Cyberbullying Detection in Social Networks

Areej Eweida, May Steinfeld-Kalisher, Sheryl Sitruk

Data Science Fellowship, Israel Tech Challenge

Abstract- While social media offer great communication opportunities and working platforms for individuals and companies, they also increase the vulnerability of receiving negative, harmful, false, or mean content causing threatening situations online. Recent studies report that cyberbullying constitutes a growing phenomenon among society. Successful prevention depends on the adequate detection of potentially harmful messages and the information overload on the Web requires intelligent systems to identify potential risks automatically. In this paper we focus on cyberbullying detection in social media text by training state of the art NLP models on posts and tweets. perform a series of binary classification experiments to determine the feasibility of automatic cyberbullying detection. We evaluate different machine learning models (SVM, Naïve Bayes and Random Forest) and finally make use of Transformer, an attention mechanism that learns contextual relations between words by applying BERT model, exploiting a rich feature set and investigate which information sources contribute the most for this particular task. Experiments on a holdout test set reveal promising results for the detection of cyberbullying-related posts. After optimization of the hyperparameters, the classifier yields a validation accuracy of 93%.

I. INTRODUCTION

Cyberbullying is a disturbing online phenomenon with far reaching consequences. It appears in different forms, and in most of the social networks, it is in textual format. Automatic detection of such incidents requires intelligent systems. Most of the existing studies have approached this problem with conventional machine learning models and most of the developed models in these studies are adaptable to a single social network at a time. In recent studies, deep learning-based models have found their way in the detection of cyberbullying incidents, claiming that they can overcome the limitations of the conventional models, and improve the detection performance. In this paper, we investigate the findings of a recent literature in this regard. We successfully reproduced the findings of this literature.

Bullying is not a new phenomenon, and cyberbullying has manifested itself as soon as digital technologies have become primary communication tools. On the positive side, social media like blogs, social networking sites (e.g. Twitter) and instant messaging platforms (e.g. WhatsApp) make it possible to communicate with anyone and at any time. Moreover, they are a place where people engage in social interaction, offering the possibility to establish new relationships and maintain existing friendships. On the negative side however, social media increase the risk of children being confronted with threatening situations including grooming 1 or sexually transgressive behavior, signals of depression and suicidal thoughts, and cyberbullying. Users are reachable 24/7 and are often able to remain anonymous if desired.

The goal of this sort of research is to develop models which could improve manual monitoring for cyberbullying on social networks.

After examining different models to achieve the proposed task we propose a deep learning method to cyberbullying detection and manage to reach 93% accuracy of automatic detection.

The remainder of this paper is structured as follows: the next section presents a theoretic overview and gives an overview of the state of the art in cyberbullying detection, whereas Section 3 describes the dataset construction. Next, we present the experimental setup and discuss our experimental results. Finally, Section 6 concludes this paper and provides perspectives for further research.

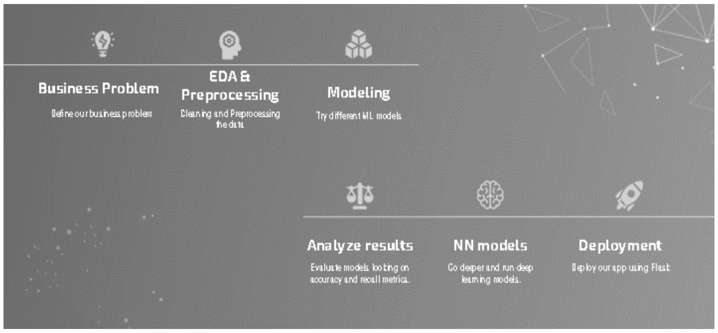

Business Problem
Define our business problem
EDA &
Preprocessing
Cleaning and Preprocessing the data
Modeling
Try different ML models
Analyze results
Evaluate models looking on accuracy and recall metrics.
NN models
Go deeper and run deep learning models.
Deployment
Deploy our app using Flask

Figure 1: Pipeline

II. RELATED RESEARCH

Cyberbullying is a widely covered topic in the realm of social sciences and psychology. A fair amount of research has been done on the definition and prevalence of the phenomenon (Hinduja & Patchin, 2012; Livingstone et al., 2010; Slonje & Smith, 2008), the identification of different forms of cyberbullying (O'Sullivan & Flanagin, 2003; Vandebosch & Van Cleemput, 2009; Willard, 2007), and its consequences (Cowie, 2013; Price & Dalgleish, 2010; Smith et al., 2008). In contrast to the efforts made in defining and measuring cyberbullying, the number of studies that focus on its annotation and automatic detection, is limited (Nadali et al., 2013). Nevertheless, some important advances have been made in the domain over the past few years.

III. COLLECTING DATA AND ANNOTATION

To be able to build representative models for cyberbullying, a suitable dataset is required. This section describes the construction of the final dataset (number-of-samples wise) that we used in the following parts of this article.

The combination is done to obtain an enriched dataset with improved quality and to avoid unbalanced datasets as noticed in the single datasets that used in the construction.

The dataset is the product of combining 3 datasets that we found online, each of the following datasets contains labeled posts or tweets that are classified as different annotations of cyberbullying (offensive, hateful, etc.) or non-cyberbullying.

A. Hate Speech

9924 posts classified as hate speech or not.

B. Labeled Tweets

11091 tweets classified are offensive and non-offensive.

C. Public Data

24,783 tweets classified as hate speech, offensive or neither, in this dataset we combined data under the labels hate speech and offensive and added them to the Cyberbullying label in our dataset.

After combining these three datasets we managed to obtain a balanced dataset consisting of 45600 labeled posts.

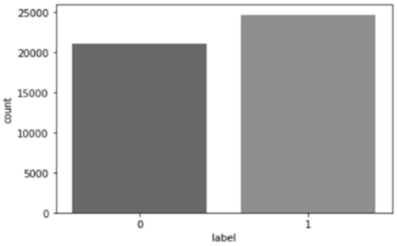


Figure 2: Histogram of total counts for each label in the dataset.

IV. EXPERRIMENTAL SETUP

In this paper, we explore the possibility of automatically recognizing signals of cyberbullying. Applying state-of-the-art approaches to cyberbullying detection (text data classification). The experiments described in this paper focus on the detection of such posts, which are signals of a potential cyberbullying event to be further investigated by human moderators.

For the automatic detection of cyberbullying, we performed binary classification experiments on different models and after performance comparison we choose BERT and Transformers model.

A. Pre-processing and Feature Engineering

In supervised learning, a machine learning algorithm takes a set of training instances (of which the label is known) and seeks to build a model that generates a desired prediction for an unseen instance. To enable the model construction, all instances are represented as a vector of features that contain information that is potentially useful to distinguish cyberbullying from non-cyberbullying content.

As pre-processing we applied the following steps using the Natural Language Toolkit (nltk) and Python basic functions.

1. Remove some conventions used on social media such as the expression “RT” that stands for a “Re-Tweet”.
2. Lemmatization: Transforming any form of a word to its root word.
3. Tokenization: Converting a sentence into list of words.
4. Remove punctuations.
5. Remove stop words.

B. DATA VISUALIZATION

The best way to understand the data we have is to plot it. We first check the length of posts in the dataset and try to find some pattern, we can see that most of the posts are composed of around 5-12 words (average number of words ~ 9 words in each post). This underlying distribution of posts length (fig. 3) poses a limitation on the structure of unseen data that optimize the performance of the predictive model. This issue well be touched in future work.

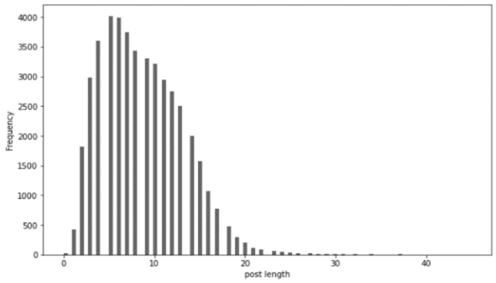


Figure 3: Histogram of the dataset posts length frequency

Identifying frequent signals and occurrences of cyberbullying is the core key of a working automatic detector, for that matter we can already sense the feasibility of automatic cyberbullying detection in social media data (fig. 4)

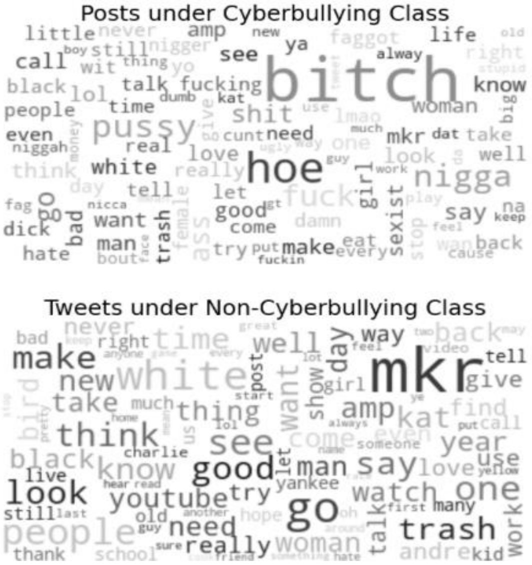


Figure 4: Most common words under each of the labels

C. Modeling

Conventional machine learning models:

Our goal is to find a baseline binary classifier with satisfactory performance. As we cannot work directly with text data when applying machine learning algorithms, we convert the texts to vectors of numeric features. To achieve this, we use Count Vectorizer to encode the data from words to numbers.

Now, in the classification process each post is an "input" and the class labels are the "output" for our predictive algorithm.

A simple and effective model for thinking about text classification in machine learning is the Bag-of-Words Model. The model is simple and throws away all the order information in the words and focuses on the occurrence of words in a dataset. This encoded vector is returned a length of the entire vocabulary and an integer count for the number of times each word appeared in the document.

Finally, after all our words were converted to vectors of numbers, we manage to split our data into train and test sets. The train-test split is of ratio of 20%.

We try to run multiple basic machine learning models.

1) Naïve Bayes
2) Random Forest
3) Support Vector Machine

Deep Learning Model:

Finally, we go deeper and try to run a SOTA model for NLP. We use the pretrained BERT model from transformers-Hugging face and this model gives us the best results. We run the model for 100 epochs with an early stopping and a flag to save the best weights. This allows us to save in memory the parameters that give us the best prediction score.

BERT is a multilingual transformer-based model that has achieved state-of-the-art results on various NLP tasks. BERT is a bidirectional model that is based on the transformer architecture, it replaces the sequential nature of RNN (LSTM & GRU) with a much faster Attention-based approach.

With this library there is no need to do all the tokenization and preprocessing on the text data as we saw before for other models. The BERT model already knows how to tokenize and preprocess the data by itself.

First it takes the text and tokenize it with the word piece encoder, then these tokens are added with segments embeddings and position embeddings to know its place in the sentence. This is the input to the self-attention layers.

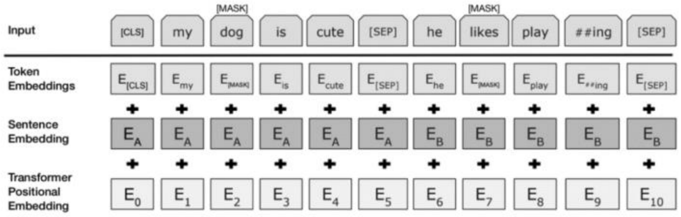

Input
[CLS]
my
[MASK]
dog
is
cute
[SEP]
he
[MASK]
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} $E_{[MASK]}$ E_{is} E_{cute} $E_{[SEP]}$ E_{he} $E_{[MASK]}$ E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
+
Sentence Embedding
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
+
Transformer Positional Embedding
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Figure 5:BERT input representation

D. Deployment

We deploy our deep learning model using the Flask API.

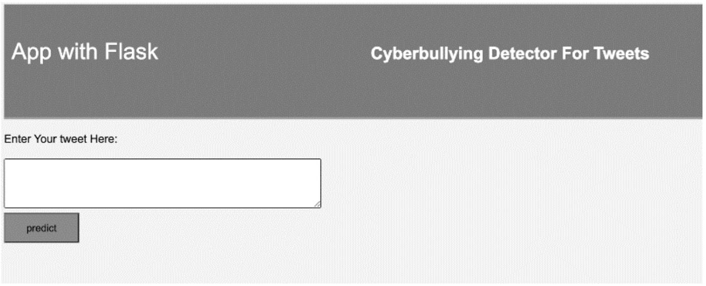


Figure 6: Cyberbullying detecting app for tweets

User's will be able to enter their tweet and be able to see an output telling them if their input is Cyberbullying or not.

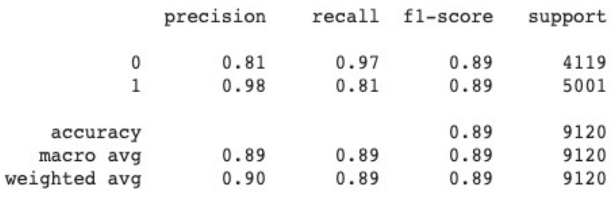
More details about the implementation is provided in our GIT repository attached in the references.

V. RESULTS

In this section, we present the results of our experiments on the automatic detection of cyberbullying-related posts using the different models described in the previous section.

Precision, Recall and F1 performance metrics are shown on the positive class (i.e., 'binary averaging') for the machine learning models (fig. 7-9), and for the BERT model we present the history of accuracy and loss function (cross entropy). Accuracy is considered as the performance metric rather than ROC curve (AUC) scores because data imbalance was dealt with previously.

Using Naïve Bayes yields the worst performance in terms of all metrics, we get accuracies of 89% and 91% for random forest and SVM classifiers respectively, and finally for the BERT model we get the highest accuracy of 93%, and however it may be considered as not much of an improvement comparing the latter machine learning model, it helps avoiding the limitations of conventional machine learning models as mentioned previously.

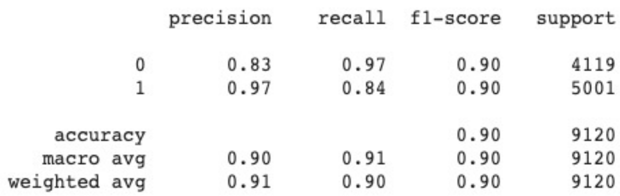


```
              precision    recall  f1-score   support

           0       0.81      0.97      0.89      4119
           1       0.98      0.81      0.89      5001

    accuracy                           0.89      9120
   macro avg       0.89      0.89      0.89      9120
weighted avg       0.90      0.89      0.89      9120
```

Figure 7:Random Forest Classification Report

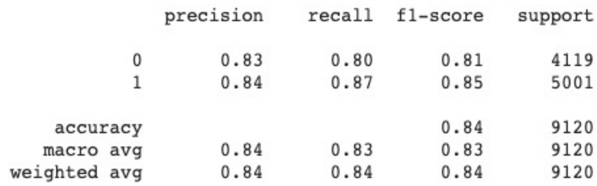


```
              precision    recall  f1-score   support

           0       0.83      0.97      0.90      4119
           1       0.97      0.84      0.90      5001

    accuracy                           0.90      9120
   macro avg       0.90      0.91      0.90      9120
weighted avg       0.91      0.90      0.90      9120
```

Figure 8: SVM Classification Report



```
              precision    recall  f1-score   support

           0       0.83      0.80      0.81      4119
           1       0.84      0.87      0.85      5001

    accuracy                           0.84      9120
   macro avg       0.84      0.83      0.83      9120
weighted avg       0.84      0.84      0.84      9120
```

Figure 9: Naive Bayes Classification Report

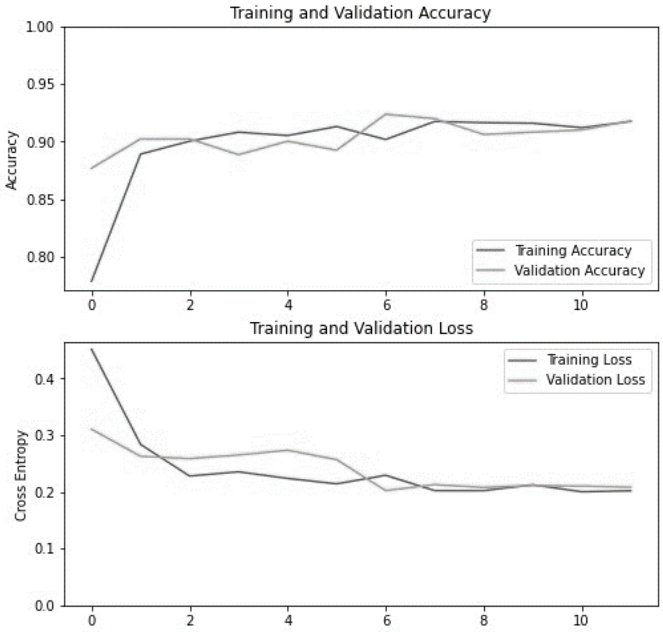


Figure 10: Accuracy and cross entropy history in BERT training

VI. CONCLUSION

The goal of the project was to investigate the automatic detection of cyberbullying-related posts on social media. A set of binary classification experiments were conducted to explore the feasibility of automatic cyberbullying detection on social media. In addition, we sought to determine which information sources and which models contribute to this task.

Our experiments reveal that the current approach is a promising strategy for detecting signals of cyberbullying in social media data automatically. After feature selection and hyperparameter optimization, the final classifier (deep learning pretrained model) achieved validation accuracy of above 93%.

A future work might include expanding our work by applying the developed methods on new datasets and different platforms. Also, transferring and evaluating the performance of the models trained on one platform to another platform.

CHALLENGES

Some of the challenges that the team faced in the project.
During the elaboration of this project, our team faced some challenges. This most important one was dealing with the reality of doing a project over zoom. At the beginning it was hard to communicate and work together over zoom and it took some time until we found the best way to divide the work between us and communicate properly.
More related to the model itself, we encounter some trouble to access high-quality data.
At the beginning, we did not find a satisfactory dataset for our task. Finally, we decided to combine multiple datasets to work with a balanced and large dataset. Also, we tried multiple models until we obtained a good accuracy, and each one of them required a different approach (preprocessing methods) of dealing with the data.

ACKNOWLEDGMENT

The work presented in this paper was carried out in the framework of the Israeli Tech Challenge (ITC) final project.

REFERENCES

[1] “Hate Speech and Offensive Language” dataset, kaggle.

[2] “Twitter Hate Speech” dataset, kaggle.

[3] “Cyberbullying Detection in Social Networks Using Deep Learning Based Models; A Reproducibility Study”, https://arxiv.org/abs/1812.08046

[4] Kaggle’s notebook “Text Data Cleaning - tweets analysis”, https://www.kaggle.com/ragnisah/text-data-cleaning-tweets-analysis.

[5] “Automatic Detection of Cyberbullying in Social Media Text”, link: https://arxiv.org/pdf/1801.05617.pdf

[6] Git repository: https://github.com/mayste/ITC_final_projct.git