Course paper proposal subjectivity mining

Kamiel Gülpen and Dante de Lang

October 11, 2021

STUDENTNUMBERS 2718984, 2585032
ASSIGNMENT 1 Course paper
GROUP 6
COURSE Subjectivity mining

1 Project motivation

A big issue with machine learning models is the fact that a large amount operate as black boxes. This means that the users of these models know the inputs and outputs but do not know how the model comes to these conclusions. The implementation of such a model can have severe consequences when it is not implemented carefully[Pappada and Pauli, 2018]. Realistic problems can occur when implementing a machine learning model such as discrimination based on race, gender, religion etc.

As the BERT model grows in popularity it is important to know on which signal words they make a decision and label sentences. For this reason we are interested in the interpretability of different BERT models. An interesting paper which could help channeling our research is a comparison study by Fortuna et al. [2021].

In our project we want to research the interpretability of different BERT models on different datasets. We want to look at three popular BERT models, namely BERT [Devlin et al., 2019], hateBERT [Caselli et al., 2021] and BERTweet [Nguyen et al., 2020]. We want to furthermore train and test these BERT models on the following datasets: HateXplain and ethos. The goal of our study is to investigate whether the BERT models act in the same way throughout all the datasets and if there is a difference in interpretability between the BERT models

2 Data

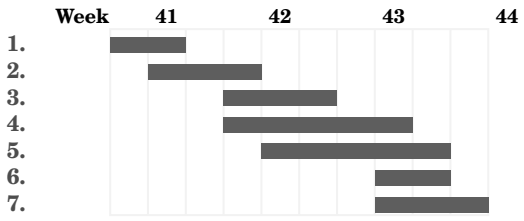
For the data we aim to use more than one dataset in order to assess the classification models in a more robust manner. One of the datasets in which we are interested is the HateXplain dataset [Mathew et al., 2020]. This dataset is presented as a benchmark dataset for explainable hate speech detection. They also worked with the LIME extension which we also want to explore within this research.

Another dataset that caught our interest is presented by Mollas et al. [2020] and contains labelled comments from Youtube and Reddit. The labelling was done using initial manual annotation which was validated using the Figure-Eight platform, in addition data was added using the Hatebusters dataset. The dataset can be used for binary and multi-label classification.

3 Project plan

Our plan is to use BERT models mentioned in section 1. For the models we will use the simpletransformer[1] and pytorch[2] packages. Our plan is to use all three models on the three datasets described in section2. We will furthermore compare the results retrieved by the models. Afterwards we want to do a interpretability analysis with the LIME [3] tool in order to interpret the model outcomes. We want to visualize the results to make them as insightfull as possible. Finally our plan is to discuss the obtained results and draw conclusion from them.

4 Time table



1. Proposal brainstorm
2. Executing a more elaborate literature search and fine-tuning plan of approach
3. Start writing the introduction data and methods section.
4. Start with building the pipeline for our final product.
5. Having the results ready and interpret them as well as implementing them into our project.
6. Discuss the results and write the discussion and conclusion section

[1] https://simpletransformers.ai/
[2] https://pytorch.org/
[3] https://github.com/marcotcr/lime

7. Reflect on our project, reread and rewrite when necessary.

References

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *ArXiv*, abs/2010.12472, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, June 2019. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Paula Fortuna, Juan Soler-Company, and Leo Wanner. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524, 2021.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*, 2020.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*, 2020.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*, 2020.

Roberta Pappada and Francesco Pauli. Discrimination in machine learning algorithms. 2018.