Project 3: Audience Questions and Answers – Milestone 3

Shanthibooshan Subramanian

Bellevue University

DSC680

Amirfarrokh Iranitablob

Here are ten questions that an audience might ask regarding the analysis of Tweet Sense.

1. What was the primary goal of the Tweet Sense analysis, and how did it influence the choice of models and techniques?

   The primary goal of the Tweet Sense analysis was to accurately classify the sentiment of tweets as positive, negative, or neutral. This goal influenced our choice of models and techniques by pushing us towards robust machine learning algorithms capable of handling text data.

2. How did you handle data preprocessing and cleaning for the tweet dataset? Were there any specific challenges you encountered?

   For data preprocessing and cleaning, we removed special characters, URLs, and stop words, tokenized and normalized the text, converted everything to lowercase, and performed lemmatization to reduce words to their base form. One of the specific challenges we encountered was dealing with slang, abbreviations, and emoticons common in tweets, which required careful handling to maintain the context and meaning of the text.

3. What were the key challenges in handling noisy or ambiguous tweet data, and how did you mitigate their impact on model performance?

   Handling noisy or ambiguous tweet data presented several challenges, primarily due to the informal and unstructured nature of tweets and the presence of sarcasm, irony, and ambiguous language. To mitigate their impact on model performance, we employed advanced preprocessing techniques like lemmatization and stop-word removal.

4. How does the inclusion of stop words impact the performance of your model, and why did you choose to remove them?

   Including stop words in the analysis can add noise to the model since these are common words that do not carry significant meaning. Removing stop words helps the model focus on words that contribute more to the sentiment, thereby improving the accuracy and efficiency of the model. This is why we chose to remove them during preprocessing.

5. Can you explain the rationale behind choosing the specific machine learning models for sentiment classification?

   We selected specific machine learning models for sentiment classification based on their strengths. Logistic Regression is simple and interpretable and works well with TF-IDF features for text classification. Bernoulli Naive Bayes and Multinomial Naive Bayes are well-suited for text data and perform effectively with the sparse features from TF-IDF, with Bernoulli Naive Bayes being particularly good for binary/boolean features and Multinomial Naive Bayes for multinomial distributed data. LinearSVC is effective for high-dimensional spaces and offers strong performance with text data.

6. Why did you choose to use TF-IDF with n-grams up to 2 (unigrams and bigrams) for feature extraction?

   We chose to use TF-IDF with n-grams up to 2 for feature extraction because TF-IDF highlights important words by considering their frequency and rarity across documents. This improves the representation of tweet data. Using n-grams up to 2 helps capture more context and relationships between words, which is crucial for understanding sentiment in short and informal text like tweets.

7. What metrics did you use to evaluate the performance of your models, and how did they compare across different models?

   To evaluate the performance of our models, we used accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation in terms of both correctness and robustness. Models like Logistic Regression and LinearSVC generally performed better in terms of precision and recall, while Bernoulli Naive Bayes and Multinomial Naive Bayes showed competitive accuracy.

8. How do you plan to extend or improve this analysis in future work, and what additional features or models might you explore?

   In future work, we plan to extend and improve this analysis by incorporating hyperparameter tuning for better model optimization, exploring deep learning models like LSTM or BERT for enhanced text understanding, adding more features such as tweet metadata (e.g., user information, hashtags), and experimenting with ensemble methods to combine the strengths of multiple models.

9. What measures have you taken to ensure the privacy and confidentiality of users' tweets, particularly considering that sentiment analysis involves analyzing potentially sensitive content?

   To ensure the privacy and confidentiality of users' tweets, we anonymized the dataset by removing any personal identifiers and only used publicly available tweets. The analysis focused on aggregate trends rather than individual tweets, and we adhered to ethical guidelines and best practices in data handling and analysis.

10. What real-world applications or implications can your sentiment analysis findings have, and how might they be used in practical scenarios?

    Sentiment analysis of tweets can have several real-world applications. For instance, it can be used in market research to understand customer opinions about products and brands, in public opinion analysis to gauge sentiment on political events or social issues, in customer service to identify and address customer complaints and feedback in real-time, and in social media monitoring to track sentiment trends and inform marketing strategies and campaigns.