UNIVERSITÀ
DEGLI STUDI
DI MILANO

Text Mining and Sentiment Analysis

Sexism Classification

Luca Graziano, ID:987996

July 5, 2023

Abstract

The following analysis acts as a first introduction to text classification problems and is the final paper for the "Text Mining and Sentiment Analysis" course for the Master's degree course "Data Science for Economics" by Università degli Studi di Milano. The dataset is composed of sexism related tweets and the main aim of the analysis is to build an efficient and well-performing binary classification solution using supervised machine and deep learning techniques. Further effort has been placed for a comparison in sexism between the English and Italian languages.

Introduction

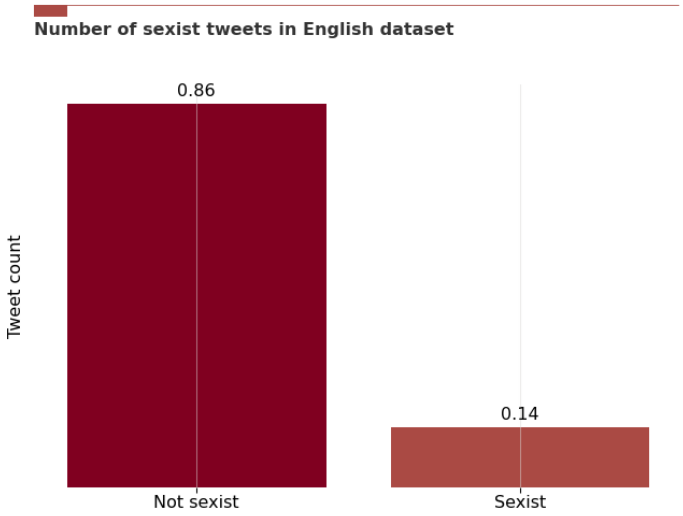
The analysis follows the data pre-processing as well as the development of several algorithms with the aim of achieving an efficient and well-performing solution to sexism identification for tweets. Finally, some interest is reserved to the comparison of the language used for sexist remarks in the Italian and English languages.

Data description

The data represent tweets that pertain to two different datasets:

- The Automatic Misoginy Identification at Evalita 2020, an Italian dataset composed of 3922 tweets labeled as sexist or not sexist
- The "Call Me Sexist" But, an English dataset composed of 13631 tweets labeled as sexist or not sexist from the Leibniz Institute for the Social Sciences

The English dataset is highly unbalanced, with the vast majority of tweets not being sexist.



Data Pre-Processing

A big portion of the analysis is about pre-processing the text data in order to prepare it for the modeling processes. Both datasets have followed the same pre-processing with the necessary changes due to the language they are based on.

First duplicates and missing values have been removed. Secondly, the removal processes has focused on the content of the tweets, specifically on:

- numbers;
- links;
- hashtags;
- user mentions;
- all forms of punctuation;

- all extra spaces;
- all emoticons;

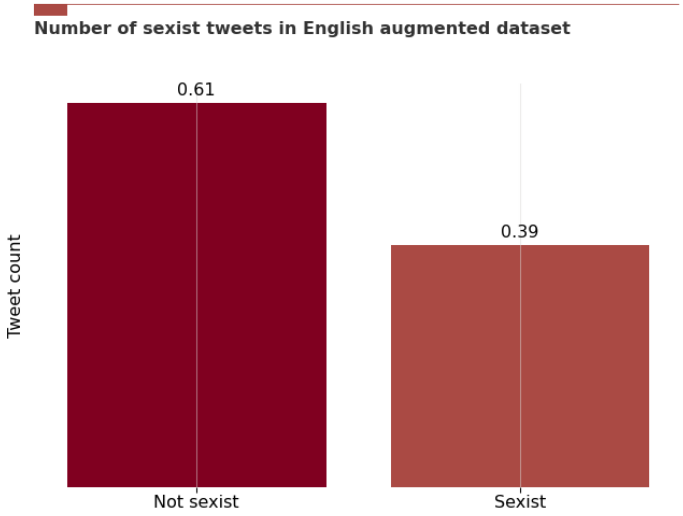
Finally, all text has been standardised to lower case letters.

Data Augmentation

One of the biggest issues with NLP problems tend to be the data itself which can be scarce or highly unpredictable. The English dataset, due to its class unbalance is a clear example of a data scarcity problem related to the positive class, that is sexist observations. In order to overcome this issue a number of techniques have been considered throughout the development. The choice fell down on data augmentation, which is very popular within the computer vision space but underexplored in NLP. The purpose of data augmentation is generating additional, synthetic training data in insufficient data scenarios. Data augmentation ranges from simple techniques like rule-based methods to learnable generation-based methods[1]. In order to be valid, augmented data has to include some diversity to improve model generalisation on future tasks. One of the most popular and effective ways to augment text data is using machine translation. This method means that the original text is translated into other languages, and then translated back to obtain the augmented text in the original language. Different from word-level methods, back-translation does not directly replace individual words but rewrites the whole sentence in a generated way[2]

The severe under-representation of the sexist class required data augmentation using six different languages: Italian, Spanish, German, French, Portoguese.

After exploiting the data augmentation technique, the class unbalance has been drastically reduced. The augmented dataset class distribution is:



Feature Engineering

The tweets content has then been TF-IDF Vectorised and train/test split before feeding it to the machine learning models. TF-IDF vectorisation has been chosen over different forms of vectorisation as it has historically proven to be highly effective and yet very simple and compute efficient.

[1]Raille, G., Djambazovska, S., Musat, C., 2020. Fast cross-domain data augmentation through neural sentence editing. arXiv abs/2003.10254. https://arxiv.org/abs/2 003.10254. arXiv:2003.10254

[2]Bohan Li, Yutai Hou, Wanxiang Che, Data augmentation approaches in natural language processing: A survey, AI Open, Volume 3, 2022, Pages 71-90, ISSN 2666-6510, https://doi.org/10.1016/j.aiopen.2022.03.001

TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word.

$$IDF = \log(N/n_k)$$

where N is the total number of documents and n_k is the number of documents in which the term k appears

$$TF = O_{k,x}/D_x$$

where $O_{k,x}$ is the number of times the word k appears in the document x and D_x is the number of words present in document x

Finally, TF-IDF results from the multiplication between TF and IDF.

$$TF - IDF = TF_{k,x} * IDF_k$$

TF-IDF vectorisation has been used for regular machine learning techniques while for the CNN GloVe vectorisation has been used. GloVe stands for global vectors for word representation and it is an unsupervised deep learning algorithm developed by Stanford that aims at capturing the semantic meaning of a word in a vector space. The GloVe algorithm leverages the statistics of word co-occurrence in a large corpus of text to learn word vectors. Specifically, it looks at how often words co-occur with each other within a context window in the training corpus. The underlying assumption is that words with similar meanings tend to have similar co-occurrence patterns.

GloVe vectors have several advantages. They capture both syntactic and semantic information, perform well on various word analogy tasks, and exhibit good generalization properties even for words not seen during training.

Modeling

In order to identify sexist tweets and find an efficient and well-performing algorithm that solves the binary classification problem, five different algorithms have been created:

- Logistic Regression
- Decision Tree
- Xgboost
- Convolutional Neural Network
- Bidirection LSTM

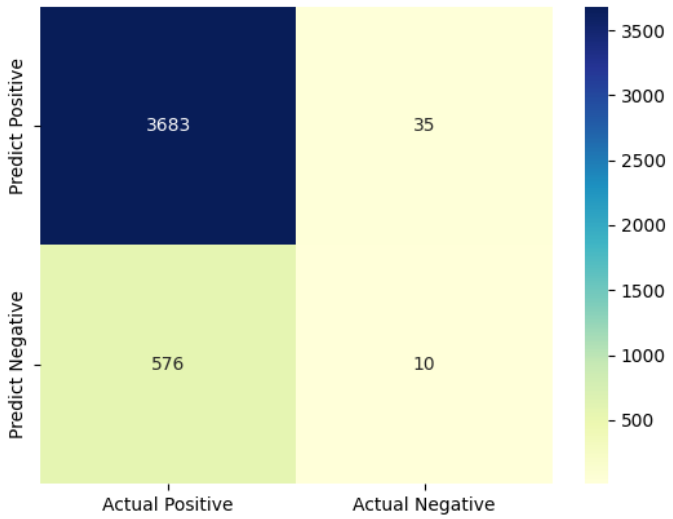
Again, all models have been run exclusively on the English dataset and for the first three models TF-IDF vectorisation has been used while for the CNN and BiLSTM glove vectorisation was preferred. All models have also been trained on the English dataset after it has been augmented trying to balance the positive and negative classes.

Given the fact the English dataset is highly unbalanced, the performance metric that has been used to compare all models is the recall in an attempt to optimise the model to identify sexist tweets. What is expected from the model is to identify most and ideally all sexist tweets and to result in a good amount of false positives, that is non sexist tweets flagged as sexist. However, it can be argued that for a social network like Twitter, identifying all sexist tweets is by far the most important thing and human moderators can be hired to ensure that flagging errors are resolved.

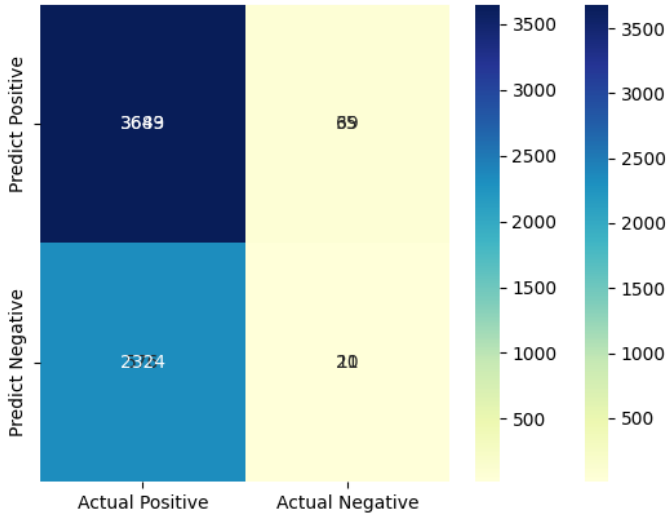
In addition to data augmentation, to further deal with the unbalanced dataset, proper class weights that mimic the distribution of the classes over the dataset have been provided to the models. With exception for the CNN and BiLSTM, class weights have been assigned by dividing the number of datapoints of one class with that of the other.

Logistic Regression

The logistic regression model is a very simple supervised learning technique that models the probability of one event taking place by having the logarithm of the odds for the event as the event of a linear combination of one or more independent variables. The logistic regression is a parametric method that estimates the coefficients in the linear combination of predictors.



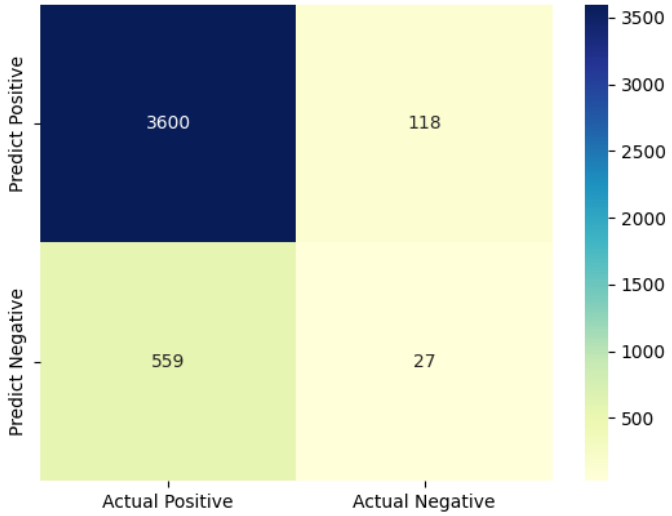
The logistic regression trained on the non-augmented dataset has performed extremely poorly with a recall on the positive class of 2%



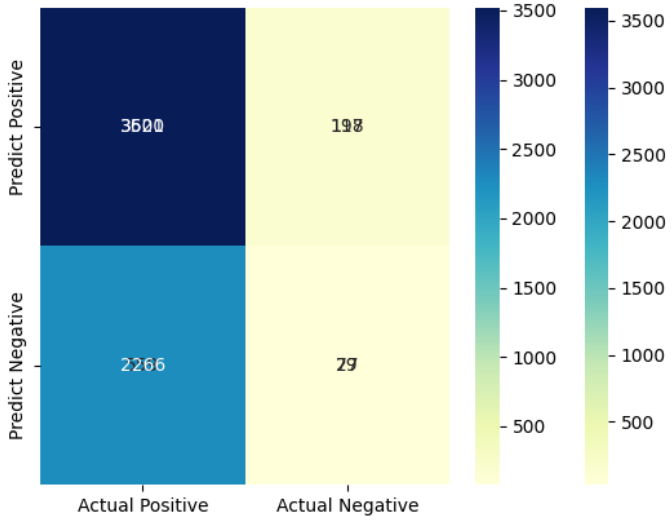
The logistic regression trained on the augmented dataset has performed extremely poorly with a recall on the positive class of 1%

Decision Tree Classifier

The decision tree is a non-parametric method of supervised learning that can be used both for regression and classification problems. Each tree is made up of nodes and leaves. The former set conditional rule using tests and through which data are split; whereas the leaves provide the resulting label based on the splits. The tests set out inside the nodes depend on one of the variables and a given threshold defined in order to create the best possible split. In order to find the best available split, either the Gini Index or Cross Entropy can be used.



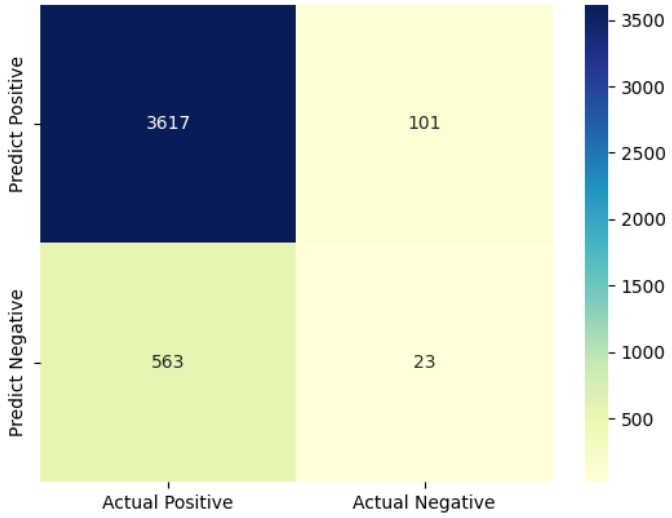
The decision tree classifier trained on the non-augmented dataset has performed extremely poorly with a recall on the positive class of 5%



The decision tree classifier trained on the augmented dataset has performed extremely poorly with a recall on the positive class of 3%

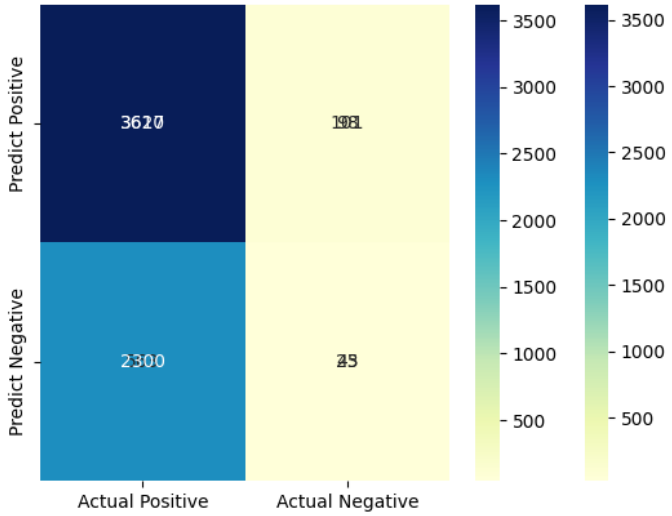
Xgboost

A Gradient Boosting Decision Trees (GBDT) is a decision tree ensemble learning algorithm that combines a multitude of weak decision tree models to build one single strong classifier.



The xgboost trained on the non-augmented dataset has performed extremely poorly with a recall

on the positive class of 4%



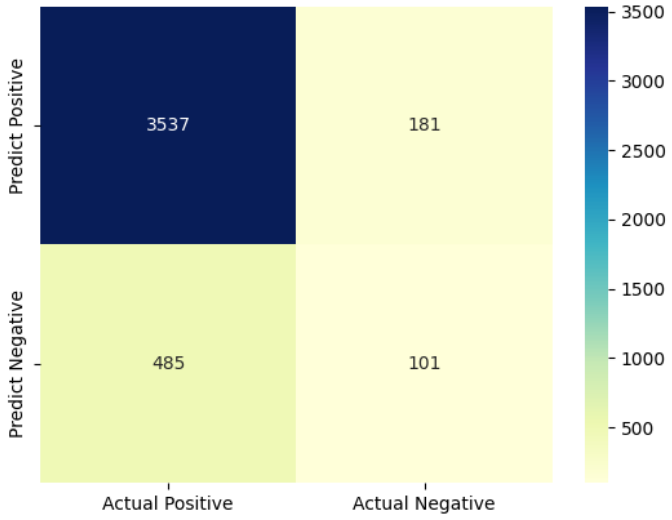
The xgboost tree classifier trained on the augmented dataset has performed extremely poorly with a recall on the positive class of 2%

CNN

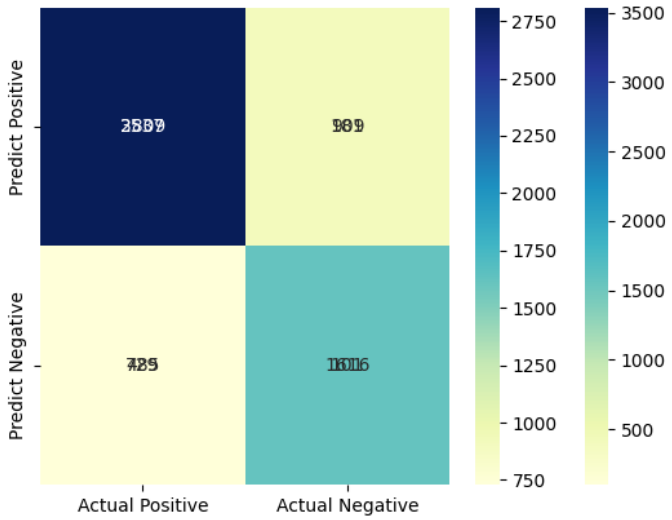
A convolutional neural network is a particular neural network tipically used to perform Computer vision tasks that can also be used for text classification tasks. A Convolution layer performs convolution, that is the process of applying filters to an input that results in activation in order to identify and learn specific features. Adding multiple convolutional layers on top of another, the CNN is able to identify higher level features.

This model presents three 1 dimensional convolution and max-pooling blocks and one dense layers. The structure breaks down as follows:

- a first 64 filters with size (3 x 3) convolution layer with a Relu activation function
- a first max-pooling layer
- a second 128 filters with size (3 x 3) convolution layer with Relu
- a second max-pooling layer
- a third 256 filters with size (3 x 3) convolution layer with Relu
- a third max-pooling layer
- one 128 nodes dense layer



The CNN trained on the non-augmented dataset has performed poorly with a recall on the positive class of 17%

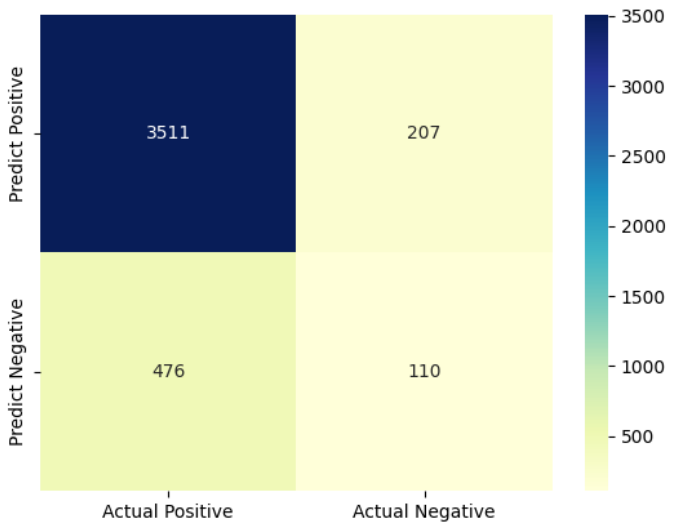


The CNN trained on the augmented dataset has performed well with a recall on the positive class of 69%

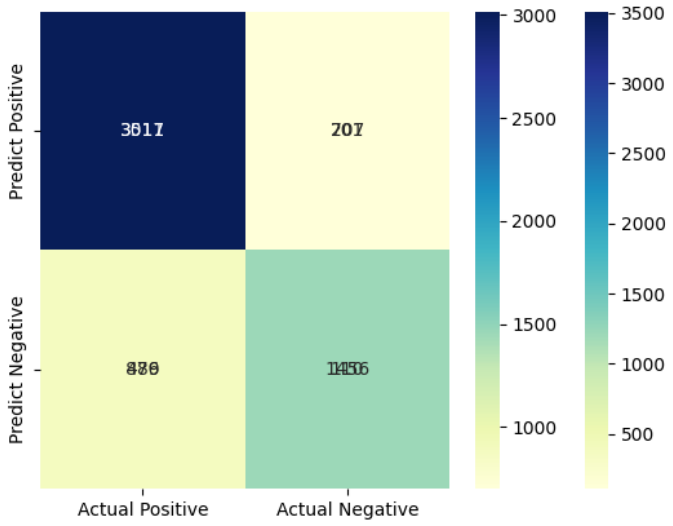
BiLSTM

An RNN allows for predicting what word comes next by not only considering the current input but also the previous input. Bidirectional LSTM (BiLSTM) is a recurrent neural network used primarily on natural language processing. Unlike standard LSTM, the input flows in both directions, and it's capable of utilizing information from both sides.

The model presents two bidirectional LSTM layers and one dropout layer.



The BiLSTM trained on the non-augmented dataset has performed poorly with a recall on the positive class of 19%



The BiLSTM trained on the augmented dataset has performed well with a recall on the positive class of 62%

Italian and English Sexism Comparison

The last part of the analysis has been developed to compare the most prominent words pertaining to Italian and English language used in sexist tweets. To do so, a wordcloud, that is a graphical representation of word frequency where the frequency is transmitted through the size of the word in the picture: the bigger the word the more frequent it is across tweets.

Before developing a wordcloud, the English text has been lemmatized, a process where all inflected words are grouped together to act as single element. Lemmatization is very useful to reduce the number of similar words present in the image as well as ensuring that the frequency of words is more in line with the actual word distribution. Additionally, a process of keywords extraction to capture the most salient keywords using KeyBERT has been performed. The algorithm uses BERT-based models to generate word embeddings, which are dense numerical representations of words that capture semantic and contextual information. These embeddings are then used to calculate the similarity between words and phrases, allowing KeyBERT to identify the most salient keywords in a document.

The English wordcloud is decently representative of sexism with words phrases such as "feminists are ugly", "a day without women" only "ya all women" or "bitch dont cut it".

On the other hand, the Italian wordlcloud has mostly references to abusive actions and verbs like "seviziare", "strozzare" or "strangolare". Finally, both wordclouds present multiple different words indicating the female sex from girl to sister and woman as expected.

Conclusion

The analysis has followed through the pre-processing and the attempt at developing efficient and well-performing machine learning models tasked with identifying sexist tweets. Additionally, a comparison between the Italian and English language in sexism comments have been carried out. Unfortunately, the quality of the data and the existing class unbalance have resulted in extremely poor sexism identification by all simpler models. Only the CNN and BiLSTM have performed decently after being trained on the data augmented version of the original dataset. Regardless, a final recall on the positive class of nearly 70% by the CNN trained on the augmented dataset is a satisfactory result for the analysis though several improvements, like including an attention layer, to the models could lead to even better performances.