Graphs

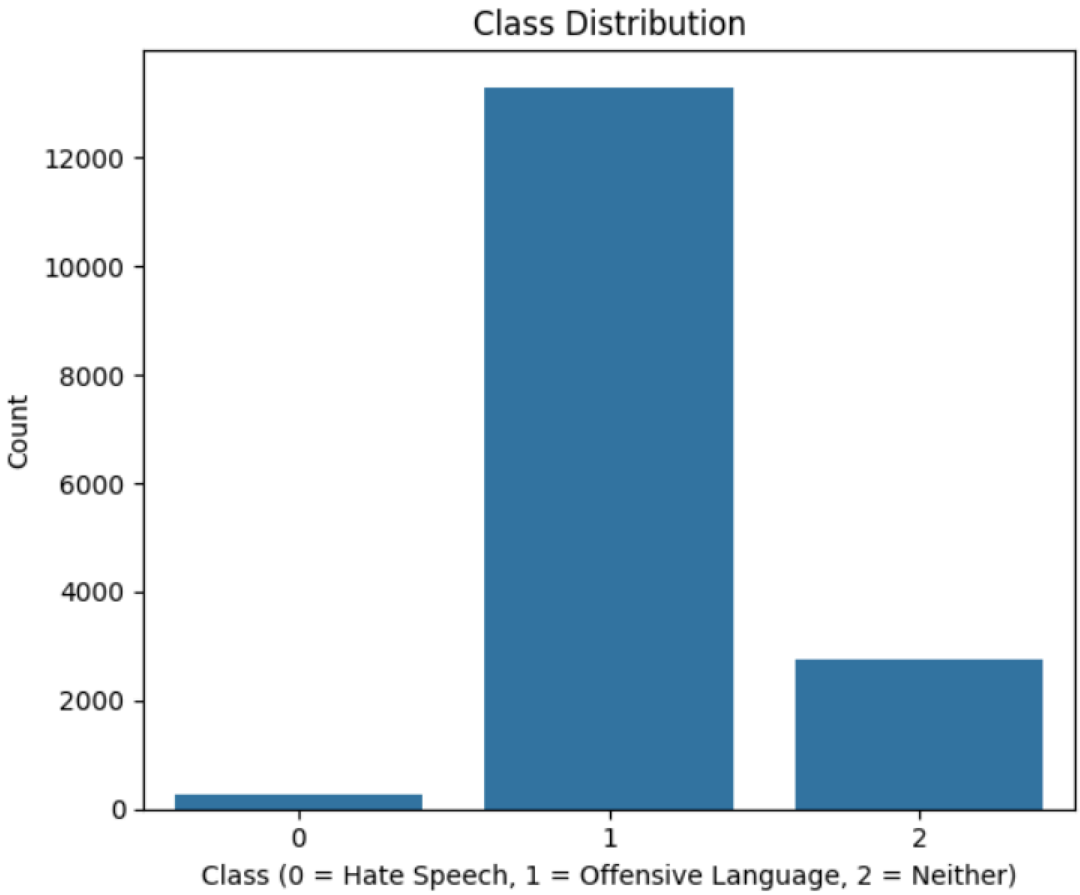


Figure 1. Class Distribution
This bar chart shows the distribution of classes in the dataset. Class 1 (Offensive Language) dominates with over 13,000 samples, while Class 0 (Hate Speech) has very few examples (~300), and Class 2 (Neither) has around 2,700. The dataset is highly imbalanced.

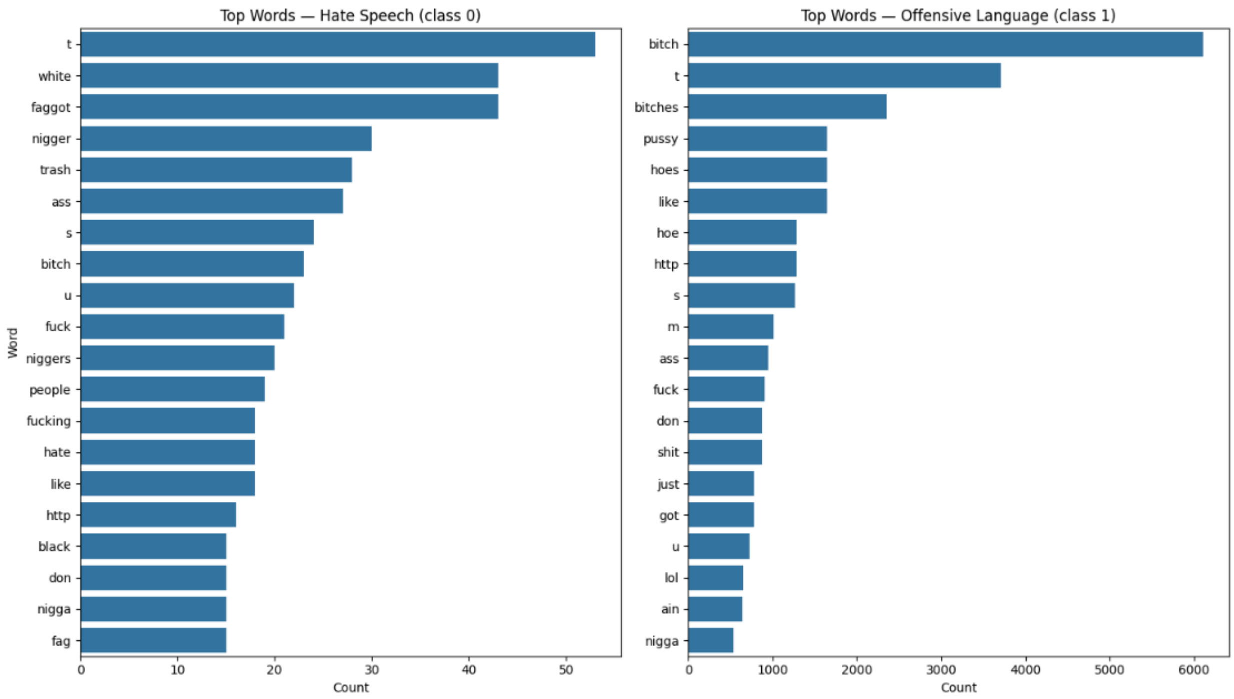


Figure 2. Top Words in Hate Speech and Offensive Language
The left panel shows the most frequent words in Hate Speech tweets (Class 0), where slurs and group-targeted terms like "f***ot," "n***er," and "white" appear prominently. The right panel shows the top words in Offensive Language tweets (Class 1), dominated by profanity and misogynistic terms such as "b*tch," "b*tches," and "pu**y."

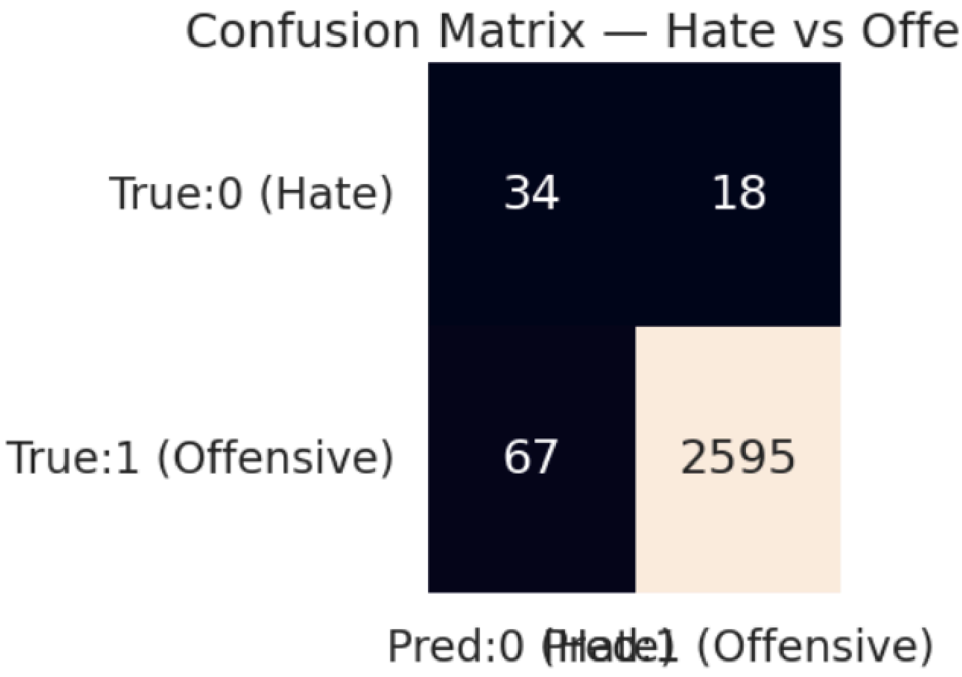


Figure 3. Confusion Matrix — Hate vs Offensive

This confusion matrix compares model predictions for Hate Speech (0) versus offensive language (1). While most Offensive tweets are correctly classified (2595), the model struggles with Hate Speech, misclassifying 67 of 101 actual Hate tweets as Offensive.

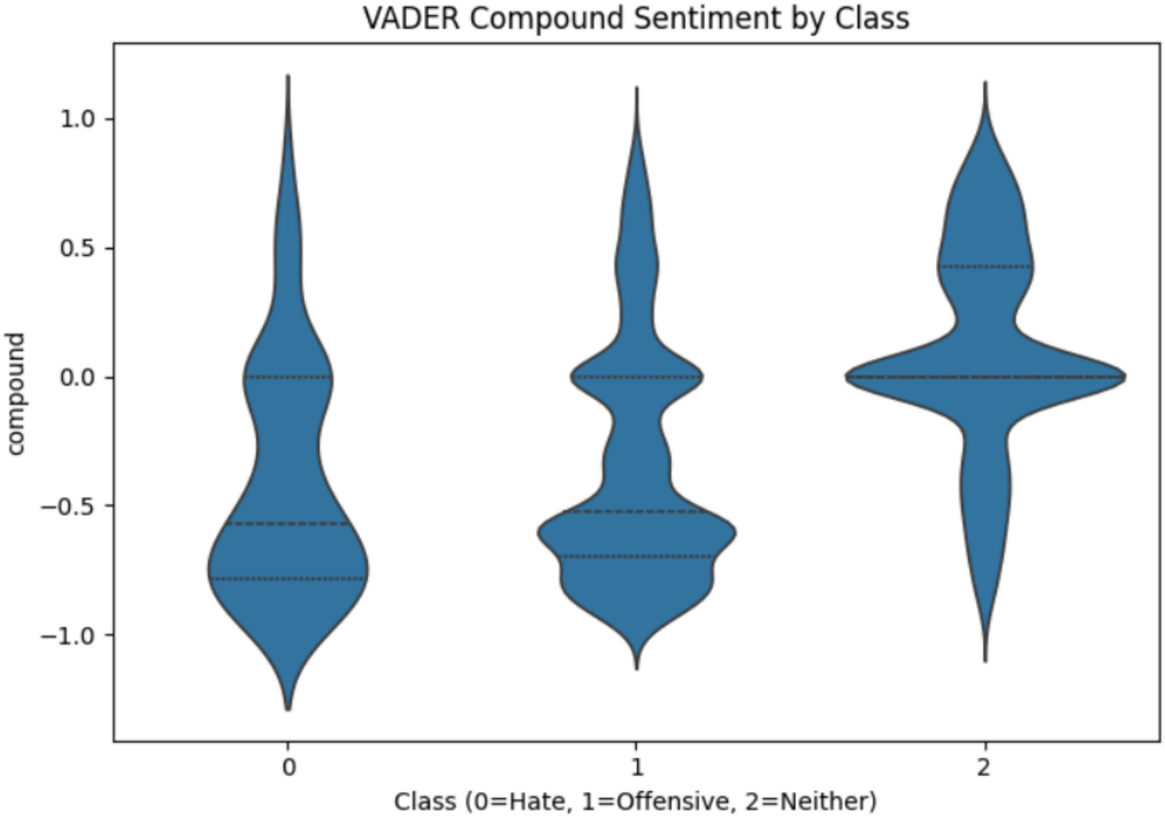
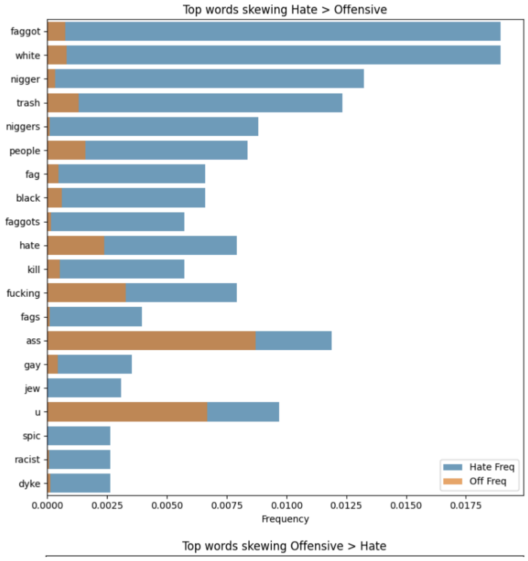


Figure 4. VADER Compound Sentiment Distributions

Violin plots show the distribution of sentiment polarity per class. Hate and Offensive tweets concentrate in negative ranges, while Neutral tweets center around zero, confirming the classifier’s distinction between hostile and non-hostile language.



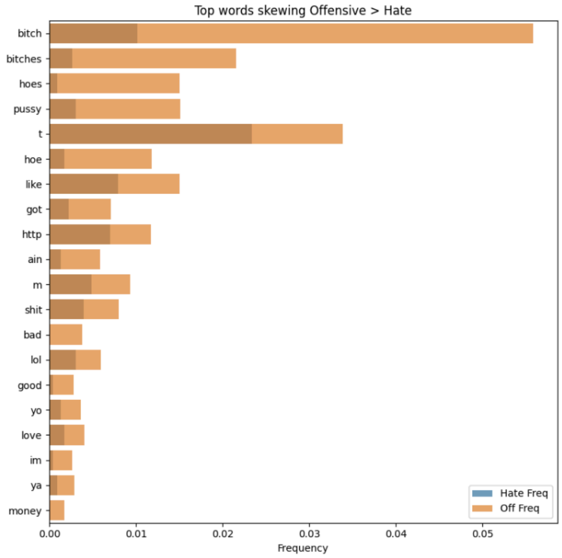


Figure 5. Top Words Skewing Hate vs Offensive (Hate > Offensive) = left graph
This chart shows words disproportionately associated with Hate Speech (blue bars) compared to Offensive language (orange). Words such as “f***ot,” “n***er,” “white,” and “jew” skew heavily toward Hate, reflecting targeted group-based slurs.

Figure 6. Top Words Skewing Offensive vs Hate (Offensive > Hate) = right graph
This chart shows words disproportionately linked to Offensive Language. Terms like “b*tch,” “b*tches,” “h*es,” and “pu**y” occur more frequently in Offensive tweets than Hate, reflecting profanity and misogynistic language rather than group-targeted slurs.