PUTTING THE 'SOCIAL' BACK IN SOCIAL MEDIA

CLASSIFYING TOXIC COMMENTS FOR A HEALTHIER ONLINE COMMUNITY

Toxic comments are a growing concern in online communities and can negatively impact individuals and communities. The rapid growth of social media has led to the spread of toxic content, including hate speech, cyberbullying, and malicious attacks. As a result, there is an increasing need for methods to automatically detect and identify such behaviour in online discussions.

This project aims to address these knowledge gaps by conducting a thorough analysis of existing methods and proposing new approaches to solving this problem.

In this way, the study contributes to the development of more effective and efficient methods for identifying toxic comments and mitigating their impact.

AUTHORS

DIXIT KAMAL
UDIT SHARMA
SATISH MAHABHASHYAM

TRENT UNIVERSITY

INTRODUCTION

Toxic comment classification is an important task in natural language processing, with applications in social media monitoring, online community management, and content moderation.

In this poster, the research on toxic comment classification is conducted, which involves exploring various machine-learning algorithms and feature engineering techniques to build accurate and efficient models for this task.

OBJECTIVE

Toxic comment classification is a complex task, and there is a lack of research and understanding in this area.

The goal of the Toxic Comment Classification project is to create a machine-learning model that can classify online comments into different categories based on their content using NLP Techniques.

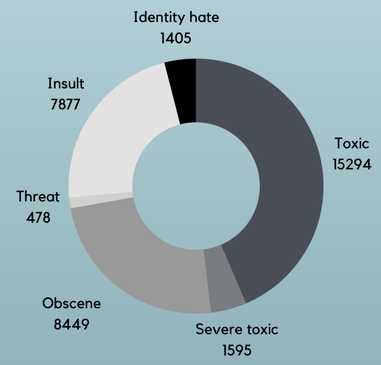
METHODOLOGY

Our approach involves pre-processing and feature extraction, including converting comments to numerical representations for machine-learning models.

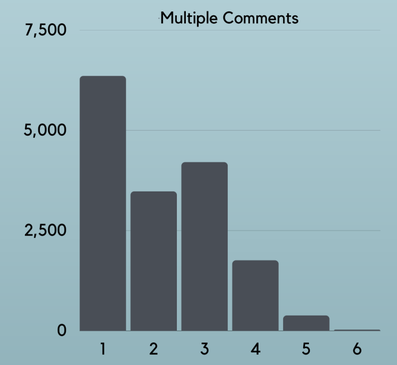
The model performance is compared using accuracy, precision, recall, F1-score, and Confusion Matrix. The study aims to determine the best model for predicting online comment toxicity levels.

ANALYSIS

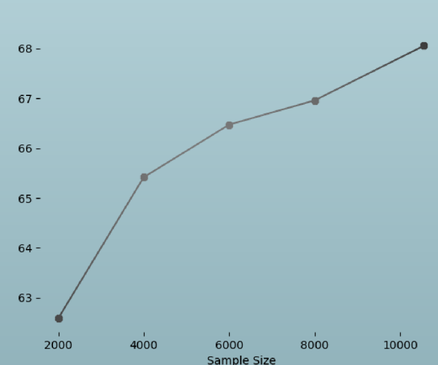
- **Data Preprocessing**
  - Removing
    - Irrelevant data
    - Duplicate Rows
    - Null Values
    - URL, Usernames & Emails
    - HTML Tags
    - Punctuations
    - Special characters
    - Digits
    - Stopwords
  - Lemmatizing Text
- **Machine Learning Model**
  - Text Vectorization
  - Undersampling dataset
  - NLP Techniques (TF-IDF)
  - Logistic Regression



Comments labelled under one category



Comments labelled under multiple category



Accuracy versus Sample Size chart

SAMPLE COMMENT: YOU BASTARD! I WILL KILL YOU!

Results after implementing NLP technique TF-IDF with Multilabel Logictic Regression:

RESULTS/FINDINGS

1. The accuracy of the model is ~ 73%.
2. The ROC area under the curve is ~94%.
3. Comment categorized as toxic? Chances are it's also obscene or insulting. These classes are strongly correlated.
4. Threats and Severe Toxicity stand-alone, least correlated of all. So, if it's one, it may not be the other.

The data used in the project is the labelled dataset from kaggle competition: Jigsaw Comment Classification.

CONCLUSION

The EDA included determining the number of comments in each category, including clean comments that were subsequently under-sampled to 25,000 from 143,346.

The TF-IDF technique was utilized in conjunction with the Multilabel Logistic Regression machine learning model to train on this data. Finally, the performance of our model was evaluated using metrics such as accuracy and ROC on the test data.

DATASET DESCRIPTION

The dataset consists of the following attributes:

1. Id: a unique identifier for each comment
2. Comment_text: the text of the comment
3. Toxic: The toxicity of comments
4. Severe_toxic: Intensity of toxicity
5. Obscene: offensive comment
6. Threat: Involves threats in the comment
7. Insult: Involves insults in the comment
8. Identity_hate