Multi-Lingual Sexist Hate Speech Detection and Classification

Ritvik Garg
(2020201071)

Rishabh Malik
(2020201074)

I. Code Links

Model	URL
Roberta	Link
XLM-R	Link
BERT and BETO	Link
Data Augmentation	Link
BERT Multilingual	Link
BERTweet	Link
Augmented BERT and BETO	Link
Class balanced BERT and BETO	Link

II. Problem Statement

Sexism against women is a cultural component, whose principle is the supremacy of men over women in different areas of life, such as in the workplace, politics, society and even the family. We find sexism in daily conversation, in the disregard for opinions expressed by women and in statements loaded with discriminatory ideologies. The invisibility, anonymity and accessibility that social media platforms provide today, have led to the inequality and discrimination against women remain embedded in society, increasingly being replicated and spread online. Understanding sexism and how it is different from other forms of harassment and hate speech gives us more potential to restraint the harm caused on digital social platforms. Researchers have made several attempts in classifying sexism to achieve more robust datasets or to achieve a better understanding of sexism from the text.

The detection of sexist content is still a difficult task for social media platforms, as it may be expressed in very different forms, but it is necessary in order to design new equality policies, as well as to encourage better behaviour in society. The problem is relatively less explored field in NLP. Our aim in this study is to create more understanding of the machine and deep learning approaches for sexism in a broad sense, ranging from objectification, explicit misogyny to other types of implicit sexist behaviours.

The sEXism Identification in Social neTworks (EXIST) dataset has been developed to address this critical social problem. It contains more than 10000 labelled texts collected from twitter and gab distributed almost equally between English and Spanish. The dataset has been compiled by web crawling based on sexist phrases identified by two gender studies experts. Measures have been taken to avoid seed phrase bias, temporal bias and user bias. The problem can be divided into two tasks.

The first task, Sexism Identification, is a binary classification problem to classify the text as sexist or non-sexist. Both the training and test dataset contains an even mixture of both classes.

The second task, Sexism Categorization, is a multi-class classification problem where texts labelled as sexist have to be classified into one of five labels: ideological-inequality, stereotyping-dominance, objectification, sexual violence, misogyny-non-sexual-violence.

III. Related Work

A. *MultiAzterTest@Exist-IberLEF 2021: Linguistically Motivated Sexism Identification*

MultiAzterTest team participated only in Task 1. They aimed to see if linguistically motivated features can help in the detection of sexism. They used three approaches : i) an approach based on language models, ii) an approach based on linguistic and stylistic features + machine learning classifiers and iii) an approach combining the previous approaches. The Language Model approach uses Bidirectional Encoder Representation from Transformer BERT for English and BETO for Spanish. For this approach, they have used the cross-entropy loss function for each of the outputs.On top of BERT, they have probed with two sequential models: i) a dropout layer to fight overfitting. ii) a linear layer, ReLU activation function and linear layer model. The second approach, the MATS-Sexism approach, consists of the outputs of the tool plus a classical machine learning classifier. They have experimented with the Sequential Minimal Optimization (SMO), Random Forest (RF) and Simple Logistics (SL) classifiers out of which SMO gave best results. The third approach is a combination of the results of the LM and MATS-Sexism, finally classifying a tweet as sexist if both approaches consider that the tweet is sexist. The LM approach obtains the best results in all the settings: both languages, and English and Spanish on their own.

B. *Sexism Identification in Social Networks using a Multi-Task Learning System*

SINAI-TL has used a Multi-Task Learning (MTL) system using the transformer-based model BERT. In the MTL model they have integrated knowledge from 3 different tasks related to sexism identification, namely, sentiment analysis (assuming that sexism texts are associated with a negative polarity), emotion analysis (assuming negative emotions such as anger, fear, sadness and disgust could be related to sexism texts while positive emotions are not) and offensive language identification (assuming sexism identification is associated with offensive language). They have used an additional dataset for these 3 tasks: EmoEvent for emotion analysis data, HatEval for hateful content data and MEX-A3t for offensive language data.

C. Automatic Sexism Detection with Multilingual Transformer Models

AIT_FHSTP used data augmentation to tackle low data samples. They used the MeTwo (Spanish dataset) and HatEval2019 dataset. In addition, they converted English to Spanish and vice-versa to augment data. To model the data, they used mBERT and XLM-R as the 2 variants of transformers. Their approach uses two different strategies to adapt the transformers to the detection of sexist content: first, unsupervised pre-training with additional data and second, supervised fine-tuning with additional and augmented data. For both tasks XLM-R with unsupervised pre-training on the EXIST data and additional datasets and fine-tuning on the provided dataset performed better.

D. Sexism Identification using BERT and Data Augmentation – EXIST2021

The paper focuses on the pre-processing techniques and data augmentation to boost results on various machine learning and deep learning methods for sexism detection and classification. It uses 'Back Translation' for augmentation of the text and transformers as well as various machine and deep learning algorithms for analysis. They applied a range of classifiers such as Logistic Regression (LR), Multilayer perceptron (MLP), Random Forest (RF), Support Vector Machine (SVM), 1 Dimensional Convolutional Neural Network (1D-CNN), Long short-term memory (LSTM) and BERT on the augmented dataset and use the one vs. rest technique for sexism categorization task. It uses tenfold cross-validation to ensure the robustness of evaluation. It was observed that with proper pre-processing, machine learning can produce competitive results in comparison to deep learning methods and tends to outperform even deep learning methods such as 1D-CNN. The best performing algorithm on both tasks was BERT with data augmentation. Among the machine learning algorithms, RF performed the best.

E. Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media

In this paper, an analysis of six publicly available datasets is done after combining them into a single homogeneous dataset. Having classified them into three classes, abusive, hateful or neither, a baseline model is created to improve model performance scores using various optimisation techniques. They discovered valuable sequential information in Twitter text for each class. When they applied a BiLSTM layer on top of a CNN layer, it could capture the sequential information as well as the low lying textual representation. Thus, improving the simple contextual BiLSTM classification with use of CNN layers. They used random search to find the parameters and hyperparameters. One key point that was highlighted is that the BERT and CNN LSTM models take much less time in comparison because of the fact that the logistic regression model requires TF-IDF features on the entire vocabulary size. The objective of this study was to bring to light a model which was trained on a large dataset of multiple languages. By experimenting on an aggregated dataset combining six datasets in English, Hindi and Code-mixed Hindi, they demonstrate that their models achieve comparable or superior performance to a wide range of baseline monolingual models.

F. HASOCOne@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection

This work focuses on hate speech detection. The main obstacle with hate speech is, it is difficult to classify based on a single sentence because most of the hate speech has context attached to it, and it can morph into many different shapes depending on the context. Another obstacle is that humans cannot always agree on what can be classified as hate speech. Hence it is not very easy to create a universal machine learning algorithm that would detect it. They solve this by using the pre-trained BERT transformer model. The pre-trained BERT models have a better word representation as they are trained on a large Wikipedia and book corpus. As the pre-trained BERT model is trained on generic corpora, they fine-tune the model for the downstream tasks. The BERT models were run on a NVIDIA RTX 2070 graphics card with an 8 GB graphics card. They also compare it with a baseline model consisting of ELMO embeddings classified by a SVM.

IV. Workflow

Dataset : EXIST2021 dataset contains 6977 training instances in English and Spanish. In total there are 3426 English and 3541 Spanish social media postings from Twitter and Gap. The test set contains 4368 instances, split into 2208 English and 2160 Spanish postings from mentioned sources. They are annotated in a binary fashion (task 1) as either sexist or non-sexist; and in a more fine-grained categorization (task 2) as:

1) ideological-inequality: The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression.
2) stereotyping-dominance: The text expresses false ideas about women that suggest they are more suitable to fulfil specific roles (mother, wife, family caregiver, faithful, tender, loving, submissive, etc.), or inappropriate for certain tasks (driving, hard work, etc.), or claims that men are somehow superior to women.
3) objectification: The text presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfil traditional gender roles (compliance with beauty standards, the hypersexualization of female attributes, women's bodies at the disposal of men, etc.).
4) sexual-violence: Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made.

5) misogyny-non-sexual-violence: The text expresses hatred and violence towards women.

The final set of sexism terms was used to extract tweets both in English and Spanish.

A. *Methodology*

1) ***Preprocessing*** *:* We have tried various methodologies for solving this task. First and foremost, we performed data cleaning for the dataset to remove unwanted noise. The following things were done as part of dataset cleaning:

1) Regex in order to filter and remove all unwanted noise in the dataset.
2) **Stemming:** Reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.
3) **Expanding Hashtags:** We replace hashtags after expanding them. For e.g. lifeisgood will be expanded and replaced with ‘life is good’.
4) **Link Removal:** URLs (or Uniform Resource Locators) in a text are references to a location on the web, but do not provide any additional information. We thus remove these too using the library named re, which provides regular expression matching operations.
5) **Contraction Expansion:** Contractions are words or combinations of words that are shortened by dropping letters and replacing them by an apostrophe. For e.g. take the text ‘I'll be there within 5 min. Shouldn't you be there too?’. It'll be changed to ‘I will be there within 5 min. should not you be there too?’ after expanding contractions.
6) **Mentions Removal:** We remove mentions/tags present in the text like @name, @username, etc.
7) **Percentage Removal:** We remove percentages from the text, as it usually isn't very informative for such tasks.
8) **Elongation Removal:** In informal writing people may use a form of text embellishment to emphasize or alter word meanings called elongation (a.k.a. ”word lengthening”). For example, the use of ”Whyyyyy” conveys frustration. We replace the elongation with known semantic replacements.
9) **Emojis Removal:** Emojis are part of our life. Social media text has a lot of emojis. We need to remove the same in our text analysis.
10) **Lower casing:** For the sake of simplicity we convert text to lowercase.
11) **Punctuation Removal:** We remove punctuations like ‘.’, ‘,’, ‘;’, etc.

NOTE : For the spanish data, we applied all the above mentioned cleaning techniques except (5).

2) ***Language specific approach for task1*** *:* For the english language used pretrained models, namely XLM-R, ROBERTA, BERT to generate features. We then trained a single learnable layer followed by softmax to give the probability for the text being classified as SEXIST/NON-SEXIST. We computed the ACCURACY, PRECISION, RECALL and F1 score (macro weighted avg) to see how well the model performs.

For the Spanish task we made use of the BETO model keeping all the other steps the same. BETO is a BERT model trained on a big Spanish corpus. BETO is of size similar to a BERT-Base and was trained with the Whole Word Masking technique.

3) ***Language specific approach for task2*** *:* For the english language we used pretrained models, namely XLM-R, ROBERTA, BERT and BERTWEET to generate features. We then trained a single learnable layer followed by softmax to give the probability for the text being classified as one of ideological-inequality, misogyny-non-sexual-violence, objectification, sexual-violence or stereotyping-dominance. We computed the ACCURACY, PRECISION, RECALL and F1 score (macro weighted avg) to see how well the model performs.

For the Spanish task we made use of the BETO model keeping all the other steps the same.

4) ***Language independent approach for task1/task2****:* We used the BERT multilingual base model to solve tasks for both languages simultaneously. The model is pre-trained on the top 102 languages with the largest Wikipedia using a masked language modeling (MLM) objective. It was introduced in paper and first released in this repository. We keep the remaining setup same as in the above two mentioned approaches.

5) ***Models Used:***

a) ***BERT****:* Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google. Has 12-layers, 768-hidden state, 12-heads and 110M parameters.

b) ***Roberta****:* A robustly optimized method for pretraining natural language processing (NLP) systems that improves on Bidirectional Encoder Representations from Transformers, or BERT, the self-supervised method released by Google in 2018. Has 12-layers, 768-hidden state, 12-heads and 125M parameters. RoBERTa uses the BERT-base architecture.

c) ***XLM-R****:* A new model from Facebook AI called XLM-R , where ‘R’ stands for Roberta, By the name it's very common to assume that XLM in addition to Roberta, instead of BERT, but it would be incorrect to say that. XLM-R is different from XLM, avoiding TLM(Translation language model) objective, it just trains Roberta on a huge multilingual dataset at a large scale. Around 100 languages was extracted from CommonCrawl datasets, i.e. 2.5TB of text data (unlabeled). XLM-R is trained only with an objective of the Masked language Model(MLM) in a Roberta way. Has 270M parameters with 12-layers, 768-hidden-state, 3072 feedforward hidden-state, 8-heads. Trained on on 2.5 TB of newly created clean CommonCrawl data in 100 languages.

d) ***BETO****:* BETO is a BERT model trained on a big Spanish corpus. BETO is of size similar to a BERT-Base and was trained with the Whole Word Masking technique. Model has 12 self-attention layers with 16 attention-heads each, using 1024 as hidden size. In total our model has 110M parameters.

e) ***Bertweet****:* BERTweet is the first public large-scale language model pre-trained for English Tweets. BERTweet is trained based on the RoBERTa pre-training procedure, using the same model configuration as BERT-base. The corpus used to pre-train BERTweet consists of 850M English Tweets (16B word tokens 80GB), containing 845M Tweets streamed from 01/2012 to 08/2019 and 5M Tweets related to the COVID-19 pandemic. BERTweet does better than its competitors RoBERTa-base and XLM-R-base and outperforms previous state-of-the-art models on three downstream Tweet NLP tasks of Part-of-speech tagging, Named entity recognition and text classification. Has 135M parameters.

6) ***Data Augmentation****:* We used machine translation to convert the Spanish text in the dataset in to English and the English text into Spanish thus doubling the text in the dataset. This was done using the opus-mt transformer model from the Helsinki-NLP collection. We then trained bert on the English text and Beto on the spanish text for both task 1 and task 2. We found that there was a slight improvement in the accuracy compared to the original dataset.

7) ***Class Balancing****:* While the classes for task one are almost balanced since the dataset has a ratio of 1800:1636 for English and 1800:1741 for Spanish, the classes for task two are very imbalanced, especially in Spanish where there are only around 200 tweets in the classes of objectification and sexual violence while there are more than 400 tweets for each of the other classes. To overcome this, we assigned more weights to the classes with less text so that they will be sampled more, thus balancing out the classes. The weight for a class was given as the (max num of elements in any class) / (num elements of that class). This was found to increase the accuracy.

8) ***Hyper-parameter Tuning****:* For the models above, we tried varying the learning rate, the number of epochs of training and the batch size. We selected 256 as the maximum length of sequence, instead of the default of 512, since all the text were shorter than this and we found that this decreased the training time. We found that a small batch size of 8 gave the best result since the models were over fitting for larger batch sizes. For deciding learning rate and number of epochs, we started with a base of 1e-5 lr for 1 epoch and ran a stochastic seek using wandb to find the best values for these hyper-parameters optimising for validation accuracy. For this we divided the training data set into training and validation data sets in approx 2:1 ratio. We found that most of the models are overfitting since they gave a low training loss while giving higher validation loss. We found that a learning rate of approximately 2e-5 for 5 epochs gave the best accuracy on validation and test sets.

B. Evaluation Metrics:

We evaluate the system outputs as classification tasks (binary and multi-class respectively) using standard evaluation metrics, including Accuracy, Precision, Recall, and macro-averaged F1-score. We make use of the class-weighted F1-score for Task 2 because the dataset is highly imbalanced.

V. Results & Analysis

Task - 1 Results (%)				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
BERT	76.86	74.23	85.58	79.50
BETO	76.52	78.47	75.60	77.01
BERT Multilingual	72.02	72.95	73.78	73.37
Roberta	60.19	71.78	82.38	76.72
XLM-R	60.01	69.43	79.45	74.10
BERTweet	76.67	73.69	86.35	79.52
BERT Augmented	76.31	76.70	78.75	77.71
BETO Augmented	73.68	71.22	83.59	76.91
BERT Class Balanced	76.67	76.24	80.65	78.38

Task - 2 Results (%)				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
BERT	59	59	59	59
BETO	65	68	65	65
BERT Multilingual	60	61	60	60
Roberta	60	62	60	59
XLM-R	52	50	52	47
BERTweet	65	65	65	64
BERT Augmented	64	65	64	64
BETO Augmented	67	69	67	67
BERT Class Balanced	61	60	61	60
BETO Class Balanced	68	68	67	67

VI. Conclusion

We can observe that the best model for task 1 is BERT for English and BETO for Spanish. Thus a language specific approach works better than a multilingual approach. We can see from the results table that data augmentation has not changed the accuracy significantly for task 1. However, in task 2, we can clearly see that the score has increased for all models when data augmentation and class balancing were done. Thus Augmented and class balanced BETO is the best model for Spanish, however for English, the best result is shown by BERTweet, which is not surprising as the corpus used to pre-train BERTweet consists of 850M English Tweets (16B word tokens - 80GB), which is suitable for our tasks. The test accuracy for all the models are less than training accuracy indicating that there is slight overfitting.

In both the tasks, our models easily surpas the baseline and our best models for task 1 would be ranked third in all the runs submitted for the shared task, beaten by less than 2 percentage points by the state of the art model. Our models for task 2 are performing better than the state of art models in the shared task by almost 8 percentage points for English and 7 percentage for Spanish.

VII. References

1) *Sexism Identification in SociSexism Identification in Social Networks using a Multi-Task Learning System.* Flor Miriam Plaza-del-Arco, et al.
2) *Automatic Sexism Detection with Multilingual Transformer Models AIT_FHSTP@EXIST2021.* Schütz Mina, et al. .
3) *MultiAzterTest@Exist-IberLEF 2021:Linguistically Motivated Sexism Identification.* Kepa Bengoetxea, Itziar Gonzalez-Dios
4) *HASOCOne@FIRE-HASOC2020: Using BERT and Multilingual BERT models for Hate Speech Detection* Suman Dowlagar, Radhika Mamidi.
5) *Online Multilingual Hate Speech Detection: Experimenting with Hindi and English Social Media* Neeraj Vashistha, Arkaitz Zubiaga
6) *Sexism Identification using BERT and Data Augmentation – EXIST2021* Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander Gelbukh.
7) *https://arxiv.org/abs/1810.04805*
8) *Overview of EXIST 2021:sEXism Identification in Social neTworks* Francisco Rodr´ıguez-S´anchez et al