uOttawa

Proposal and Problem Formulation

Toxic Wikipedia Comments Classification

Group ID: 5

DTI5125[EG] Data Science Applications [LEC] 20235

Submitted By:

Esraa Fayad
Ahmed Nasser
Aya Metwally
Abdelrahman Ali

University of Ottawa, Canada

July 15, 2023

Problem Formulation

Problem: Developing a classifier that can accurately identify/recognize toxic comments. Given a dataset of toxic comments.

Challenge: The shortage of data., so Insufficient data prevents the development of accurate classifiers.

Solution: Use data augmentation techniques to increase the size of the dataset. This will allow us to train a classifier that is more accurate.

Pipeline: apply the augmentation techniques that will mention in the Methodology part to increase the size of the dataset and apply our classification algorithms to determine whether this is a toxic comment or not then if this is a toxic comment then will determine the level of toxicity to build our models, after building all models we will deploy it with simple GUI using Python.

Methodology

The dataset that was utilized to construct the classification model was obtained from the competition website and comprised both the training and testing datasets. We have six different classes for the label: toxic, severe toxic, obscene, threat, insult and identity hate. To classify the datasets into the 6 given classes, we started from data pre-processing and feature engineering and ran three learning algorithms (Logistic Regression, Support Vector Machines, and Bidirectional LSTM). We also used two data augmentation algorithms (Easy Data Augmentation and Backtranslation). Three different models are then used with their results compared. Easy Data Augmentation (EDA): This technique involves making simple changes to the text of a comment, such as changing the capitalization, adding or removing punctuation, or swapping words with synonyms. Backtranslation: This technique involves translating a comment from one language to another and then translating it back to the original language. This can help to identify patterns in the text that are not obvious in the original language. We can use scikit-learn and Keras with TensorFlow backend for training LR, SVM, and Bi-LSTM models, and the use of Google Translate API for Backtranslation.

After applying our classification algorithms to determine whether this is a toxic comment or not then if this is a toxic comment then will determine the level of toxicity to build our models.

We will deploy our models in a GUI python Framework like PQT5 to build the GUI and generate a simple interface to be tested by some users.

Data Description and Data Sources

Our dataset is the "Wikipedia Toxic Comments" dataset which we obtained from Kaggle through the Kaggle Comment Classification Challenge. The dataset contains comments along with corresponding labels indicating different types of toxicity, such as threats, insults, obscenity, identity-based hate, and more. The aim of the challenge is to develop machine learning models that can effectively categorize comments and accurately predict their toxicity levels. Analyzing this dataset

can provide valuable insights into understanding and addressing the issue of online toxicity, enabling the development of systems to promote healthier and safer online interactions.

Data sources from several platforms These platforms could include social media platforms, discussion forums, news article comment sections, or any other online spaces where users can interact and leave comments.

The dataset aims to capture real-world instances of toxic language and offensive behavior commonly found in online discussions.

Evaluation Methods

Firstly, we apply the F1 score to evaluate the performance of our algorithms (SVM – BI LSTM - Logistic Regression). We use a subset of our dataset and apply data augmentation techniques (EDA – BT) to demonstrate our ability to achieve high results with a small amount of data. Additionally, we display all the important words associated with each model.

Then, we utilize TSNE to visualize our dataset in the most effective manner. Afterward, we apply classification algorithms to the entire dataset to determine if the comments are toxic or not. We assess the performance of the algorithms using a Confusion Matrix and Classification Report, measuring F1-score and accuracy. Finally, we display the important words associated with each model.

Finally, if a comment is determined to be toxic, algorithms will produce the level of toxicity. To evaluate this part of the algorithms, we utilize multiple classes Confusion Matrix, Classification Report, measuring F1-score and accuracy. Subsequently, we display the important words associated with each model.

Results Expectations

Our project's results include demonstrating the performance of the algorithms with data augmentation, visualization of the dataset, accurate classification of toxic comments, and assessment of the toxicity level through clustering algorithms. Additionally, the important words associated with each model are displayed, providing insights into the identified toxic comments.

References

Rastogi, C. (2020, July 2). Can We Achieve More with Less? Exploring Data Augmentation for Toxic Comment Classification. arXiv.org.https://arxiv.org/abs/2007.00875.

Wang, Z. (2022, February 15). Toxic Comments Hunter: Score severity of toxic comments. arXiv.org. https://arxiv.org/abs/2203.03548v1.

Zhao, Z. (2021, September 6). SS-BERT: Mitigating identity terms bias in toxic comment classification by utilizing the notion of "Subjectivity" and "Identity terms." arXiv.org. https://arxiv.org/abs/2109.02691v1.

Toxic Comment Classification Challenge.https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data.