Research Journal - Milestone 2

Ionescu Diana & Dumitrescu Andrei

1. ***ConvAbuse***

ConvAbuse [1] is a custom dataset specially designed for the task of detecting abusive language in conversations. The dataset was built by using conversations between different users and three conversational AI bots. These three bots are: *Alana V2, CarbonBot* and *ELIZA*. The first 2 bots are text based, while *ELIZA* is voice based. Additionally, *Alana V2* and *ELIZA* are classified as chat-bots, while *CarbonBot* is a system designed to convince the user to buy carbon offsets. The dataset is made out of conversation between these bots and users during different periods of time. Samples that include testing conversations were discarded. In order to form the dataset, the researchers propose to extract samples based on a list of keywords that are abusive. There are multiple samples selected based on how many sentences contain such words. In order to obtain a balanced dataset the researchers also extract samples that don't contain any of those keywords. After the process of sampling, different annotators assert for each extracted sample if it is considered abusive or not and also label it with an abuse type (sexism, racism, homophobia, etc.). The dataset contains a total of 6837 samples with a total of 20,710 ratings, as there are multiple annotators, and around 27% of the samples were found abusive.

The authors of the paper also suggest multiple experiments for benchmarking their dataset. We found these experiments really useful and a good starting point for solving the task of abusive language detection regarding the dataset. There are proposed three different baselines: a random classifier which outputs a label uniformly at random, a keyword filtering method which labels the samples based on the keywords used for sampling and a "jigsaw" method which uses a commercial system to label the samples. Additionally, there are proposed three machine learning methods: one based on a *Support Vector Machine* applied on the bag-of-words representations of the samples, a *Multi-Layer Perceptron* with a hidden layer which uses the same representation as mentioned previously, and lastly, a method based on *BERT* embeddings where the *BERT* model is fine-tuned for 4 epochs. The metric used for measuring the performance was the macro F1 score.

2. ***Pay "Attention" to Your Context when Classifying Abusive Language***

We find this paper [2] important as it explores multiple aspects of classifying abusive language. First of all, it analyses how different types of attention can affect the performance of deep learning models regarding the researched task. Second of all, it introduced us to 4 more datasets that we could use in our project. Lastly, it presents multiple experiments with architectures based on Bi-LSTMs and word embeddings. In the following paragraphs we will give more details regarding these topics.

In the paper, there are defined 6 different types of abusive language: *Racism, Sexism, Hate Speech, Offensive Language, Harassment* and *Personal Attack.* All these types will be present in at least one of the following 4 datasets. The datasets are: *D1 [3], D2 [4], D3 [5]* and *D4 [6]*. The *D1* dataset consists of 15844 tweets which are labelled with three classes: *Racism, Sexism* and *None*. The *D2* dataset is made out of 25112 tweets which are also

classified into three different classes: *Hate Speech, Offensive Language, Neither.* The *D3* dataset has 20362 tweets binary labelled into: *Harassment* and *Non-Harassment*. Lastly, the *D4* dataset is a bit different from the rest, as it consists of 115K Wikipedia comments, classified into: *Personal Attack* and *None.* Additionally, the paper offers some insight into how these datasets are balanced.

The paper also offers a perspective into how different types of attention mechanisms can affect the learning process of the models. The two main types of attention that the paper focuses on are: *contextual attention* and *self-attention*. The first mentioned one differs from the later one by introducing an additional context vector that is used in computing the attention and that is learned during the training process. The experiments presented in the paper suggest that the *context attention* performs better than the *self-attention.*

The main proposed architecture consists of a stacked Bi-LSTM model which also incorporates the attention mechanisms mentioned above. There are multiple experiments run using different word representations. For the Twitter-based datasets there were experiments run using: *Twitter-specific embeddings* and *ELMo embeddings*. In the experiments run on the Wikipedia dataset the words were represented either by the *fastText embeddings* or by *ELMo embeddings*.

3. ***Data Integration for Toxic Comment Classification***

Due to the overall evolution of toxic expressivity, a wider variety of labels and classifications were being discovered. Despite these numerous types, a great number of datasets have been released for training and testing data. The main throwback of this diversity is that it comes in different formats, with different class labels, such as racism, hate, abuse. Therefore, the stated paper [7] presents a collection of forty datasets in the form of a software tool, easing the classification of toxic comments. What is more, it proposes compatible datasets, languages, platforms and class labels in order to facilitate the training and testing process.

The current paper proposes a software that succeeds in easily accessing the labels based on a simplified API and a unified data that can be filtered with the use of metadata. The software has access to nearly thirteen languages and plenty of toxic behavioural labels (racism, sexism, aggression, hate etc.). What is more, the tool can be used to combine multiple datasets.

The unified dataset was based on a collection containing all subtypes of toxicity: offensive and abusive language, aggression and comments that would make the user leave a conversation. The tool excludes WhatsApp-type conversations, since the users are supposed to know each other. In order to integrate datasets retrieved from GitHub repositories, web pages, google drive, coming in various formats, the data was converted in a standardised csv format. Class labels are encoded in different ways and mapped so that labels with the same meaning are classified as the same label, remaining 126 labels. The main data source used was Twitter. Some of the datasets were retrieved from Twitter posts by filtering using specific key words or hashtags. Although there is a huge number of labels, the authors of the paper propose a way of splitting the data into binary annotated samples as non-toxic or toxic. In total there are 812,920 samples out of which 293,844 are considered toxic.

4. ***HateBERT***

HateBERT [8] is a recent model trained for the abusive language detection task. What is interesting about this model is that it combines semi-supervised learning with supervised learning in order to solve the task. More concretely, the authors of the paper propose in their approach a complete re-training step of the popular *BERT* model before fine-tuning it for our specific task.

HateBERT was first retrained using the original *Masked Language Model* objective on the *Reddit Abusive Language English dataset (RAL-E). RAL-E* is a custom dataset built by the authors of the paper that incorporates 1,492,740 comments retrieved from different banned subreddits. The subreddits chosen were banned due to heavily promoting abusive and hateful content, so most of the extracted comments represent use of abusive / hateful speech. The main idea behind the retraining is that in this way we will obtain a new flavour of *BERT* specially designed to be biassed to abusive language. This new model obtained is called *HateBERT* and it uses the original base version of *BERT*. The model was retrained for 100 epochs and it is made publicly available in order for others to use it.

Additionally, the authors of the paper perform fine-tuning experiments for the abusive language detection task. The experiments consist of fine-tuning simple classifiers for the task which use the classical *BERT* and *HateBERT* models. The idea behind these benchmarks is to compare the performance of the two models and showcase how *HateBERT* outperforms the original model. These experiments were run on three different datasets: OffensEval 2019 [9], AbusEval and HatEval [9]. All of these datasets consist of tweets and are quite imbalanced. The metric used to compare the two classifiers is the macro F1 score.

5. ***Abusive Language on Social Media Through the Legal Looking Glass***

This research paper [10] focuses on building a new dataset entitled *Abusive Language on YouTube* (*ALYT)* which focuses on YouTube comments. The dataset contains 20,215 annotated samples which represent comments left by users to different videos.

In order to build the dataset, the researchers chose three main controversial topics of videos that can spark abusive language in the comment section. These three topics are: Gender Identification, Veganism and Workplace Diversity. Additionally, for each topic there are selected videos from three different categories: personal videos, reaction videos and official videos. After the set of videos is established, all the comments left in the comment section to those videos are extracted. The comments that are replies or contain external links to other websites are discarded. From the remaining set of data the authors extract at random 20,215 samples and annotate them. This is done in order to better capture the real distribution. The data is annotated on a scale of grades ranging from 1 to 7, where higher the grade the more abusive is the comment considered. Comments which are annotated with scores between 1 and 3 are not considered abusive. In total, approximately 11.42% of the total samples are considered abusive comments.

The authors of the paper also present some classification experiments run on this dataset. They propose two baselines to classify the comments into *abusive* and *non-abusive* classes. These baselines are a logistic regression and a baseline based on *BERT*. The second

mentioned baseline uses a special flavour called *BERTweet*, which was trained on tweets. The metrics used to compare the performances of the model are: precision, recall and F1-score.

6. ***Conclusion***

During the research activity for the first milestone we identified multiple datasets that we could use, different approaches for solving the abusive language detection task and what metrics we could use. As far as we observed, most of the approaches focus only on one type of online social interaction such as: tweets, comments, conversations etc.

Following this milestone, we would like to set firstly as one of our goals the task to compare multiple models on how they perform on different online social interaction contexts. We want to experiment on the datasets described above, as they present a large variety of online environments, and to focus on the English language. As a baseline, we would like to experiment with SVMs, simple Neural Networks and Regressions. After building these baselines, the next step would be to build and replicate more advanced experiments such as the one which uses Bi-LSTMs and a new type of attention and others that use *BERT embeddings* such as *HateBERT*. This step can be seen as a first experiment iteration over the task. Our last goal would be to combine *HateBERT* with the *contextual attention* described above in order to see how it affects the performance of the model. We want to combine these two ideas as we found them really interesting and consider that they have a huge potential.

7. ***References***

[1]: Curry, A. C., Abercrombie, G., & Rieser, V. (2021, November). ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Detection in Conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 7388-7403).

[2]: Chakrabarty, T., Gupta, K., & Muresan, S. (2019, August). Pay “attention” to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online* (pp. 70-79).

[3]: Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop* (pp. 88-93).

[4]: Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 512-515).

[5]: Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., ... & Wu, D. M. (2017, June). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference* (pp. 229-233).

[6]: Wulczyn, E., Thain, N., & Dixon, L. (2017, April). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web* (pp. 1391-1399).

[7]: Risch, J., Schmidt, P., & Krestel, R. (2021, August). Data integration for toxic comment classification: Making more than 40 datasets easily accessible in one unified format. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 157-163).

[8]: Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2020). Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.
[9]: Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.
[10]: Bertaglia, T., Grigoriu, A., Dumontier, M., & van Dijck, G. (2021, August). Abusive Language on Social Media Through the Legal Looking Glass. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (pp. 191-200).