ANLY-580 Natural Language Processing Project proposal

Joshua Gladwell, Valeria Vera Lagos, Weiye Zhu

Automatic detection of misogynistic content of news articles through supervised classifiers

Problem

The responsibility of the media should be to objectively inform the public without attempting to influence their opinions but so far, media has been incapable of doing this. Communication platforms in Mexico have openly stated comments as “It was her fault”, “Her parents should have picked her up”, “Surely she set herself on fire”, “Look how she was dressed up that day”, “She might have fallen and bumped her head” among news articles of occurred femicides [1].

Media is portraying women who have been victims of any kind of violence as stigmatized, guilty, untrustworthy, or sexualized. This method of data governance creates a sense of revictimization in the survivors' or witnesses' families and fosters another form of violence against women: media violence.

An analysis done by [2] shows that out of 1,300 reports of violence against women, published by more than 20 media platforms from Argentina, 897 of them revictimized the sufferer women. Such approaches present dangerous interpretive frameworks that may end up justifying violent acts, perpetrating the misogynistic and discriminative culture the country is holding or even jeopardizing judicial investigations.

Existing approaches

Automatic detection of misogyny is a hard Sentiment Analysis task. It is difficult to distinguish it from sexist or aggressive language towards women, given that it can be subtle and dependent on the social context where it takes place.

Until now, most of the analysis of this problem are focused on English texts and the small amount of Spanish texts research doesn’t distinguish the text origin between Spanish spoken countries. Therefore, cultural, and linguistic differences can complicate the classification. Further difficult are multiple approaches that have generated data set of tweets in Spanish labeled as misogynistic or not, reduce the misogynistic detection to short texts since tweets can only hold 280 characters.

Finally, Sentiment Analysis detection of news articles has been performed in the past [3] focusing on “negative” or “positive” sentiment but no detection of misogyny in news articles has been performed on any language.

Proposal

Train multiple supervised models to classify from misogynistic and non-misogynistic texts obtained from news articles from Mexico in Spanish after applying different transforming and modelling techniques to those and analyzing their results.

Results will be evaluated in terms of accuracy, precision, and a deep understanding of how the classifier is detecting both subtle and clear misogyny.

Methods to achieve objective:

i. **Data collection:**
Maria Salguero is a Mexican activist who has collected multiple femicides articles published in Mexico. To have a bast dataset, her set of news and other news articles will be manually collected.

ii. **Data labeling:**
The text from each article will be split into paragraphs and sentences for them to be manually label as “misogynistic” or “non-misogynistic”. Different approaches for labelling had been defined*, based on professor feedback we will select one:*

Approach 1 - Binary classification:
If a paragraph contains 1 misogynistic sentence, then the article will be classified as misogynistic.

Approach 2 - Binary classification:
If a paragraph contains N number of misogynistic sentences the paragraph will be marked as misogynistic. If the article contains N number of misogynistic paragraphs, then the article will be classified as misogynistic.

Approach 3 - Multiclass classification:
A scale of ranges will be defined to classify if an article is misogynistic based on the number of misogynistic and the number of misogynistic paragraphs.

iii. **Data cleaning:**
The text will be cleaned through the standard NLP cleaning task such as punctuation, numbers, hypertext and stop words removals and lowercase transformation for all characters.

Based on the state-of-the-art results, lemmatization and stemming techniques have not improved the performance of sentiment classification in texts Spanish, thus multiple sets will be generated to test the results of those techniques.

iv. **Data modelling:**

Different words modelling techniques, that had shown good performance on Sentiment Analysis, will be implemented. The objective of the work is to do an exploratory analysis of words modelling techniques results for misogyny detection.

To do so, representations that loose semantic representation such as BoW with techniques to solve such problems like TF-IDF will be implemented. Models that preserve semantic representation will also be tested, for instance: BERT, Word2Vec.

We have a bast dataset of twitter data labeled as misogynistic or not. We are looking for feedback on, should we train the words embedding with that data? Should we test with a pretrained model with data in Spanish?

v. **Data classification:**

Multiple supervised models will be implemented. *We are looking for feedback on best models to choose from.*

vi. **Evaluation criteria:**

Since we have labeled data, results will be evaluated in terms of accuracy, precision, and a deep understanding of how the classifier is detecting both subtle and clear misogyny by testing it with text examples labeled as “subtle” and “clear” misogyny.

Hardware considerations

(Ask professor if we need Hardware specifications for transformers)

Biggest unknowns that might dictate the success or failure of this project:

i. Since we will be using transformers and a big amount of data is needed, the manual collection and labeling of the texts might need a huge amount of time.
ii. Lack of work to compare our results to.
iii. Bias results in terms of what can be labeled as “misogynistic” or not.

Possible extra development

i. Twitter bot to perform as server for our classifier.

ii. Classifier to return the reasons behind the prediction.
iii. Test the classifier with English data.

Results

Live demo and slide deck and oral presentation.

Bibliography

[1] “Fue su culpa: por qué es vital evitar la revictimización en los casos de feminicidio” https://www.infobae.com/america/mexico/2022/09/05/fue-su-culpa-por-que-es-vital-evitar-la-revictimizacion-en-los-casos-de-feminicidio/

[2] “Medios revictimizantes”. Palacios, Claudia, https://www.eltiempo.com/opinion/columnistas/claudia-palacios/medios-revictimizantes-observatorio-de-medios-y-genero-149406

[3] Chen, a. W. Global journal of advanced engineering technologies and sciences sentiment analysis of news articles and its comments: A Natural Language Processing Application, http://gjaets.com/Issues%20PDF/Archive-2018/May-2018/3.pdf