Toxic Comments Classification

Project Progress and Milestones

Siva Swaroop Vardhineedi - 016835312
Sri Charan Reddy Mallu - 017419779
Mahek Virani - 017446936
Sohan Leburu - 017408768

Introduction

Our project focuses on identifying and categorizing toxic comments across various digital platforms into six toxicity levels: toxic, severe toxic, obscene, threat, insult, and identity hate. Utilizing Natural Language Processing (NLP) techniques and machine-learning models, we aim to automate the moderation process to maintain a healthy online community environment. The performance of our models is measured using the F1-score to ensure accuracy and reliability.

Achievements So Far

Data Acquisition and Preprocessing:

- Successfully sourced the dataset from the Kaggle Toxic Comment Classification Challenge, consisting of 159,571 Wikipedia comments labeled for toxicity.
- Implemented data visualization techniques to understand the distribution of classes, revealing a significant imbalance among them.
- Preprocessed the data through normalization, punctuation removal, word tokenization, stop words removal, and IP address stripping using TfidfVectorizer for vectorization.

Model Development and Evaluation:

- Explored and implemented three machine-learning models: Logistic Regression, Support Vector Machines (SVMs), and Multinomial Naïve Bayes, splitting the dataset into training, validation, and test sets (60:20:20).
- Tuned the models for optimal performance with specific combinations of n-grams and max features, achieving our best results with 20,000 unigram and bigram word features combined with 10,000 tri, quad, and pent-grams.

GUI Development:

- Developed a Graphical User Interface (GUI) using the PyScript Python library, HTML, CSS, and JavaScript, enabling real-time classification of custom comments and model evaluation visualization.

Baseline Modules

TfidfVectorizer:

Serves as a critical preprocessing module, transforming textual data into a suitable format for machine-learning models. This vectorization allows for the removal of less informative text parts and normalization of the data.

Logistic Regression, SVM, and Multinomial Naïve Bayes Models:

These models form the backbone of our project, enabling the classification of comments into various toxicity levels. Their performance and tuning are crucial for the project's success, with SVM showing the best performance based on F1 scores.

GUI:

Our GUI acts as the project's interface, allowing users to interact with our models. It enables the demonstration of the models' capabilities in real-time, enhancing user experience and model accessibility

Challenges Encountered

Data Imbalance: Significant class imbalance presented challenges in model training and accuracy. Strategies like adjusting class weights and oversampling are being considered to mitigate this.

Model Selection and Tuning: Finding the right balance between model complexity and performance was challenging. Ongoing efforts include experimenting with additional models and tuning parameters for better accuracy.

Plans Moving Forward

Enhancing Model Accuracy: Continue refining our models, particularly addressing the challenges posed by data imbalance and exploring advanced techniques like ensemble methods.

Expanding Dataset Sources: To improve model robustness, we plan to integrate datasets from other social media platforms, ensuring our models can generalize well across different content types.

Neural Network Exploration: With increased computational resources, we aim to experiment with neural network architectures, such as RNNs or pre-trained models like BERT, for potentially superior performance

References

[1] J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, and Y. Burov, “Information resources processing using linguistic analysis of textual content,” in 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 2, 2017, pp. 573–578.
[2] A. A. Putri Ratna, A. Kaltsum, L. Santiar, H. Khairunissa, I. Ibrahim, and P. D. Purnamasari, “Term frequency-inverse document frequency answer categorization with support vector machine on automatic short essay grading system with latent semantic analysis for japanese language,” in 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), 2019, pp. 293–298.
[3] K. Dubey, R. Nair, M. U. Khan, and P. S. Shaikh, “Toxic comment detection using lstm,” in 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), 2020, pp. 1–8.

Source Code Links

GitHub Repository for Project Code

Additional Resources

Dataset used for the project:
https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data