Team Members:

Carlos Escamilla, Tyler Heflin, Andres Melendez, Delilah Slabaugh

DSC450

19 July 2025

# Stakeholder Q&A

To anticipate potential inquiries during a presentation or review, we compiled ten key questions along with answers prepared in advance:

1. What preprocessing steps were most effective in cleaning the text?

Removing HTML tags, special characters, and stopwords significantly improved data quality. These steps reduced noise, normalized the input, and allowed the models to focus on meaningful terms rather than irrelevant formatting or filler words.

1. Why was **TF‑IDF** chosen over word embeddings?

TF‑IDF was selected because it offers a simple, interpretable feature representation that works well with traditional linear models. It performs strongly on smaller datasets and does not require the large computational resources or pretraining that embeddings often need.

1. How were missing or corrupt reviews handled?

Rows containing missing or null review text were dropped from the dataset. This maintained data integrity and ensured the models were trained on complete, reliable input without introducing bias or noise from incomplete entries.

1. Which model performed best and why?

DistilBERT achieved the highest performance. Its transformer-based architecture effectively captures contextual relationships and subtle linguistic cues, which are especially valuable for sentiment analysis tasks.

1. Were there signs of overfitting in any models?

No significant overfitting was observed. Cross-validation showed consistent performance across training and validation folds, indicating that the models generalized well.

1. How would this approach scale to larger datasets?

Logistic Regression and Random Forest models scale efficiently with larger datasets and benefit from additional data. DistilBERT can also handle larger datasets but would require more computational resources, such as greater GPU memory and longer training times.

1. Could sentiment be multi-class (neutral, etc.)?

Yes. The same approach can support multi-class sentiment classification by redefining the label set to include additional categories, such as neutral or mixed, and retraining the pipeline.

1. What are the limitations of this dataset?

The dataset consists solely of text reviews, which may not fully represent the diversity of audience opinions. Cultural references, sarcasm, and context-dependent expressions can be misinterpreted by models, potentially reducing accuracy.

1. How would you improve model accuracy?

Accuracy could be improved by fine-tuning transformer-based models, experimenting with larger or domain-specific embeddings, integrating additional metadata such as user ratings, or applying advanced preprocessing techniques like lemmatization and phrase detection.

1. What business applications does this have?

This sentiment analysis can enhance recommendation engines, monitor audience reactions in real time, guide product or content improvements, and inform marketing strategies by tracking sentiment trends over time.