Final Project

Natural Language Processing

Deadline: 04/25/2025

Total Points: 150

Dataset: Twitter US Airline Sentiment Dataset

Part 1: EDA & Feature Engineering [50 Points]

a. Exploratory Data Analysis [15 Points]

- Load the dataset, check dimensions, and handle null values.
- Explore post/tweet length distributions and label proportions.

b. Text Visualization [15 Points]

- Word clouds per class
- Top *n* frequent word occurrence
- Tweet/post length histograms

c. Preprocessing & Feature Engineering [30 Points]

- Clean, tokenize: lowercasing, punctuation removal, stopwords, stemming/lemmatization
- Create at least two representations:
    - Bag of Words
    - TF-IDF

Part 2: Model Building [50 Points]

Choose and train four models preferable two simple/explainable models (e.g., Logistic Regression, Naive Bayes, Decision Tree) and two black-box models (e.g., LSTM, RNN, BERT)

a. Model Training [25 Points]

- Train and evaluate all 4 models using the same feature set
- Use appropriate train/test split or cross-validation
- Report accuracy, precision, recall, F1-score, and confusion matrix

b. Architecture & Implementation Details [10 Points]

- Describe the pipeline for each model: vectorization -> model -> evaluation; provide details of the type of text vectorization used (e.g., TF-IDF, embeddings), model architecture (e.g., Logistic Regression with L2 regularization, LSTM with 128 units + dropout)
- For neural models: describe the architecture (layers, activation functions, dropout, etc.) and training details (optimizer, batch size, epochs, etc.)

c. Initial Observations [15 Points]

- Which models performed better and are the performance differences significant?
- How did simple/explainable models perform in comparison with black-box models? Can you identify any area where simple models fail or performed poorly?

Part 3: Interpretation, Error Analysis & Trade-Offs [50 Points]

a. Error Analysis [20 Points]

- Review 10–15 misclassified examples (preferably across classes)
- Discuss ambiguities in the text and patterns in errors (short tweets, sarcasm, class confusion)

b. Class Sensitivity & Confusion Patterns [15 Points]

- Analyze per-class precision, recall, and F1-score
- Visualize confusion matrices for all models
- Discuss how class imbalance or text structure might impact model performance

c. Explainability vs Performance: Comparison [15 Points]

- Compare explainable vs black-box models, is the performance gain from black-box models *worth the loss in interpretability*?

Submission Instructions

You can use NumPy, pandas, Matplotlib, scikit-learn, pytorch, tensorflow etc required libraries for your assignment, make sure to understand it conceptually and answer the questions with proper analysis

- Please upload a zipped file named 'FinalProject_YourName' to Dropbox. The zip file should contain the following items: your working directory containing a dataset, code file(s) (preferably .ipynb format) (please use relative paths when reading/importing the dataset), a PDF-format report, and a README.txt.
- Please ensure that your code is error-free; there should be no errors or warnings when running it on my machine.
- If you are using a programming language other than Python, please provide a README.txt file explaining how to run your code to obtain the result. Mention any libraries or dependencies that need to be installed before running the code, if necessary.

Academic Integrity

Discussion of course contents with other students is an important part of the academic process and is encouraged. However, it is expected that course programming assignments, homework assignments, and other course assignments will be completed on an individual basis (unless specified otherwise). Students may discuss general concepts with one another, but may not, under any circumstances, work together on the actual implementation of any course assignment. If you work with other students on “general concepts” be certain to acknowledge the collaboration and its extent in the assignment. Unacknowledged collaboration will be considered dishonest. “Code sharing” (including code from previous quarters) is strictly disallowed. “Copying” or significant collaboration on any graded assignments will be considered a violation of the university guidelines for academic honesty. If the same work is turned in by two or more students, all parties involved will be held equally accountable for violation of academic integrity. You are responsible for ensuring that other students do not have access to your work: do not give another

student access to your account, do not leave printouts in the recycling bin, pick up your printouts promptly, do not leave your workstation unattended, etc. If you suspect that your work has been compromised notify me immediately. If you have any questions about collaboration or any other issues related to academic integrity, please contact me immediately for clarification. In addition to the policy stated in this syllabus, students are expected to comply with the Wright State University Code of Student Conduct (http://www.wright.edu/students/judicial/conduct.html) and the portions Pertaining to Academic Integrity http://www.wright.edu/students/judicial/integrity.html) at all times. Note: In cases where there is suspicion of academic dishonesty, the professor and teaching assistant reserve the right to address the matter by calling in the student for an in-person question and answer session.