CS3 Rubric — Hate Speech Identification Case Study

DS 4002 Fall 2025 - Vivian Jiang
Due Dec 09
Submission Format: Upload link to Gitub Repository on UVA Canvas
Individual Assignment

Why am I doing this? This case study guides you through building a complete data science workflow to analyze harmful speech in social media text. You will explore a labeled dataset of tweets, identify meaningful linguistic patterns, and build a modeling pipeline to distinguish between speech categories. By completing this project, you will practice:

- Preparing real, messy text data
- interpreting linguistic signals
- building and comparing classification models
- explaining results clearly and responsibly

This assignment emphasizes methodological clarity, reproducibility, and critical reasoning about model performance and uncertainty.

What am I going to do? You will work with a dataset of tweets labeled into different speech categories. The GitHub repository (https://github.com/vivianjjiang/CS3-DS4002-HateSpeech) provides scripts, references, and instructions to help you begin. After accessing the dataset, you will explore the text to understand label frequencies and identify key linguistic patterns. You will then prepare the dataset using standard text preprocessing steps and build two classification models: a transparent baseline model and a context-aware transformer model. After training and evaluating both models, you will compare their accuracy in distinguishing among the categories. Finally, you will interpret your results by examining errors, highlighting influential linguistic features, and assessing each model's strengths and limitations. You will present your findings and code in a clear, well-organized GitHub repository.

Your final deliverables will include:

- a data dictionary
- well-documented and commented source code
- written summary reflection questions (3-4 sentences each)
- a complete GitHub repository with all materials

Tips for success:

- Begin with a careful review of the dataset; a good EDA will guide your modeling.
- Keep your code organized and well-commented so the workflow is easy to follow.
- Build and test the baseline model first to ensure your pipeline works, then add the transformer model.
- Compare models thoughtfully and explain *why* their performance differs.

How will I know I have succeeded? You will meet expectations when you successfully follow and complete the criteria in the rubric below.

Spec Category	Spec Details
Formatting	One GitHub Repository (submitted via link on Canvas) - Create a new GitHub repository for this assignment titled ‘CS3_HateSpeech’ that contains: - README.md - Source Code File - REFERENCE.md
README.md	- Brief Summary of what you have produced for the case study (doesn’t need to be detailed; 3-4 sentences) - 1-2 sentences about the data source - Data dictionary
Source Code File	Well-documented Google Colab Notebook file that contains the code used to execute: - Data loading: load the provided tweet dataset. - EDA: include two EDA visualizations and clearly label all figures, tables, or outputs. - Text Preprocessing Steps - Baseline Model - Implement an interpretable model (ex: logistic regression) - Evaluation metrics: accuracy, macro-F1, and confusion matrix - Transformer Model - Run a small transformer-based model (eg. DistilBERT or BERT) - Evaluation using the same metrics as the Baseline Model. - Model Comparison: Side-by-side table of performance metrics - Final Summary Cell: The notebook should end with a short Markdown summary answering: - What did you learn about how models detect harmful speech? - Which model handled nuance better, and why?
REFERENCES.md	Markdown File titled ‘REFERENCES.md’ with citations of any resources referenced in helping you create your model in IEEE documentation style.

Acknowledgements: Thank you Professor Alonzi for providing the rubric structure