DATASET DESCRIPTION

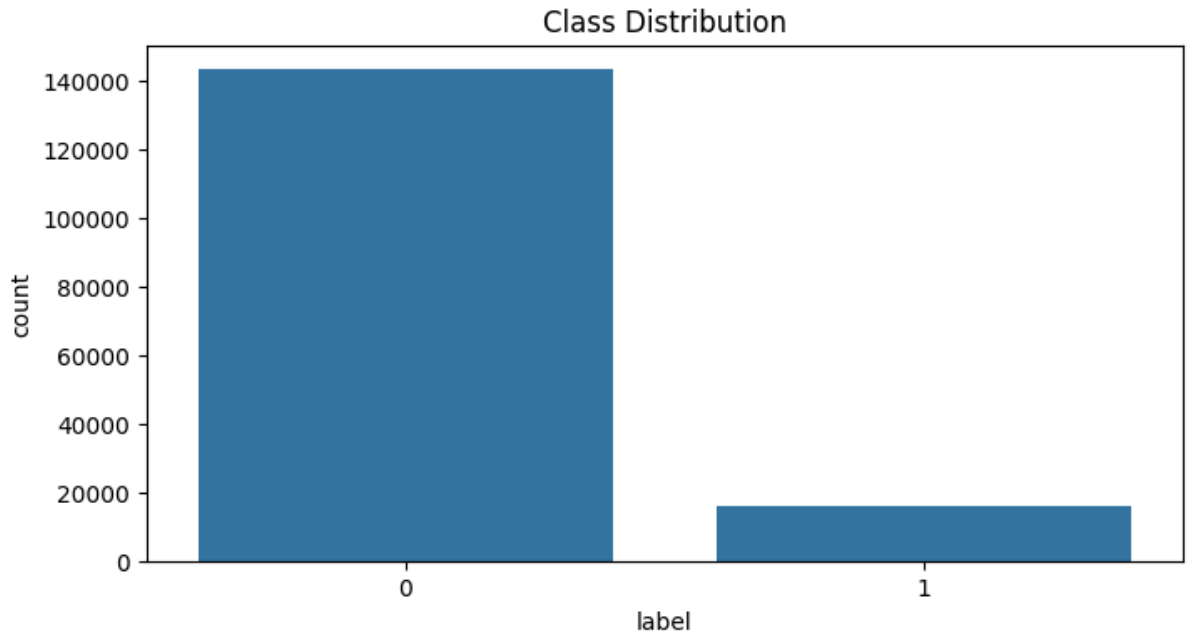
We are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of toxicity are:

- toxic
- severe_toxic
- obscene
- threat
- insult
- Identity_hate

EXPLORATORY DATA ANALYSIS

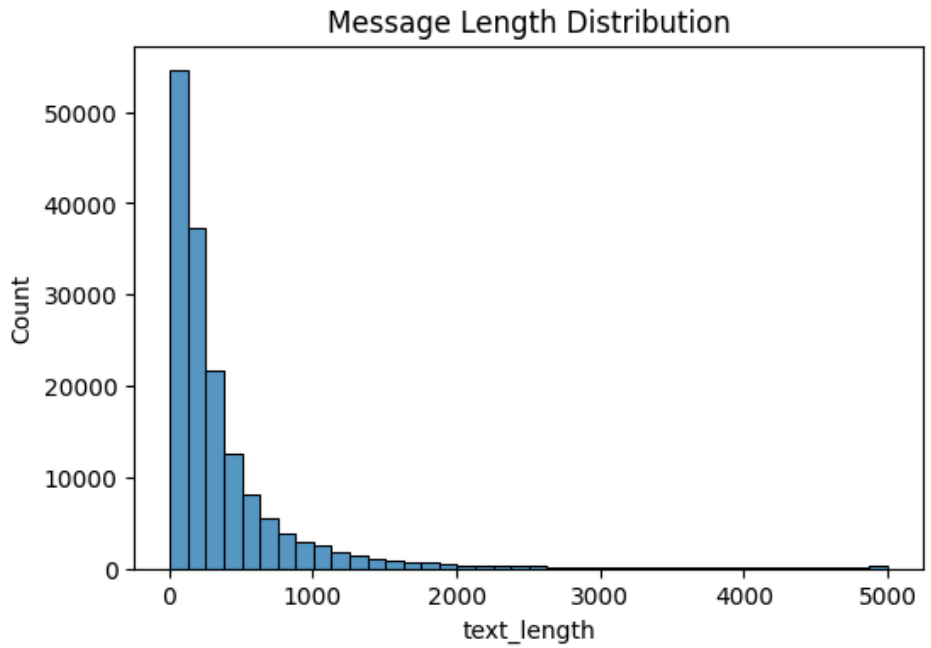
1. Class Distribution

The dataset is highly imbalanced. Most messages belong to class **0 (non-toxic/normal)**, while only a smaller portion belongs to class **1 (toxic)**. This indicates that toxic messages are relatively rare compared to normal ones. Because of this imbalance, special care may be needed during model training to avoid bias toward the majority class.



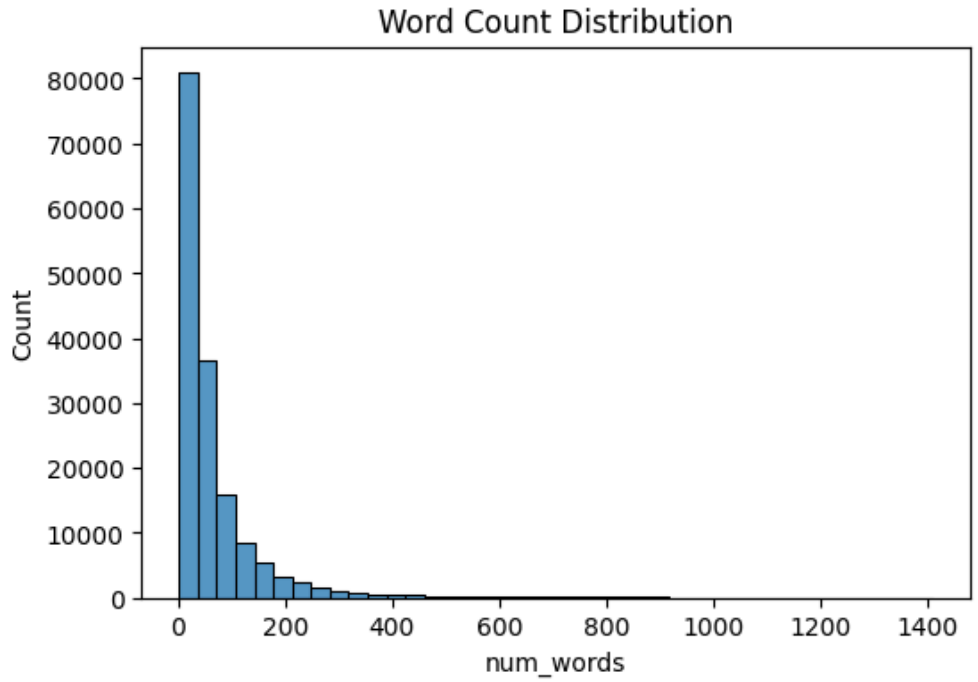
2. Message Length Distribution

Most messages are short, with text length concentrated at the lower end. A few messages are significantly longer, creating a right-skewed distribution. This shows that typical user comments are brief, while only a small number contain long detailed text.



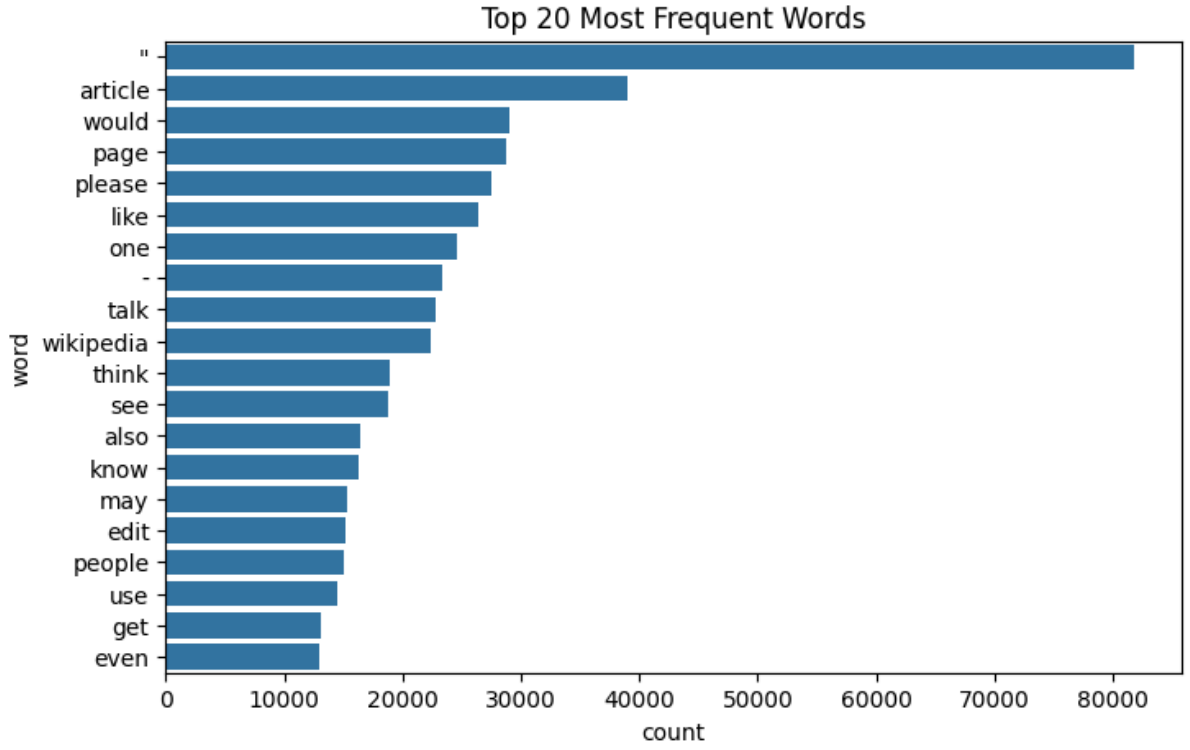
3. Word Length Distribution

The majority of comments contain very few words, typically under 50. Only a small number of comments are long, resulting in a right-skewed distribution. This indicates that most users prefer short, concise messages, while longer comments are rare outliers.



4. Top 20 Most Frequent Words

The most common words include neutral conversational terms such as “article”, “page”, “please”, “like”, and “wikipedia”. These words are general-purpose and appear frequently across both toxic and non-toxic comments. This suggests that keywords alone may not be enough — context is important when detecting toxicity.



PREPROCESSING

Before training the model, several preprocessing steps were applied to clean and standardize the text data. First, all characters were converted to lowercase to ensure that words like “Bad” and “bad” were treated identically. Next, punctuation marks, numbers, and special characters were removed because they do not contribute meaningful information for toxicity detection. Stopwords such as “is”, “the”, and “and” were removed to reduce noise and focus on important words. Finally, the cleaned text was converted into numerical form using the TF-IDF vectorization technique. TF-IDF assigns higher weights to words that are important in toxic comments and lower weights to very common words. These preprocessing steps helped improve model accuracy and efficiency by reducing redundancy and highlighting informative patterns in the text.

MODEL EXPLANATION

For this project, a Logistic Regression model was used for classification. Logistic Regression is widely used for text classification problems because it is simple, fast, and works well with

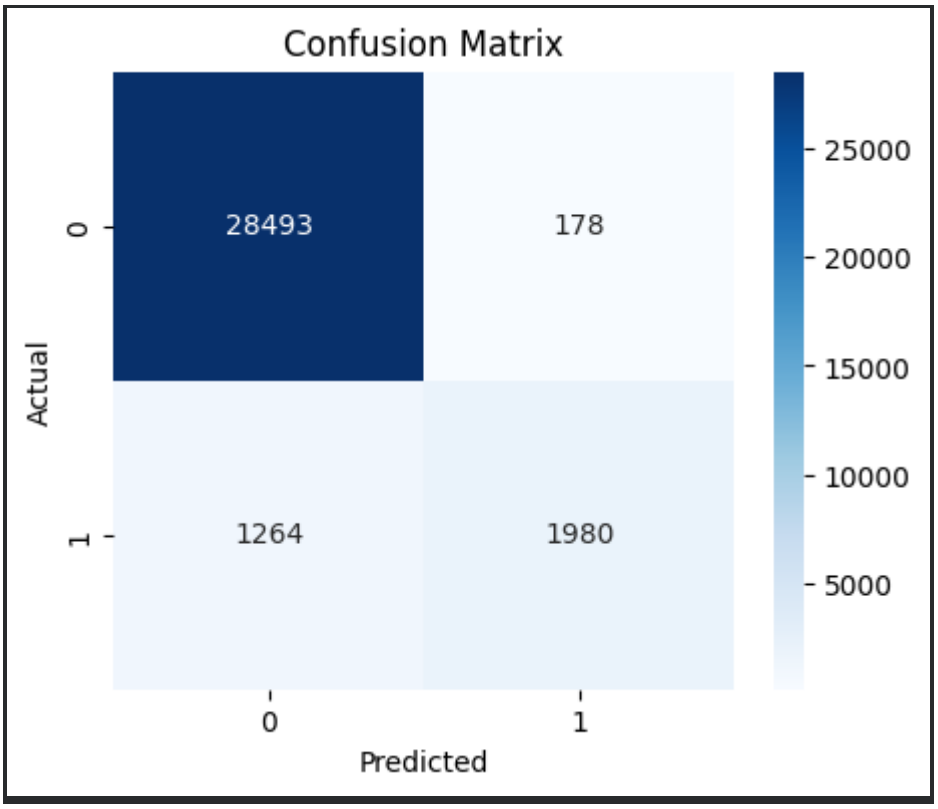
high-dimensional data such as TF-IDF features. The model learns relationships between words and toxicity labels by estimating probabilities for each class (toxic vs. non-toxic). If the probability exceeds a certain threshold, the message is classified as toxic; otherwise, it is labeled as non-toxic. Compared to more complex deep-learning models, Logistic Regression requires fewer computational resources while still achieving strong performance, making it an appropriate choice for real-time deployment.

EVALUATION WRITE-UP

```
Accuracy: 0.9548174839417202
Precision: 0.917516218721038
Recall: 0.6103575832305795
F1 Score: 0.7330618289522399
```

```
Classification Report:
```

	precision	recall	f1-score	support
0	0.96	0.99	0.98	28671
1	0.92	0.61	0.73	3244
accuracy			0.95	31915
macro avg	0.94	0.80	0.85	31915
weighted avg	0.95	0.95	0.95	31915



EXPERIMENTAL TESTING

Experimental testing showed that the model performs well on clean sentences but occasionally misses toxic comments that use slang, negation, or masked spellings (e.g., ‘stup1d’). Adding punctuation or intensity sometimes increases the probability of toxicity. These experiments highlight that context and informal writing styles significantly affect model behavior.

```
print("---- Noisy Text ----")
check("You are stup1d!!!")
check("Y0u are s0 dumb")
check("!!!!!!!")
check("....")
```

```
---- Noisy Text ----
Input: You are stup1d!!!
Prediction: Non-Toxic

Input: Y0u are s0 dumb
Prediction: Toxic

Input: !!!!!!!
Prediction: Non-Toxic

Input: ....
Prediction: Non-Toxic
```