Advance NLP: Hate Speech detection using Transformers (Deep Learning)

Group name: VerbalVigilantes

Team member names: Zohra Bouchamaoui

Email: zohra.bouchamaoui@outlook.com

Country: United Kingdom

Company: Data Glacier

Specialisation: NLP

PROBLEM DESCRIPTION

Hate speech is defined as any form of verbal, written, or behavioural communication that uses derogatory or discriminatory language to insult or attack an individual or a group based on attributes such as religion, ethnicity, nationality, race, colour, ancestry, gender, or other identity factors. In this project, our objective is to design a machine learning model, utilising Python, that can accurately detect instances of hate speech.

Hate speech detection typically falls under the umbrella of sentiment classification. To train a model capable of discerning hate speech in a given text, we will utilise a dataset commonly used for sentiment classification. Specifically, for this task, we will train our hate speech detection model using Twitter data, with the aim of identifying tweets that contain hate speech.

BUSINESS UNDERSTANDING

In the era of online communication, the prevalence of hate speech has risen dramatically. Such language, typically targeting an individual or group based on attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender, can lead to societal harm and discord. Businesses, particularly social media platforms, have a significant responsibility to monitor and moderate such content to maintain a safe and inclusive online environment for their users.

The task at hand – detecting hate speech using advanced Natural Language Processing (NLP) techniques and Deep Learning (Transformers) – presents a considerable opportunity from a business perspective. Implementing a successful hate speech detection model can help businesses proactively flag and remove harmful content, thus protecting their user base and upholding community standards.

For social media companies like Twitter, which is our data source for this project, an efficient hate speech detection system can be crucial. Twitter, with its large volume of user-generated content, needs a reliable way to identify and moderate hate speech. Manual moderation for such a massive amount of data is not feasible, and therefore, a machine learning-based solution can be of great value.

Moreover, this project can help businesses in ensuring legal compliance. Various countries have strict regulations against hate speech, and companies failing to regulate such content can face legal consequences. Therefore, an effective hate speech detection model can help in avoiding potential legal issues and maintain a positive brand image.

Lastly, the solution can be adapted to various languages and regional nuances, making it a versatile tool for global platforms. This project can also pave the way for more complex NLP tasks, such as detecting more subtle forms of toxicity online, like cyberbullying or misinformation.

In summary, the successful execution of this project will enable businesses to maintain a healthier online community, comply with legal requirements, protect their brand image, and potentially expand their user base by promoting a safer digital environment.

PROJECT LIFECYLE & DEEADLINES

This project can be broken down into the following phases:

Week 1 (Due 19 June 2023): Project Initialisation and Setup

During this phase, we'll gather all necessary details about the team, develop a detailed problem statement, establish a clear business understanding, and create an initial project lifecycle with deadlines. We will also generate a preliminary data intake report, set up our GitHub repository, and start familiarising ourselves with the Twitter dataset used for this project.

Week 2 (Due 26 June 2023): Data Understanding and Issues Identification

This week involves deeply understanding the data that we have for analysis. We will identify potential issues within the data such as NA values, outliers, and any skewed features. We will also decide on and document the approaches that we will take to mitigate these issues. Progress updates will be made to our GitHub repository.

Week 3 (Due 2 July 2023): Data Cleansing, Transformation, and NLP Pre-processing

In this phase, we will clean, transform, and pre-process the data, focusing on tackling the issues identified in Week 2. For NLP pre-processing, we will be employing techniques like regular expression cleaning and different featurisation methods.

Week 4 (Due 9 July 2023): Exploratory Data Analysis (EDA) and Feature Selection

We'll perform an EDA on the cleansed and transformed data, gathering insights about the data that could be important for our model. We will document our findings and select the features to be used in the model based on these insights.

Week 5 (Due 16 July 2023): Model Selection, Training, and Recommendation

During this week, we will select our base Transformer model, train it on the data, and optimise it. We will also develop an EDA presentation for business users and technical users, recommending the most suitable models based on our findings.

Week 6 (Due 23 July 2023): Model Refinement and Performance Evaluation

In this week, we will refine our models, try different configurations, and evaluate their performances. We will ensure that the models align with the business requirements. This will also involve developing a comprehensive report and updating our GitHub repository.

Week 7 (Due 30 July 2023): Project Finalization and Presentation

In the final week, we will wrap up the project by reviewing and selecting the best-performing model. We will finalise our report and deliver a PowerPoint presentation summarising our project and findings. All code and deliverables will be updated and finalised in the shared GitHub repository.