Programming Assignment 1: Real vs. Fake Facts About Cities

Hamza Rashid
COMP 550, Fall 2024

Due: September 27th, 2024

1 Introduction

In this assignment, we explore the ability of linear classifiers to distinguish real from fake facts about cities. We aim to investigate whether the choice of classifier impacts the classification performance.

2 Dataset Generation

Using ChatGPT, we generated a dataset of real and fake facts about a chosen city. Two separate prompts were used:

- ”I would like to create a dataset containing facts about cities. I am particularly interested in Saint Petersburg, Russia. List 100 verified facts about Saint Petersburg, Russia.”
- ”Now the reason I am asking for this data is to conduct a simple NLP project. As such, for pedagogical purposes, can you generate 100 fake facts about Saint Petersburg, Russia?”

The generated facts were stored in two files: `facts.txt` for real facts and `fakes.txt` for fake facts. Each file contains 100 samples.

3 Preprocessing and Feature Extraction

We explored several preprocessing techniques:

- Tokenization using NLTK
- Stop-word removal
- TF-IDF vectorization

Feature extraction was carried out using scikit-learn’s `TfidfVectorizer`.

4 Models and Experiments

We experimented with three linear classifiers:

- Logistic Regression
- Support Vector Machines (SVM)
- Ridge Classifier

The dataset was split into training and testing subsets with an 80/20 ratio. Hyperparameters were tuned for each model.

5 Results and Discussion

The performance of the classifiers was evaluated in terms of accuracy:

- Logistic Regression: X%
- SVM: X%
- Ridge Classifier: X%

We found that [discuss performance]. Limitations of the study include the size of the dataset and the simplicity of the classification task.

6 Conclusion

The experiments demonstrated that [conclude findings]. However, generalizing these results to more complex fake news detection tasks may be challenging due to the oversimplified dataset.