BSP 6- A deep learning-based method for Hate speech detection on social media content

Tuesday 23rd May, 2023 - 14:39

André Daniel Dussing
University of Luxembourg
Email: andre.dussing.001@student.uni.lu

This report has been produced under the supervision of:
Salima Lamsiyah
University of Luxembourg
Email: salima.lamsiyah@uni.lu

***Abstract*—The project is about the development of an approach to detect hate speech on social media content using a deep learning approach. With the rise of online harassment and hate speech, it has become crucial to identify and mitigate such harmful behaviors. To tackle this problem, the proposed method employs Natural Language Processing (NLP) techniques along with deep learning models to analyze short messages on social media and determine whether they contain hate speech. By utilizing deep learning algorithms, this approach aims to improve the efficiency and accuracy of hate speech detection, which can help in addressing this problem on social media platforms more effectively.**

1. Main objectives of the BSP

The main objective of my 6^{th} BSP is to efficiently detect hate speech given social media content. In order to achieve that goal we need to split the main objective into sub-goals and questions, which are related to those sub-goals. The sub-goals and questions for my BSP 6 will be:

1 Why do we do this? What are the ideas and reasons behind the project? (section 1).
2 Understand the (social media) data we want to classify. (section 3).
3 Think about the features we want to extract from our data set in order to classify our data. (section 3).
4 Understand and apply basic steps to solve a NLP problem:
   a Collection of data
   b Text pre-processing
   c Text representation
   d Machine Learning model selection and training
   e Evaluation of the selected model
   f Deployment the model
5 Look if there are already solutions, which are similar to our problem (section 3).
6 Analyse and understand how existing solutions work (section 3).
7 Try to use/implement the existing solutions to solve the problem (section 4).
8 Try increase the accuracy and efficiency and adapt the model to requirement changes made during the project. Iterate over Step 3 and 4. (section 4 and section 3).

Social media has undoubtedly had a significant impact on society, and there are many positive aspects of its rise. However there has also been an increase in negative aspects, particularly with regards to hate speech. Therefore, it is essential to prioritize hate speech detection on social media platforms. Detecting and removing hate speech is crucial not only for protecting individuals from harm but also for promoting a healthy and productive online discourse.

2. the main competencies necessary to master before you start the BSP

Before starting the project, it is essential to have a basic understanding of natural language processing (NLP) and machine learning concepts. Knowledge of programming languages, such as Python, and libraries commonly used in NLP, such as NLTK and SpaCy, is also needed. Additionally Understanding the basics of Neural networks and their applications in NLP will be helpful in developing an efficient hate speech detection model. It is also important to have a good understanding of the social media platform(s) where the model will be deployed, as this will affect the choice of data sources and features to be extracted.

3. Description of the scientific deliverable

The first step ,when trying to solve a problem , is to clearly define the problem statement. In this case we are looking for hate-speech. Hate is defined as ”public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation”. [3] .Despite the definition the same sentence can in one case be considered hate speech and in another case a normal comment depending on the situation. To tackle this ambiguity I read and analysed the paper ”Hate Speech and Counter Speech Detection: Conversational Context Does Matter” [5], which concludes that the context does matter in Hate Speech detection. In the paper they use a dataset from Reedit, which provides context in terms of preceding comments. At first I will use this dataset to train a machine learning model, although other datasets, which highlight different aspects than the conversational context might be used later on (see [2] [4]).When considering to select a language model architecture I choose BERT (Bidirectional Encoder Representations from Transformers),since it has been trained on a large corpus of text data and has demonstrated strong performance on various NLP tasks. In the final scientific description I will have a close look at BERT and explain how transformers work in general. The specific implementation as well as the pre-processing of the data ,training and evaluation of the model will be found in the technical section (see 4).

4. Description of the technical deliverable

As already mentioned in the scientific section BERT will be used to analyze the short messages in the Reddit dataset and to determine if those messages are neutral, counter hate speech or hate speech. More specifically I will use HATEBert, is a variant of the BERT model that is specifically designed for hate speech detection.HATEBert outperformed other state-of-the-art hate speech detection models on several benchmark datasets, demonstrating its effectiveness in identifying hate speech. The researchers also released a pre-trained HATEBert model, which can be fine-tuned on specific hate speech detection tasks with minimal training data. [1] I will use the Transformers library from Hugging Face. The library is built on top of the PyTorch and TensorFlow frameworks and offers easy-to-use interfaces for fine-tuning pre-trained models on custom datasets, as well as for building and training new models from scratch. In the technical deliverable I will follow the steps already mentioned in section 1 subgoal 4 (basic steps to solve a NLP problem).I will try to achieve the highest possible performance. Based on the fist evaluations I will consider adapting the process/planed steps (stemming, tokenization, word embeddings or the machine learning model and libraries). The evaluation will consider factors such as the time it takes to train the model, its scalability, and its overall performance, as measured by metrics like the F1-score.

References

[1] Tommaso Caselli et al. “HateBERT: Retraining BERT for Abusive Language Detection in English”. In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 17–25. DOI: 10.18653/v1/2021.woah-1.3.

[2] Thomas Hartvigsen et al. “ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3309–3326. DOI: 10.18653/v1/2022.acl-long.234. URL: https://aclanthology.org/2022.acl-long.234.

[3] *Hate speech*. URL: https://dictionary.cambridge.org/dictionary/english/hate-speech.

[4] Ioannis Mollas et al. “ETHOS: a multi-label hate speech detection dataset”. In: *Complex & Intelligent Systems* (Jan. 2022). ISSN: 2198-6053. DOI: 10.1007/s40747-021-00608-2. URL: https://doi.org/10.1007/s40747-021-00608-2.

[5] Xinchen Yu, Eduardo Blanco, and Lingzi Hong. “Hate Speech and Counter Speech Detection: Conversational Context Does Matter”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 5918–5930. DOI: 10.18653/v1/2022.naacl-main.433. URL: https://aclanthology.org/2022.naacl-main.433.

5. Appendix

All images and additional material go there.