Molly Kaipus, Katie Martin, Berenice Rascon

CSCI 425: Python Machine Learning

Final Project Self Grade Reflection

Problem Definition & Motivation

a. 5

As a team, we clearly defined a real-world and relevant problem: analyzing customer sentiment in hotel reviews. Early in the project, we worked together to frame the objective in a way that aligned with practical applications, such as helping hotels better understand customer satisfaction and improve services based on feedback.

b. 5

We emphasized the importance of online reviews in influencing a traveler's decision on which hotel to stay at. Looking at hotel ratings greatly impacts a customer's choice of hotel and in turn, it is very important for a business to acquire data/information from these reviews. This helped ground our technical work and justified why a sentiment analysis model would be useful and impactful. Our project reflects how effectively we communicated this motivation and how we aligned the problem with our machine learning goals.

Data Acquisition & Preprocessing

a. 5

Our team gathered and prepared the data to ensure it was ready for machine learning modeling. We performed thorough data cleaning, which included removing HTML tags, punctuation, and special characters, as well as applying techniques like stopword removal. These preprocessing steps helped normalize the review text and reduce noise, making our input data much more useful for analysis.

b. 5

For feature engineering, we implemented TF-IDF vectorization as a team because it allowed us to assign higher importance to unique, sentiment-rich words, unlike simpler bag-of-words models. We also worked together to reframe the review scores into binary sentiment labels to make the classification task more focused and interpretable.

c. 5

Additionally, we ensured the dataset was properly split into training and testing sets (using an 80/20 ratio), so we could evaluate our model's performance accurately and avoid data leakage.

Model Selection, Evaluation, & Justification

a. 5

Algorithm selection was based on class examples and research for what would work best with the text data. Furthermore, we investigated the best methods of feature engineering and data processing for textual data, specifically reviews.

Initially, our goal was to build a 5-class sentiment classifier for star ratings (1–5), but the model performed poorly (around 50–60% accuracy) due to the complexity and overlap in language between adjacent ratings. To improve performance within project time constraints, we reworked the task as binary classification, grouping 3-star reviews with the negative class, as aligned with our abstract goals. This simplification significantly improved model accuracy and interpretability, though we recognized the linguistic ambiguity of neutral (3-star) reviews as a limitation. In later modeling steps, we created a new model using a more balanced dataset by removing 3 star reviews to reduce the ambiguity of language and improve model performance (90-95% range).

b. 5

Beyond standard model training, we created an optimized logistic regression model using GridSearchCV to fine tune parameters, which would become our best performing model. From the substantial tuning and additional data processing and engineering we applied, such as the ambitious decision to remove 3-star reviews entirely, we achieved a significant boost in accuracy in all the models, showcasing our thoroughness.

c. 5

We used more than accuracy to evaluate the performance of the models. We also discussed precision, recall, and F-1 score values as numeric evaluators. Additionally, we created charts displaying the top features influencing sentiment for each model and confusion matrices to gain a deeper understanding of model behavior along with performance. Our use of both numeric evaluators and visual diagrams to analyze performance shows that we did an in-depth evaluation of the models.

Creativity & Innovation:

a. **4**
We preprocessed the data, applied appropriate feature engineering, and tested multiple machine learning models based on research and learnings from class. Our project pipeline was not a direct replication of tutorials or class examples, as the vectorizer choices, preprocessing steps, model selection, and comparative analysis reflect our own original problem-solving approach and exploration.

b. 4
Additionally, we went beyond standard model accuracy by identifying and visualizing the top features that influenced sentiment for each model. Instead of treating the model accuracy results as is, we explored further to produce visualizations and explanations that were more human interpretable.

Presentation:

a. 5
Our PowerPoint slides included talking points to aid the audience in engaging in our discussions and insights. We also included diagrams, charts, and graphs to help explain the data and the behavior of models. We derived profound results from our accuracy improvement efforts and were able to discuss thought-provoking insights on experimentation in finding optimal processing and engineering. Additionally, we provided valuable visualizations, like top feature charts and confusion matrices, to make the insights more human explainable and perceptible.

b. 5
Our presentation was structured around the project guidelines, the judge's rubric, and the START presentation guidelines. Each group member discussed their situation, tasks, actions, results, and takeaways based on their contributions to the project. Our presentation was well-practiced, organized, informative, and fluid in transition between each member to produce a well-rounded demonstration of our semester project. Additionally, we were able to answer all the audience's questions.