**CS101 project report**

**Team member: James Ma, Xianmai Liang, Ao Chan, Shichao He**

**Executive Summary**:

Our project embarked on the crucial task of mitigating online toxicity by crafting an advanced comment classifier, specifically tailored for pinpointing severe toxic comments. Utilizing a suite of cutting-edge machine learning algorithms—Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest. Our models underwent rigorous evaluation, predominantly utilizing the F1 score as a benchmark. Remarkably, Logistic Regression reached a score of **0.9402**, KNN at **0.8614**, SVM notched up **0.9379**, and Random Forest excelled with a score of **0.9411** in identifying **severe** toxic comments.

**Introduction**:

Our strategy encompassed the deployment of a diverse array of machine learning models, including Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest. The core objective was to fine-tune these models to efficiently and accurately flag severe toxic comments, which are typically laden with aggressive and offensive content.

The F1 scores were pivotal in evaluating our models, with Logistic Regression, KNN, SVM, and Random Forest demonstrating their prowess through scores of 0.9402, 0.8614, 0.9379, and 0.9411, respectively. These figures are indicative of our models’ proficiency in balancing precision and recall, a critical factor in accurately identifying severe toxic comments while effectively reducing the instances of false positives and negatives.

In this report, we delve deeper into the intricacies of our methodologies, the experimental setup, and the outcomes achieved. We not only discuss the strengths and limitations of each model but also extend our analysis to include other types of comments.

**Data collection:**

Each sentence will be given 1 and 0 to their toxicity. 1 means it’s toxic, 0 means not.

The dataset is focused on the classification of toxic online behavior. It includes the following columns:

**id**: A unique identifier for each comment.

**comment\_text**: The text of the comment.

**toxic**: A binary indicator (0 or 1) showing whether the comment is toxic.

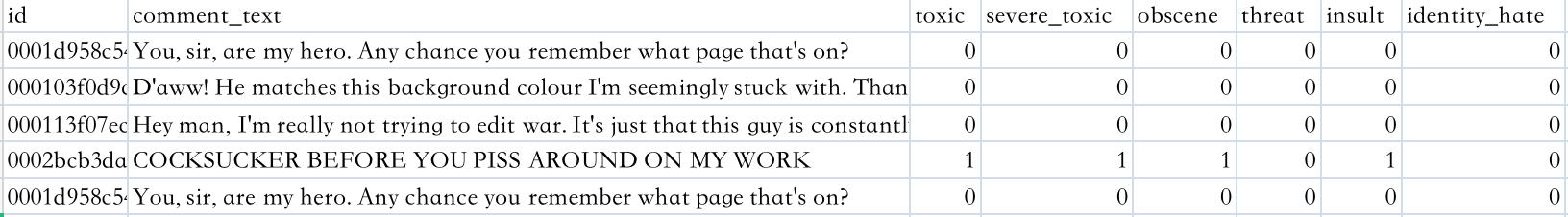
**severe\_toxic**: A binary indicator for severely toxic comments.

**obscene**: A binary indicator for obscenity in the comments.

**threat**: A binary indicator for threats present in the comments.

**insult**: A binary indicator for insults in the comments.

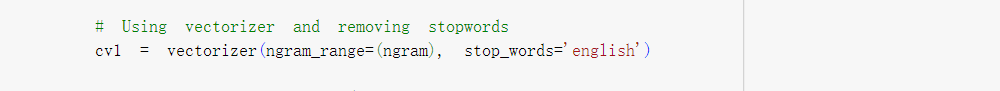
**identity\_hate**: A binary indicator for comments that express hate towards a particular identity.



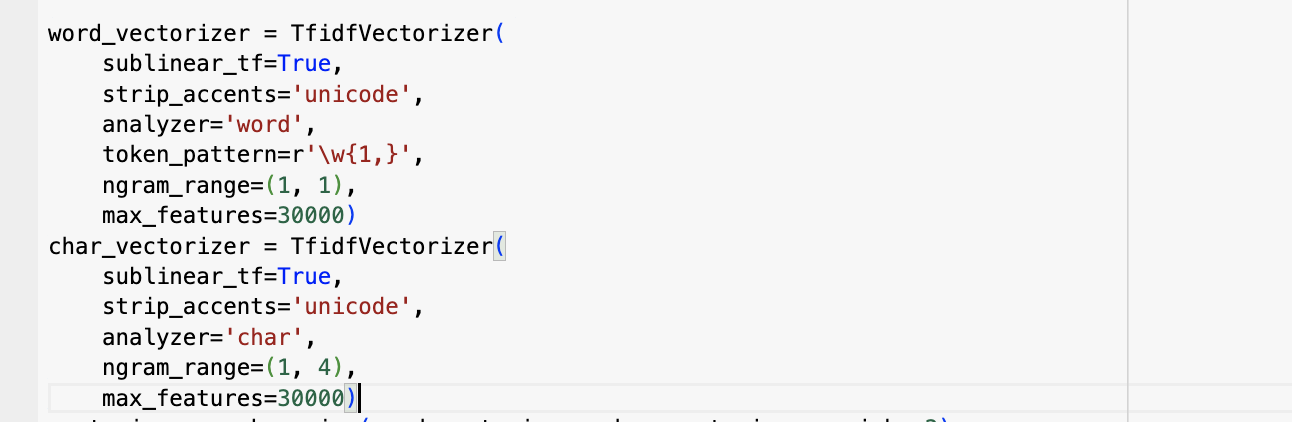
The presence of multiple specific toxicity indicators (like severe toxicity, obscenity, threats, insults, and identity hate) allows for a nuanced understanding and classification of online comments. This can be particularly useful in moderating online forums or social media platforms to maintain a healthy and respectful discussion environment.

**Preprocessing:**

Word2Vec Embeddings for Text Representation



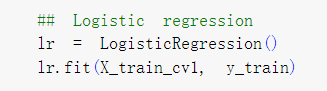
Also tried tf-idf:



**Experiment:**

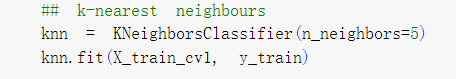
**Logistic Regression:**

Logistic regression is a statistical method used for binary classification, where the outcome variable is categorical and has only two possible values, typically denoted as 0 and 1. The logistic regression model predicts the probability that the dependent variable belongs to a particular category. It's widely used in various fields, including medicine, economics, and machine learning.



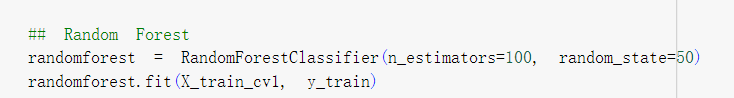
**KNN:**

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for classification and regression tasks. It is a simple and intuitive algorithm that works based on the similarity of data points. The basic idea behind KNN is to classify a new data point based on the majority class or average value of its k nearest neighbors in the feature space.



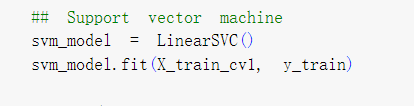
**Random Forest:**

Random Forest is an ensemble learning method used for both classification and regression tasks. It operates by constructing a multitude of decision trees during training and outputs the class that is the mode (classification) or the mean (regression) of the individual trees.



**Support Vector Machine:**

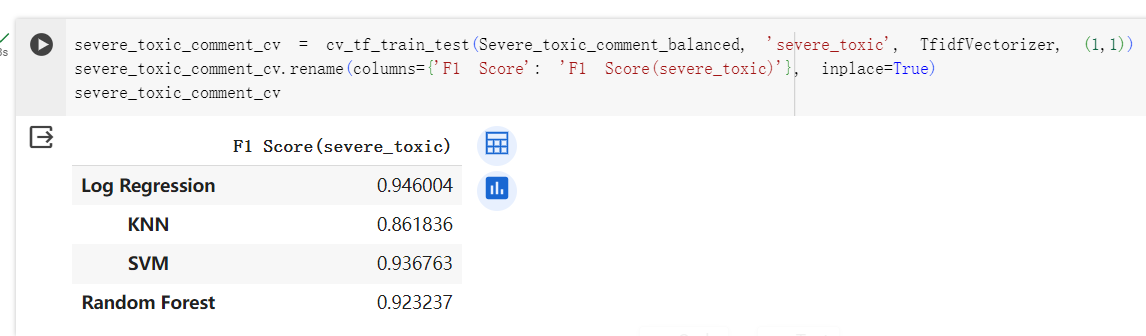
Support Vector Machine (SVM) is renowned for its efficacy in classification tasks. Its ability to delineate decision boundaries by identifying optimal hyperplanes makes it an indispensable tool in scenarios demanding precision in separating diverse classes within a dataset.



**Conclusion:**

**A. Model Accuracy and Effectiveness**

The analysis focused on four machine learning models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest. The evaluation metric used, F1 Score, revealed the following results for the classification of severe toxic comments:



These scores indicate that Logistic Regression achieved the highest F1 Score, signifying its effectiveness in classifying severe toxic comments.

The analysis provided valuable insights into the performance of different models in the context of toxic comment classification. Logistic Regression emerged as the top-performing model, showcasing its suitability for the task.

**B. Implications and Applications**

Potential Use Cases Beyond Toxic Comment Classification

While the primary focus was on toxic comment classification, the high-performing Logistic Regression model holds potential for broader applications in sentiment analysis, content moderation, and online discourse monitoring.

Contribution to Online Safety and Community Moderation

The findings contribute to enhancing online safety by providing a robust model for identifying severe toxic comments. The application of such models can aid community moderators in efficiently managing and maintaining healthier online discussions.

**C. Future Work**

Identified Areas for Improvement

Despite the success of the Logistic Regression model, areas for improvement were identified in the performance of KNN and Random Forest models. These areas may include refining feature selection, optimizing hyperparameters, or exploring alternative algorithms.

Suggestions for Future Research and Model Enhancements

Future research endeavors could delve into enhancing the model's interpretability, understanding the contextual nuances of toxic comments, and developing ensemble models that combine the strengths of multiple algorithms.