeled as hate speech. This is an issue presented in both projects, so it would esencial to build bigger and more balanced datasets in order to avoid that bias.

V. Conclusion

Free-Form Visualization

The input dataset has the “likes” feature, which presents the likes one tweet obtained. My decision was to disregard that feature, because I believed it does not represent value to the model.

Nevertheless, I noticed that those tweets, which are labeled as hate speech, are really popular. So, once the Hate Speech Detection model is improved, it would be an interesting idea to use the “likes” feature to feed the model and also to analyze the predicted classification, in a more social than technical analysis.



```
tweets_amlo[tweets_amlo['hate_speech'] == 1]
```

	user	date_time	likes	tweet	tokenized_tweet	offensive	sentiment	hate_speech
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	0.000003	1
1	ESilmarillion	2021-02-24 04:44:46	11862	#ElPeorPresidenteDeLaHistoria\nUn #burro\n@lop...	peor presidente historia presidente rompa pacto	0	0.001479	1
5	LaGabyOr	2021-02-24 02:06:00	1154	@lopezobrador_ En México, país de feminicidios...	méxico país feminicidios acusado violación deb...	0	0.077832	1
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	0.062880	1
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	0.008363	1
...	...	...	...	...	...	...	...	...
13548	angeluz123	2021-02-18 17:41:59	798	Lo que debería ser todos los países, naciones ...	debería ser países naciones independientes sob...	0	0.031802	1
13576	miltongamaliel	2021-02-18 16:17:49	57358	@EnriqueAlfaroR Gracias a las negociaciones de...	gracias negociaciones gobierno federal plan na...	0	0.080028	1
13578	hugomachinehead	2021-02-18 16:10:59	3416	No mmar, que el presidente @lopezobrador_ citó...	mmar presidente citó hitler uff padre hombre h...	0	0.089351	1
13582	Bambina11594766	2021-02-18 16:07:19	550	@elbrujodelrock @espartaco_II #AMLOmasFuertequ...	ii omas fuerteque nunca lujo presidente amlo m...	1	0.195859	1
13583	Bambina11594766	2021-02-18 16:06:31	550	@Mike_Oviedo #AMLOmasFuertequeNunca \n#AMLOLuj...	oviedo omas fuerteque nunca lujo presidente am...	1	0.124157	1

Reflection

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

Once the data was processed, it was used to train a Neural Network Binary Classifier model with PyTorch. The data was splitted in train and test data, so the model could be trained and the classifying results could be measured by comparing it with test data.

The more interesting part of the development was to test different hyperparameters and evaluate the model loss. SageMaker enabled the option to perform many tests in little time. Thus it was not difficult to decide to train the final model with 3000 epochs, and got a great performance.

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

Improvement

The most important issue to improve in all Hate Speech Detection projects is to build bigger and more balanced datasets, which mean better data to train. Also, it could be interesting to extract more features like the longest common subsequence (LCS) just for offensive words, in order to detect greater sense of hate speech.

Due to this project language, it is significant to point that there are fewer resources for Natural Language Processing (NLP) in Spanish. Therefore, an area of improvement is to create those resources.

VI. Bibliography

[1] 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.

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

[3] 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.