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

Batch code: LISUM21

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.

DATA UNDERSTANDING

For analysis, we have two datasets: train_tweets and test_tweets. The train_tweets dataset consists of three columns: id, label, and tweet. It contains 31,962 rows of data. The id column represents the unique identifier for each tweet, the label column indicates whether the tweet contains hate speech (1) or not (0), and the tweet column contains the actual text of the tweet. The test_tweets dataset has two columns: id and tweet, with 17,198 rows.

1. **Data Structure:** Both the train_tweets and test_tweets datasets are structured as tabular data, where each row represents a unique tweet and each column represents a feature of the dataset.
2. **Target Variable:** In the train_tweets dataset, the 'label' column serves as the target variable, indicating whether a tweet contains hate speech (1) or not (0). This variable will be used for training and evaluating the performance of the machine learning model.

3. **Text Data:** The main feature of interest in the datasets is the 'tweet' column, which contains the actual text content of the tweets. This text data will be pre-processed and analysed to extract meaningful insights and patterns.
4. **Missing Values:** It is important to examine if there are any missing values in the datasets. Missing values can impact the analysis and modelling process, and appropriate strategies need to be employed to handle them effectively.
5. **Outlier and Skewed analysis:** To be able to perform these analysis on the train dataset, we need to focus on the numerical columns, which in this case is only the 'label' column. The 'id' column is an identifier and the 'tweet' column contains text data. Since the 'label' column is binary (0 or 1), it does not make sense to perform outlier analysis as there are only two possible values. Skewed analysis is also not applicable to binary data.

TYPE OF DATA

For the Twitter data analysis, we have textual data in the form of tweets. Each tweet is a sequence of characters, which can include text, hashtags, mentions, punctuations, and emojis. The data also includes additional columns such as 'id' and 'label', which provide unique identifiers for each tweet and indicate whether the tweet contains hate speech or not.

The data is structured in a tabular format, where each row represents a single tweet and each column represents a specific attribute or feature of the tweet. The 'id' column serves as a unique identifier, the 'label' column provides the classification label, and the 'tweet' column contain the actual text content of the tweet.

PROBLEMS WITH DATA & APPROACHES TO SOLVING THEM

Based on the type of data we have, we can identify some potential problems:

1. **Class Imbalance:** It is important to examine the distribution of labels in the 'label' column to check if there is a significant class imbalance between tweets containing hate speech (label=1) and those that do not (label=0). Class imbalance can pose challenges in model training and evaluation, as it may lead to biased results and affect the performance of the analysis.

   The label distribution in the train dataset indicates that there are 25,917 instances labelled as ‘Positive’ and only 6,045 instances labelled as ‘Negative’. This significant disparity in class frequencies can impact the performance and accuracy of machine learning models trained on this dataset.

2. **Noise and Irrelevant Information:** Twitter data often contains noise in the form of typos, abbreviations, slang, special characters, URLs, and mentions. Additionally, not all tweets may be relevant to the analysis at hand. It is important to preprocess the data and remove or handle these noisy elements appropriately to ensure accurate analysis.

   For this, we can perform an exploratory data analysis to gain insights into the dataset (analyse the distribution of tweets, examine different features, look for unusual patterns or inconsistencies). Additionally, text pre-processing techniques can be applied to clean the textual data. This may consist of removing special characters, punctuation, URLs, and unnecessary whitespace. By standardising the text, we can reduce noise and focus on the relevant information.

   Another approach to take is Stopword Removal. We can use a predefined list of stopwords or NLTK’s stopwords corpus to remove these words from our text data.

3. **Missing Data:** The dataset may contain missing values in any of the columns, such as 'id', 'label', or 'tweet'. It is crucial to identify and handle missing data appropriately before proceeding with the analysis. Missing data can affect the quality and reliability of the results if not addressed properly.

In our dataset, there are no missing values so we will not be applying any techniques to fill in missing values.

4. **Outliers in Textual Data:** Although outliers are typically associated with numerical data, in the case of textual data, outliers can refer to extreme or unusual tweets that deviate significantly from the general patterns observed in the dataset. Identifying and handling such outliers, if present, is important to ensure the robustness of the analysis.

   One approach to identify outliers is to analyse the length of the text in the dataset. These may include extremely short or excessively long texts compared to the typical length of the majority of texts. This can be visualised using a histogram or box plot.

   Another approach is vocabulary analysis. Outliers may consist of rare or unique words that appear very infrequently compared to the rest of the dataset. This can be done by calculating the word frequency distribution and identifying words that occur significantly less than others.

   Topic Modelling, such as Latent Dirichlet Allocation (LDA) or Non-Negative Matric Factorisation (NMF) could also be used to discover latent topics in the textual data.

   One more approach is sentiment analysis. This approach will be used to identify outliers based on sentiment polarity. These could include texts with extremely positive or negative sentiments compared to the majority of the dataset.