**Sentiment Analysis on Technology and Transportation in San Antonio**

Robert Santos

**Introduction**

In this report, we present our approach to analyzing community sentiment on technological innovation and transportation in San Antonio using Reddit data from the subreddit r/Sanantonio. The goal is to develop machine learning models that can classify sentiments expressed in the community toward technological advancements and transportation developments as positive, negative, or neutral. Our analysis aims to provide insights to the city of San Antonio for decision-making and policy planning regarding technology adoption and transportation initiatives.

**Methodology**

**Data Collection and processing:**

Both technology and transportation-related data were collected from the r/sanantonio subreddit. This data was used to do a prediction on whether a statement was technology related, transportation related, or neither. These statements included posts discussing various topics from animal rights to political commentary. The annotations wrote down were gathered by two other groups, which annotated the statements based on our guidelines.

**Feature Engineering:**

We experimented with several features to capture the sentiment in the text, including predicting the presence of keywords for technology-related terms and transportation-specific terms. Additionally, we considered punctuation marks such as exclamation points and the count of capital letters in the statement too. Furthermore, lexicon-based features were incorporated to enhance sentiment analysis by incorporating sentiment scores obtained from lexicons. These were used in the modeling process.

We explored the logistic regression model due to its ease of interpretability, and the LSVM model due to the class separation. Ideally, more knowledge on machine learning itself could lead us to use better models suited for this type of data set. The data was then split 80/20, which we used for testing our models. Each model was trained on the training data, then tested either on the test set or using cross validation. During this testing, grid search was used to find the optimal tuning parameter for these models. **Evaluation Metrics:**

We evaluated the performance of each model on the validation set using the following metrics: Precision, Recall, and F1-score. These were our results. For the Micro results, the 3 results using the Micro metric were the same.

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic on Test Data w/o Lexicon | Linear SVC on Test Data w/o Lexicon | Logistic with CV w/o Leixcon | Linear SVC with CV w/o Lexicon |
| Macro F1: 46.88 | Macro F1: 52.59 | Macro F1: 43.48 | Macro F1: 43.41 |
| Macro Recall: 48.17 | Macro Recall: 52.03 | Macro Recall: 45.10 | Macro Recall: 44.66 |
| Macro Precision: 45.70 | Macro Precision: 55.67 | Macro Precision: 45.88 | Macro Precision: 42.72 |
| Micro Scores: 68.66 | Micro Scores: 70.15 | Micro Scores: 65.67 | Micro Scores: 64.18 |
| Logistic on Test Data With Lexicon | LSVC on Test Data With Lexicon | Logistic with CV w/ Lexicon | LSVC with CV with Lexicon |
| Macro F1: 52.40 | Macro F1: 56.11 | Macro F1: 45.86 | Macro F1:45.48 |
| Macro Recall: 51.64 | Macro Recall: 49.85 | Macro Recall: 46.90 | Macro Recall: 46.52 |
| Macro Precision: 57.81 | Macro Precision: 50.64 | Macro Precision: 45.10 | Macro Precision: 44.78 |
| Micro Scores: 69.51 | Micro Scores: 67.16 | Micro Scores: 67.12 | Micro Scores: 66.67 |

**Experimentation:**

We performed logistic regression and LinearSVM on the test split with and without the lexicon function. We then repeated these experiments with cross-validation sets, both with and without lexicon features. We then repeated these experiments with cross-validation sets, both with and without lexicon features. This approach helps us get results and to see if they are reliable. Ideally, there would be ways to verify the reliability of these models, such as measuring the standard deviation and confidence intervals for each fold. But looking at these results, a logistic model with lexicon features seems to perform the best. It is not significantly better than the LSVC model, but it is an improvement.

**Error Analysis:**

Additionally, errors were considered when dealing with this analysis. Common errors include misclassification of transportation or technology related statements, posts containing mixed sentiments, and sentiment indicators (e.g., emojis) not being effectively captured by the models.

**Results and Conclusion**

Based on our results, the logistic regression model with lexicon features would be the best model. The model testing with the stated features, which were used to predict whether a statement was either technology related, transportation related, or neither gave satisfactory results. However, I would suggest future improvements to be made when dealing with data like this. If more knowledge of different models were had, we could have gotten an even better model. Verifying the reliability of these models as well should be done. Overall, this analysis has shown how important it can be for companies to do sentiment analysis.