Technical Report

Author: Stephen Garcia, Carlos Araujo, Erica Rodriguez

Date: May 07, 2025

Course/Organization: [e.g., Data Analytics Capstone – UTSA]

1. Executive Summary

This project analyzes data annotation techniques and their impact on machine learning model performance. It involves creating annotation guidelines, extracting a dataset, and comparing annotations from two groups. Three machine learning models are employed on the dataset to categorize posts based on technology and politics.

2. Introduction

This project has two distinct parts: the first part focuses on data annotation, while the second part delves into model training and evaluation. The goal is to understand how different annotation methods affect the performance of machine learning models in various tasks.

3. Data Annotation

We created annotation guidelines and extracted a dataset of one thousand reddit posts. The dataset was given to two groups for annotation: Group A and Group B. Each group was tasked with annotating the posts based on specific criteria. We chose Technology Related vs Non-Technology related posts and Political vs Non-Political posts. The annotations were then compared to assess the consistency and reliability of the two groups' annotations. We also calculated the inter-annotator agreement to evaluate the level of agreement between the two groups. The inter-annotator agreement was calculated using Cohen's Kappa statistic.

4. Methodology

A gold standard dataset was created from the annotations from both groups; the final dataset consists of 1000 reddit posts with annotations labeled based on their category:

- Category 1: Technology Related vs Non-Technology Related
- Category 2: Politics Related vs Non-Politics Related

Three distinct machine learning models were employed on the dataset. A lexicon-based model was utilized to categorize the posts based on their content; where each post is

labeled as either technology-related or politics-related. A linearSVC (bag-of-words) model to learn patterns and classify post in the dataset. A BoW with Feature Engineering to incorporate additional selected features alongside word frequencies to improve classification performance (i.e., number of exclamation points, count of capitalization words).

5. Code Implementation

The first code block defined a class called LexiconClassifier. This class is designed to determine whether a given text has representations of technology or politics. The class uses a lexicon-based approach, where it checks for the presence of specific keywords related to technology and politics in the input text.

The LexiconClassifier class is trained on a dataset of reddit posts. The training process involves iterating through the dataset and checking for the presence of keywords in each post. The model then calculates the accuracy of its predictions by comparing them to the actual labels in the dataset.

The second code block splits the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance. The split is done using a 80-20 ratio, meaning that 80% of the data is used for training and 20% for testing. The LinearSVC model is trained on the same dataset as the LexiconClassifier, but it uses a different approach to classify the posts. The training process involves creating a vector representation of each post and using a linear classifier to predict its label. The model's performance is evaluated by comparing its predictions to the actual labels in the dataset.

The BOW model is a bag-of-words model that uses the frequency of words in the posts to classify them. The model is trained on the same dataset as the LexiconClassifier and LinearVC models, but it uses a different approach to classify the posts. The training process involves creating a bag-of-words representation of each post and using a classifier to predict its label. The model's performance is evaluated by comparing its predictions to the actual labels in the dataset. Furthermore, we are adding in a feature engineering model that will use exclamation marks as a feature to classify the posts. The model is trained on the same dataset as the other models, but it uses a different approach to classify the posts. The training process involves creating a feature representation of each post based on the presence of exclamation marks and using a classifier to predict its label. The model's performance is evaluated by comparing its predictions to the actual labels in the dataset.

The process was repeated with the political and non-political reddit posts; the goal is to understand how the models perform on a different dataset and whether the results are consistent with the previous analysis. Then the models will be trained and evaluated using the same approach as with the technology reddit posts. The performance of the model will be compared to assess their effectiveness in classifying political posts.

6. Discussion

The technological classification models performed better than the political classification models.

- The **lexicon model** delivered the highest precision but lower recall. It is dependent on the lexicon and may not generalize well to unseen data.
- The **BoW + Capitalized Words** model had the best overall F1 score.
- Adding cues (like ! or capitalized words) improved classification over BoW alone.

The political classification models performed much worse than the technology classification models. The lexicon model had very low precision and recall, indicating that it was not good at classifying political posts relative to that of the technology posts. My assumption is that there is a very broad lexicon of words that can be used to classify political posts, and the model is not able to account for that.

- The lexicon model performed poorly overall, missing nuanced political language and showing low recall.
- The BoW-only model had the highest precision, meaning it was cautious and rarely mislabeled non-political posts.
- Adding exclamation point or capital word count features had a noticeable benefit during cross-validation (higher validation F1), but test F1 remained similar to the base BoW model and actually decreased relative to the BOW-only model.
- Overall, the BoW alone is sufficient for identifying political language in many cases, and additional features offer marginal gains but compared to the technology classification, the models are not as effective.

7. Conclusion and Recommendations

- Summary of what was learned
- Implications for policy, safety, or decision-making
- Actionable recommendations (if applicable)

8. Appendix

- Charts or tables referenced in-text
- Full data dictionary (optional)
- Code snippets or macro references

9. References

- Cite any data sources, academic papers, or tools used.