Authors:

1. Catherine Monoue Konga
2. Alioune B. M. Dihanka
3. Aman Kassahun Wassie

Gender Based Violence Tweet Challenge Classification

Introduction

One of the main areas of interest in machine learning is classification models that can identify tweets related to gender-based violence in order to present evidence to policymakers and law enforcement agencies in online conversations, particularly in social media where, due to the "culture of silence", victims of violence are afraid, ashamed, or intimidated to discuss their experiences with others and often do not report their experiences directly to authorities. Given the enormous amount of user-generated tweets each day, the problem of automatically detecting this type of content in real time becomes a fundamental problem, especially for female victims and one that Twitter has not been able to help. In our project, we focus on detection to create a machine learning algorithm that classifies tweets about GBV into one of five categories: sexual violence, emotional violence, harmful traditional practices, physical violence, and economic violence.

Problem statement

Violence is a complex phenomenon that is difficult to define and is generally understood differently by different people. Most people think about physical violence when the word violence is mentioned in a discussion. However, it is just one amongst a wide list of exaction that can be classified as violence. According to the Newfoundland and Labrador, we have nine types of violence: *Physical violence, sexual violence,emotional violence, psychological violence,spiritual violence,cultural violence, verbal abuse,financial abuse and neglect.* In USA, the National Center for Injury Prevention and Control[1] released a report in Novemver 2018 which states that in 2015, 43,6% (52,2 million) of women experienced some form of

[1] *https://www.nsvrc.org/sites/default/files/2021-04/2015data-brief508.pdf*

contact sexual violence whereas 24,8% (around 27,6 million) of men experienced it. By the meantime, one in five women experienced completed or attempted rape at some point of their lifetime while this is the case for 1 out of 14 men. Those statistics give us an overview of the problem. Also, the gang rape committed in India has made it clear that the gender-based violence is a serious problem. Therefore institutions have been working on ways to reduce it.

It is within this context that the Gender based violence tweet classification challenge has been launched on Zindi. Indeed the anonymity guaranteed by social media platforms has encouraged many victims and witnesses of violence to express themselve. Thus, it is a great opportunity that can be used to better understand this scourge and fight it.

Objective

The objective of the project is to do a classification between different kinds of gender based violence based on a dataset collected from social media Twitter. This study will ultimately help on reporting violences based on gender.

Hypotheses

The main problem is the imbalance between different classes in the dataset. This bias in the training dataset can influence many machine learning algorithms, leading some to ignore the minority class entirely. This is a problem as it is typically the minority class on which predictions are most important. To solve this problem we hypothesise two techniques. These techniques are Oversampling and Pseudo-labeling. Oversampling is to duplicate examples in the minority class. Pseudo labelling is the process of using the labelled data model to predict labels for unlabelled data. We used both these techniques to solve the imbalanced dataset. In addition to this, we used a pre-trained Bert model to do a knowledge transfer.

Data exploration Methodology

- Data analysis

Our work attempts to provide a classification model for gender based violence which can deal with unbalanced data. Indeed the collected data has shown ***unbalanced classes*** which is an important problem that all the machine learning algorithms for classification encounter. It has been noticed that directly applying a model on unbalanced data makes it focus only on the majorities classes. The model does not try to learn parameters that may help to predict minority classes and therefore it is impossible to predict them.

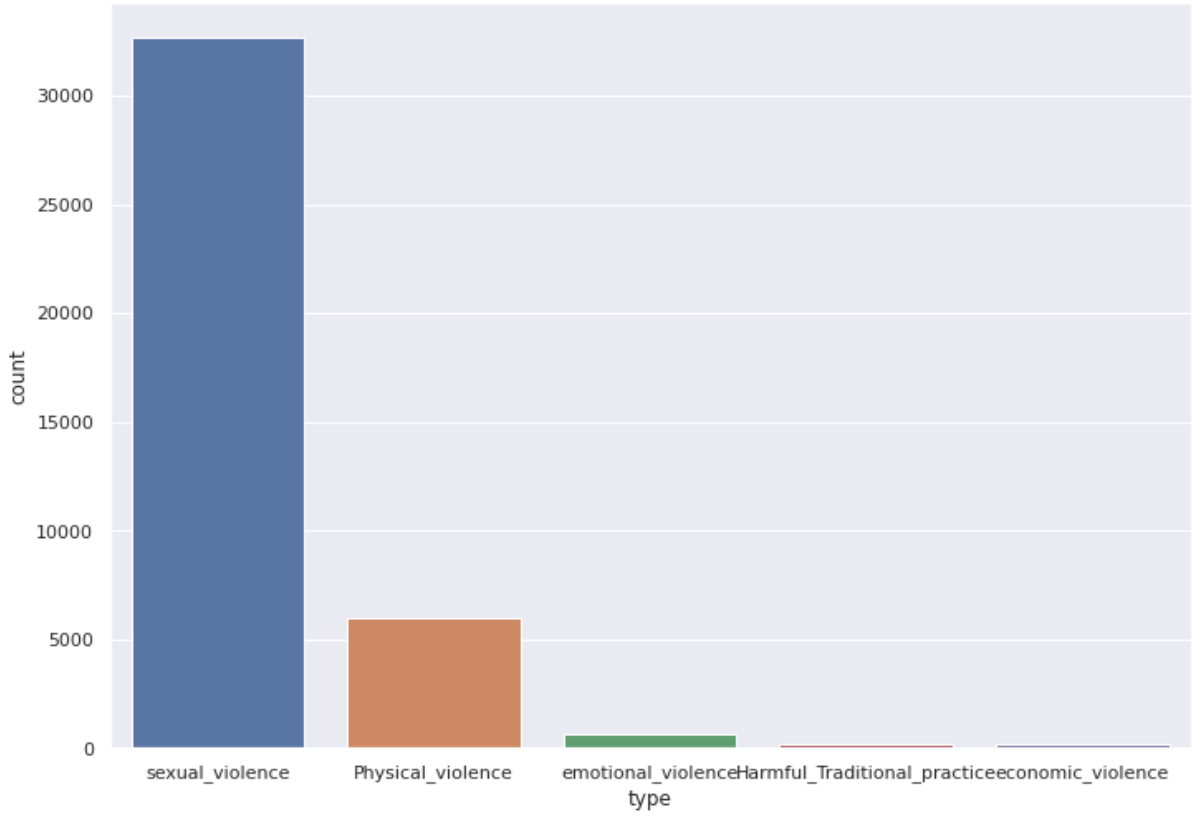

Fig 1. Label distribution

We did a length statistical analysis and the maximum length of words in a training example is 70 and the minimum is 3. The average length is 38.

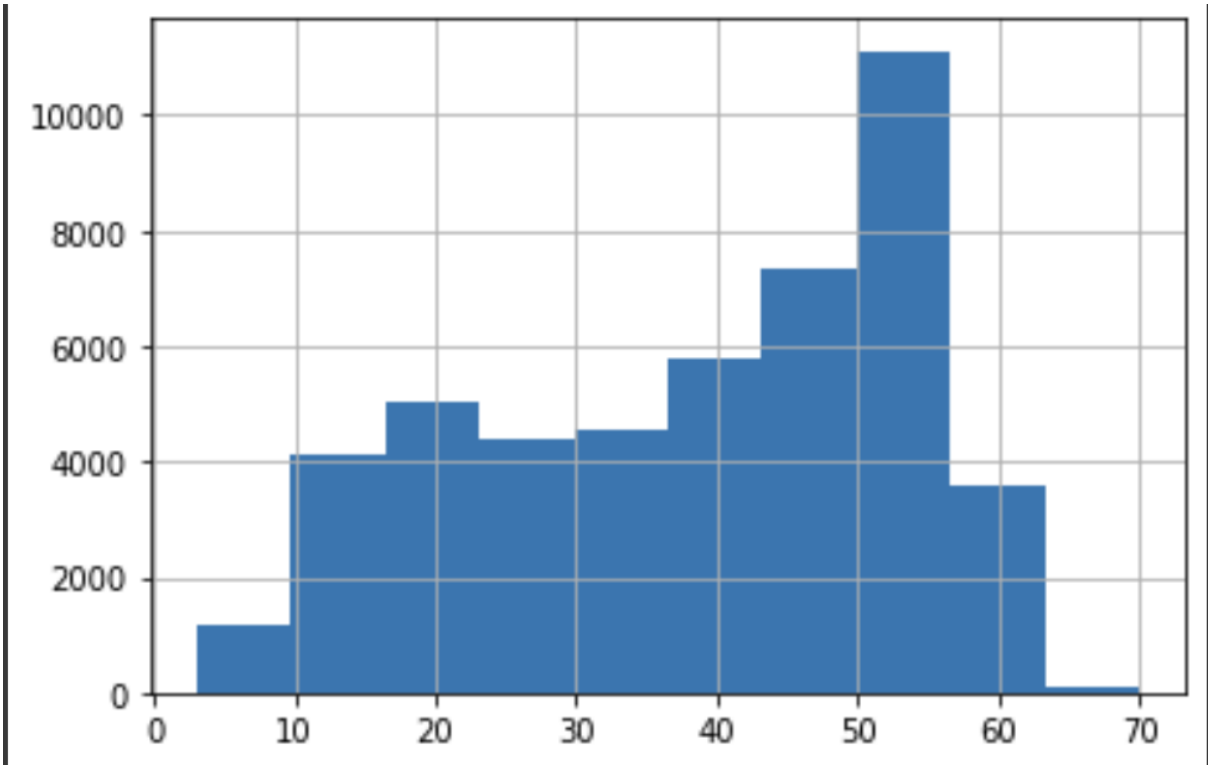

Fig.2 Length statistic analysis

- **Data processing**

We used the texthero library to lowercase sentences and remove digits, urls, punctuations and html tags. To lemmatize the sentences we used WordNetLemmatizer and we used the TweetTokenizer to tokenize the sentences.

- **Training steps**

In order to solve the unbalanced classes problem, we first use the SMOTE, and ADASYN oversampling method to get a balanced distribution of the labels and apply a Multinomial Naive Bayes. It helps to achieve an accuracy around 59% on the public leaderboard of Zindi which was an improvement of the model without an oversampling method (52% of accuracy). Since the Naive Bayes model is a simple model, we decided to look for another one more complex. Our first Idea was the BERT model since it is a model which captures well the content and the meaning of words in sentences. However, we faced a big problem, it was not possible to combine the Bert model directly with the SMOTE method as we did with the multinomial Naive Bayes. Our research led us to the " classification on imbalance dataset using BERT EMBEDDING" paper. There, the authors were proposing to oversample the embedding of the BERT model, an approach that we decided to try.

Our attempts to write a script able to modify oversampling the embedding were not successful. We propose instead a model that uses a sentence transformers model and dense layerS to address the classification task. The training step of our model is followed:

1. Get an embedding with the sentence transformers model
2. Oversample the embedding obtained
3. Train our dense layers with cross entropy loss

This model gives us an accuracy around 72% , which is a great improvement of our previous model. Among the other methods that we hear about, we try to implement "pseudo labelling " in order to build a balanced dataset. We combined this method with the BERT model and got around 86% of accuracy on the public leaderboard.

Related papers

We have been inspired by:

- The paper classification on imbalanced dataset using BERT EMBEDDING
- The degree project ***"Data augmentation in solving data imbalance problems"***

Results

Our result is explained in the following table

Model	Loss	Validation accuracy	Test accuracy
Multinomial Naive bayes	0.348	88.8	52.24
Multinomial Naive bayes with oversampling	0.221	96.3	59.38
Bert (dropout =0.1, activation=sigmoid, epoch =3, batch_size=4)	0.0012	99.93	76.23
Bert (dropout =0.5, activation=sigmoid, epoch =1, batch_size=16)	0.048	97.77	66.65
MiniLM + 5-layer NN with early stopping and oversampling	0.0352	99.12	72.96
Bert (dropout =0.1,	0.0040	99.76	91.06

activation=sigmoid, epoch =3, batch_size=4) with pseudo labelling			

References

1. https://www.nsvrc.org/resource/2500/national-intimate-partner-and-sexual-violence-survey-2015-data-brief-updated-release
2. https://www.diva-portal.org/smash/get/diva2:1521110/FULLTEXT01.pdf
3. https://github.com/UKPLab/sentence-transformers