Technical Report

Author: Stephen Garcia, Carlos Araujo, Erica Rodriguez

Date: May 07, 2025

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

# 1. Executive Summary

A high-level overview of the project objectives, methodology, key findings, and recommendations. Keep it under 250 words.  
  
Example:  
This report presents an analysis of crime incidents across Bexar County educational institutions, using data from the Uniform Crime Reporting (UCR) Repository. The goal was to evaluate offense patterns by agency, identify high-risk environments, and assess the distribution of violent vs. non-violent crime. Key insights were visualized through comparative bar charts, per-capita metrics, and crime-type segmentation.

# 2. Introduction

This project is a comprehensive analysis of data annotation techniques and their impact on model performance. 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. 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 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.

# 6. Discussion

- Interpretation of results in context  
- Limitations of the data or methods  
- Assumptions made  
- Unexpected insights or discrepancies

# 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.