Locating Tweets

Using NLP and Classification Models

Dataset

Dataset

- **Twitter**
- *Twint* searched tweets for a user defined subject
- *Twint* located tweets from two user defined cities

Dataset

- Subject, first city, and second city is dynamic.
- User input can change to any subject to search Twitter.
- User input can choose which two cities to search in.

Dataset

Subject: Trump

City 1: Seattle

City 2: Jacksonville

- 10,000 recent tweets from each city
- Total of 20,000 tweets

Cleaning & Vectorizing

Cleaning

- Lowercased all words
- Removed URLs and special characters
- Lemmatized the remaining words

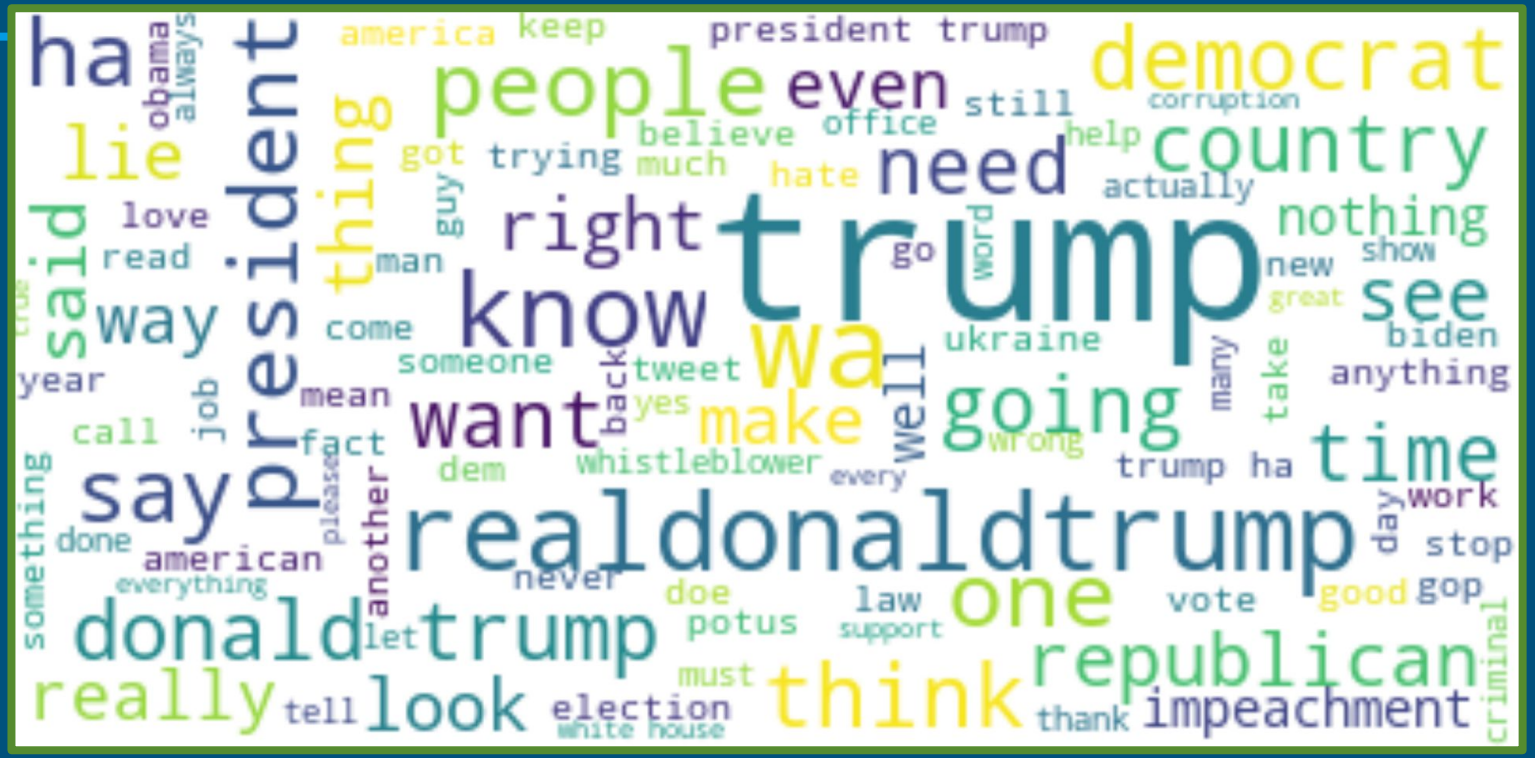
Vectorizing

- Used both CountVectorizer and TF_IDFVectorizer
- Similar performance with both
- Defaulted to **TF_IDFVectorizer**

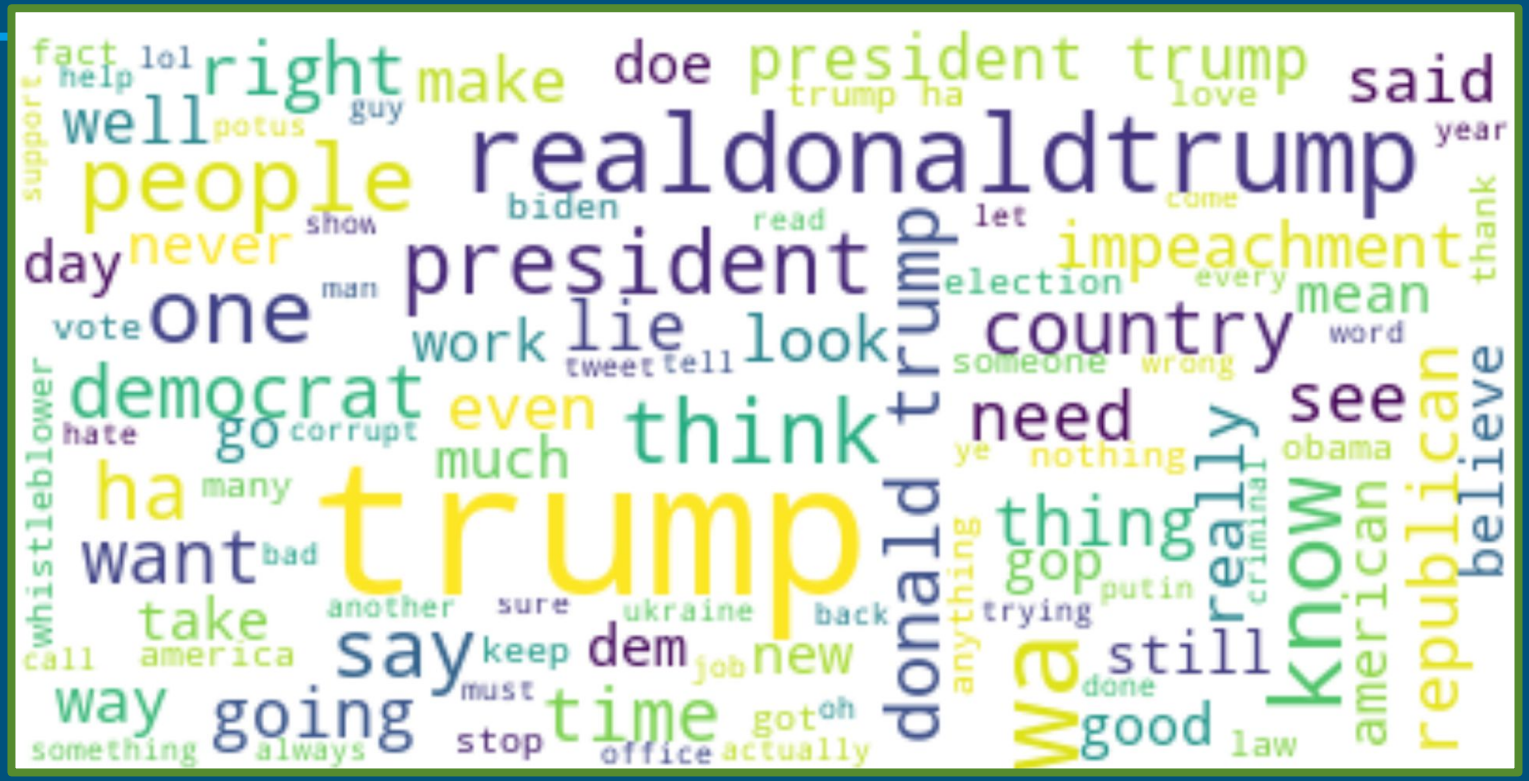
Exploring the Data

Word Cloud - Seattle

Word Cloud - Jacksonville

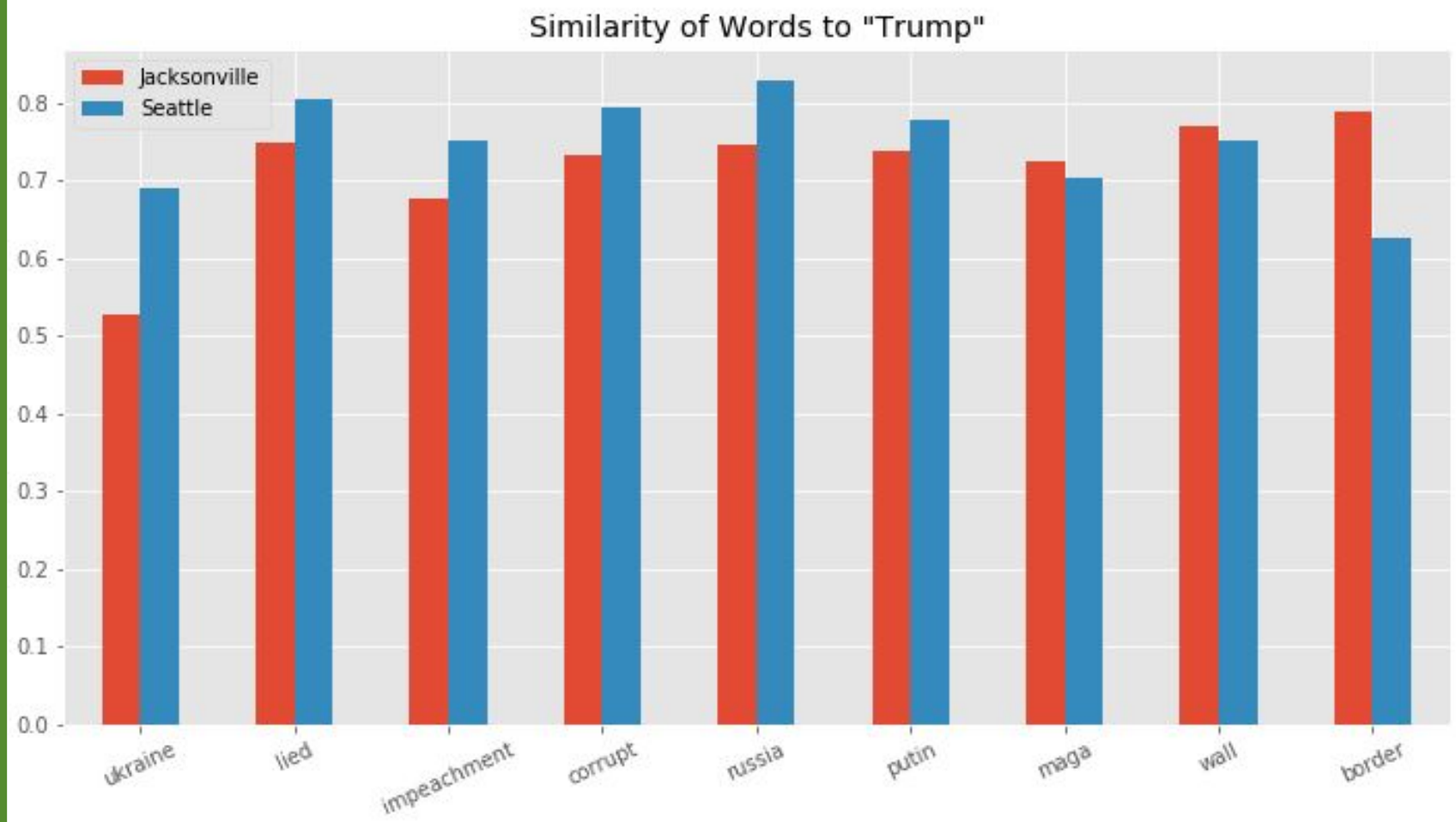


Word Cloud - Combined



Word2Vec

- Used word2Vec for exploring the data
- Found word associations
- Found frequently occuring words


Similarity of Words to "Trump"
Jacksonville
Seattle
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
ukraine
lied
impeachment
corrupt
russia
putin
maga
wall
border

Classification

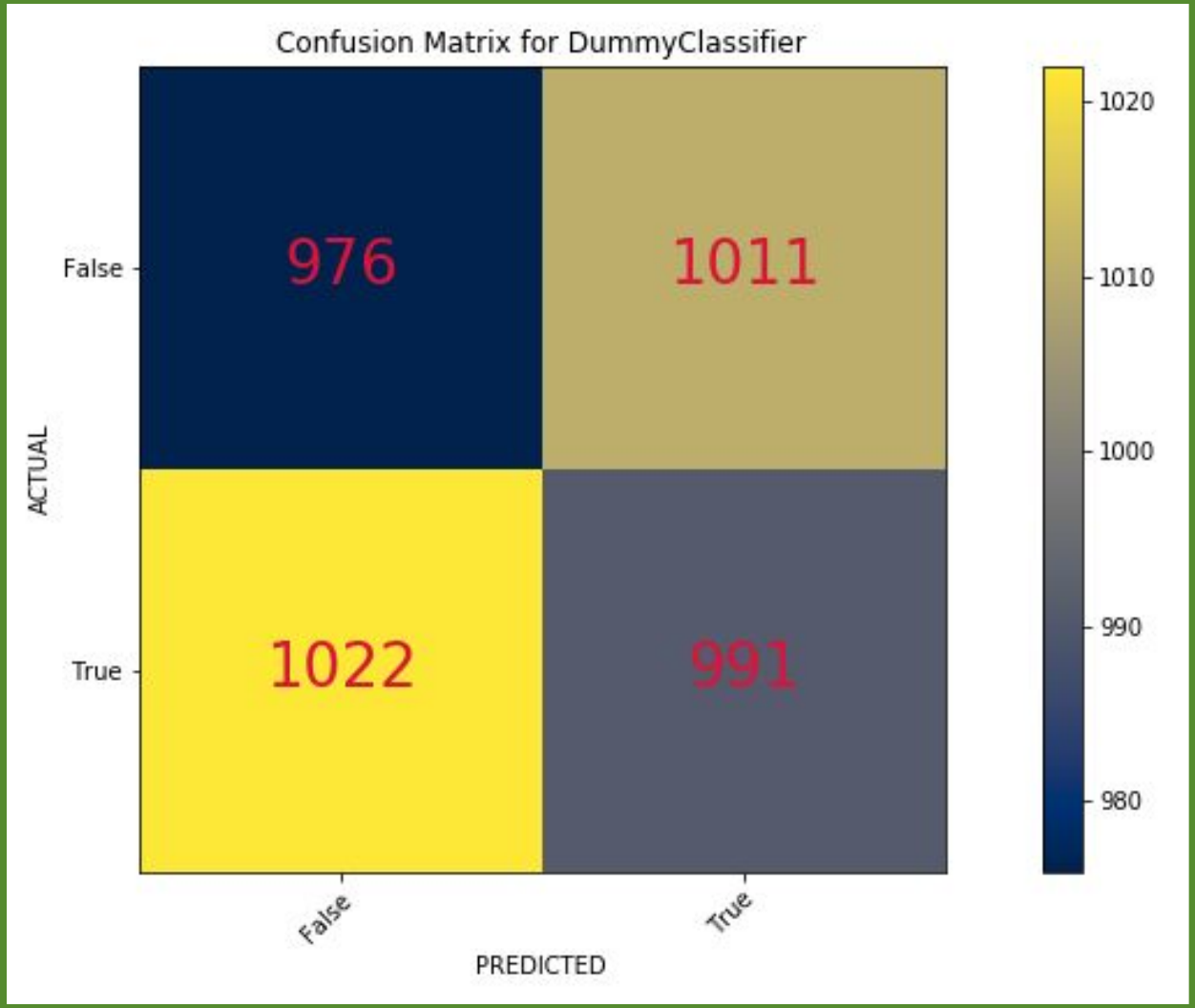
Classification Models

- Dummy Classifier - Baseline
- Random Forest
- Naive Bayes
- Logistic Regression
- Support Vector Machine

Results: Dummy Classifier

Training Score: 50%

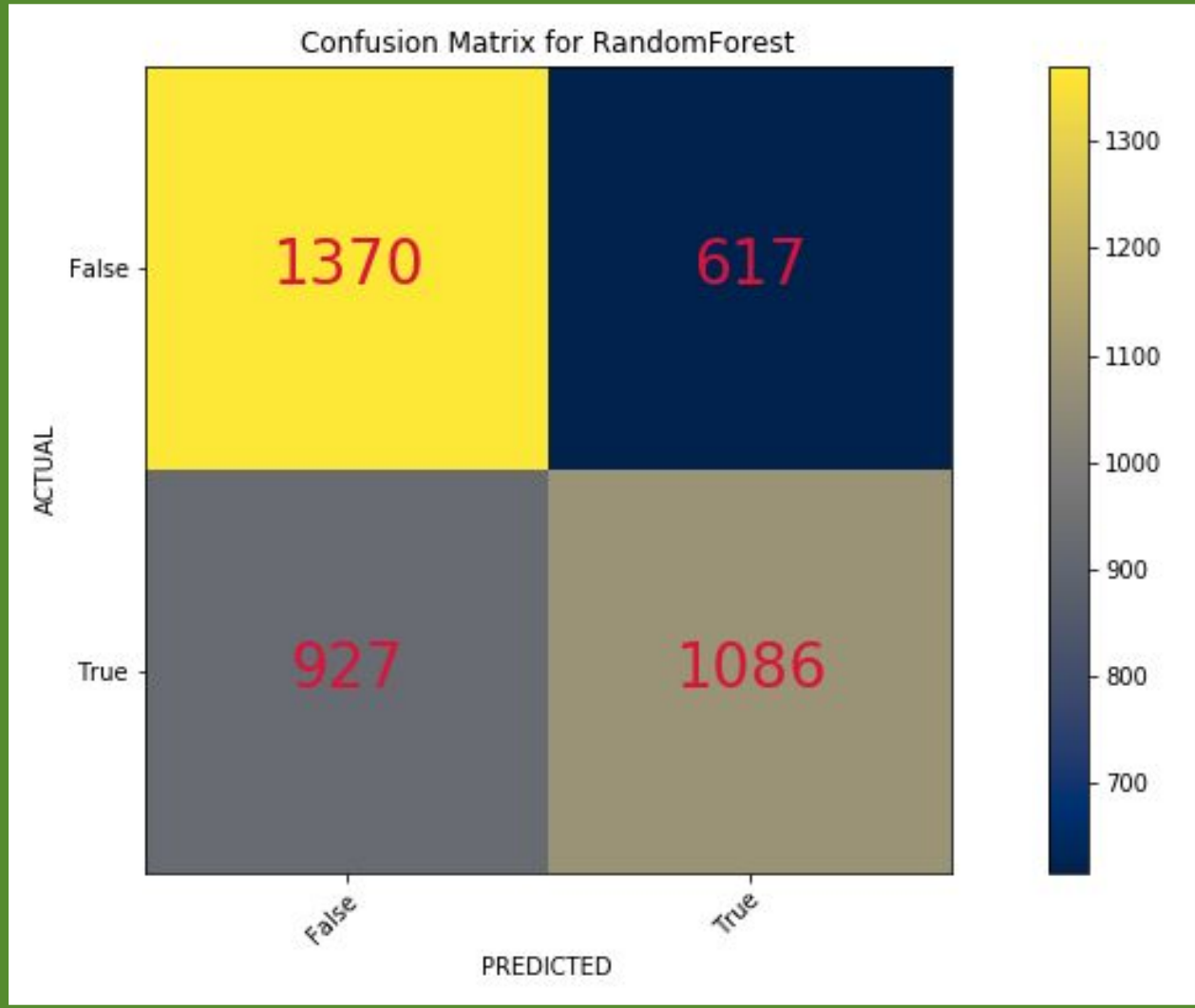
Accuracy Score: 49%



Results: Random Forest

Training Score: 96%

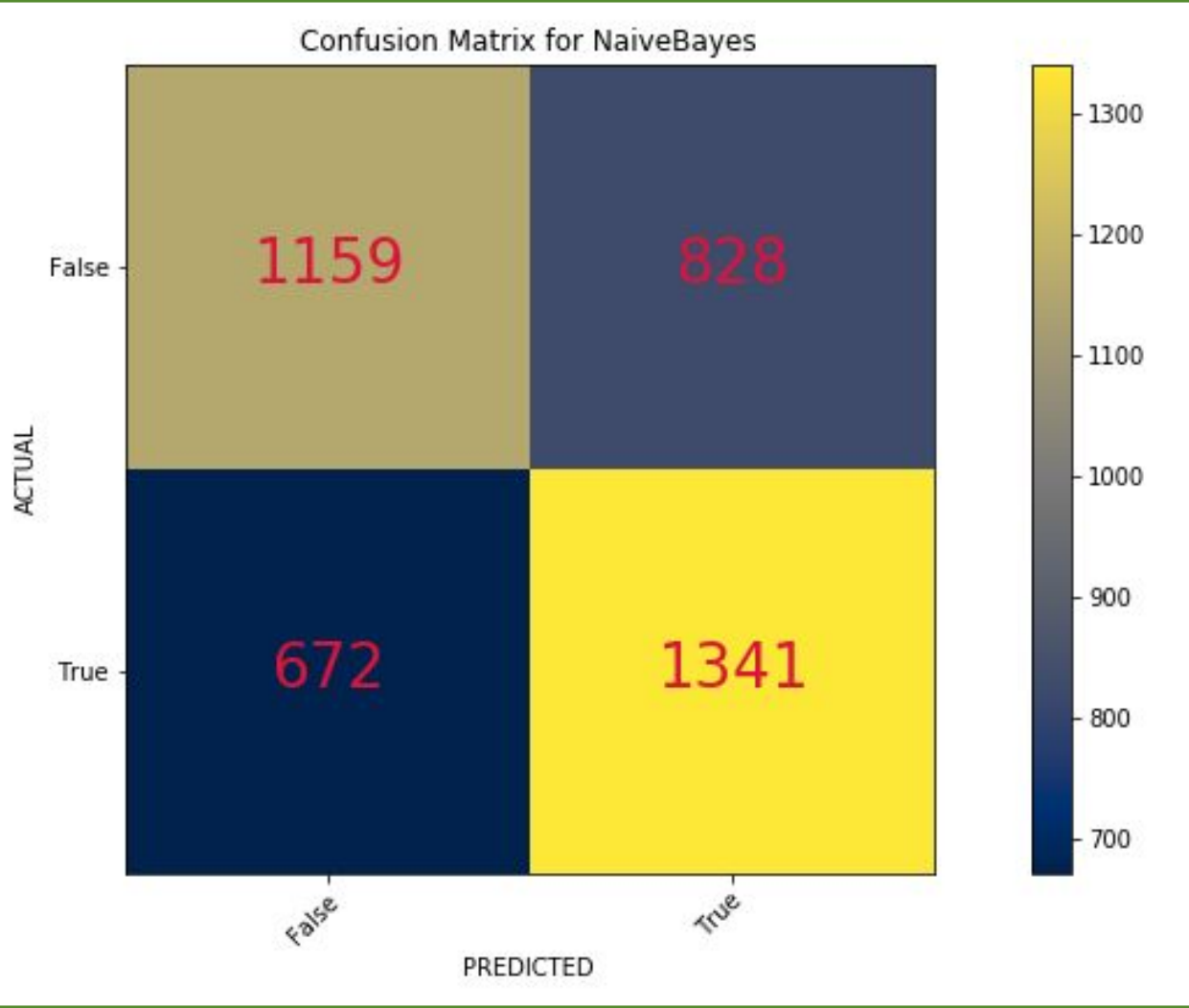
Testing Score: 60%



Results: Naive Bayes

Training Score: 79%

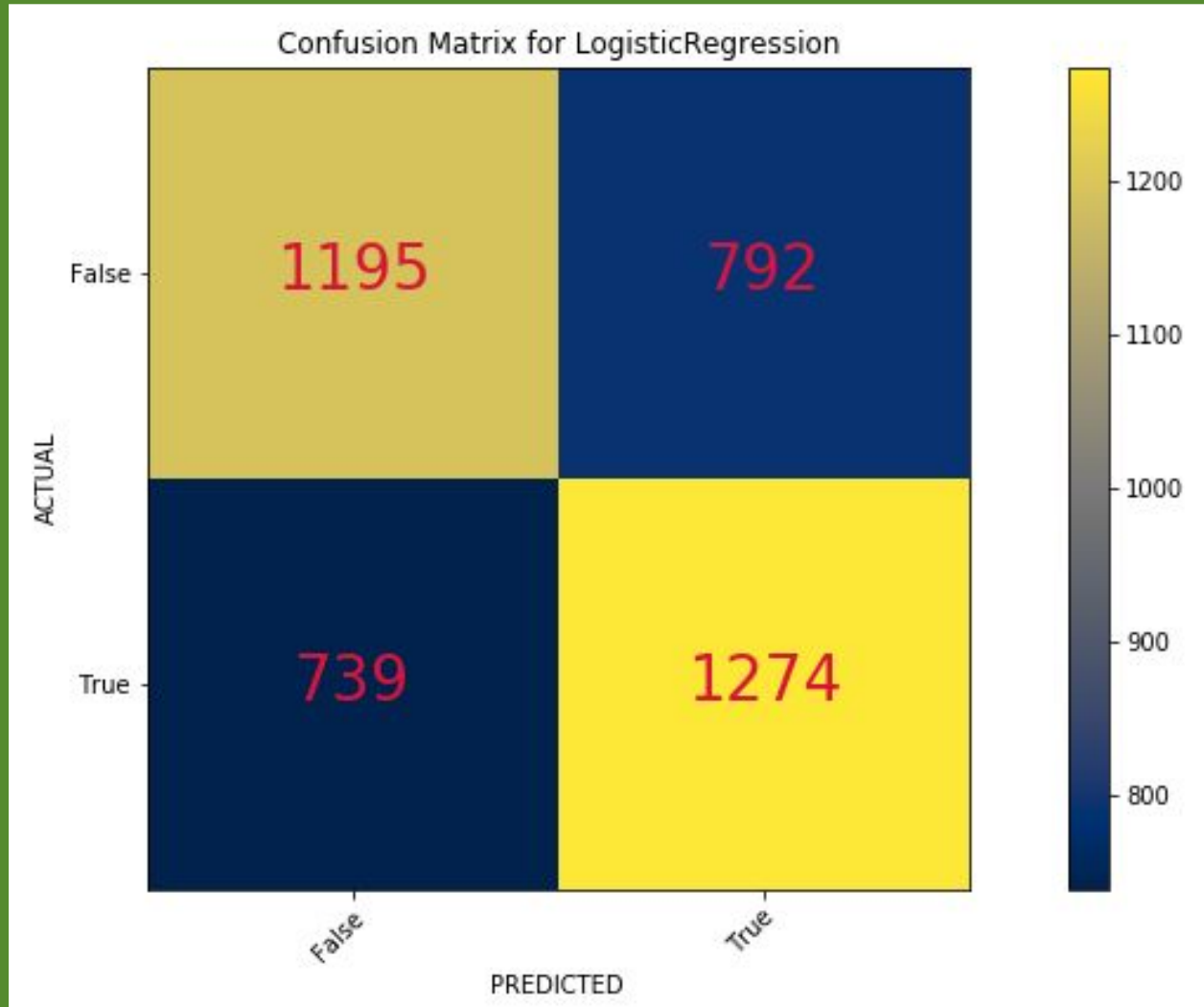
Testing Score: 62%



Results: Logistic Regression

Training Score: 82%

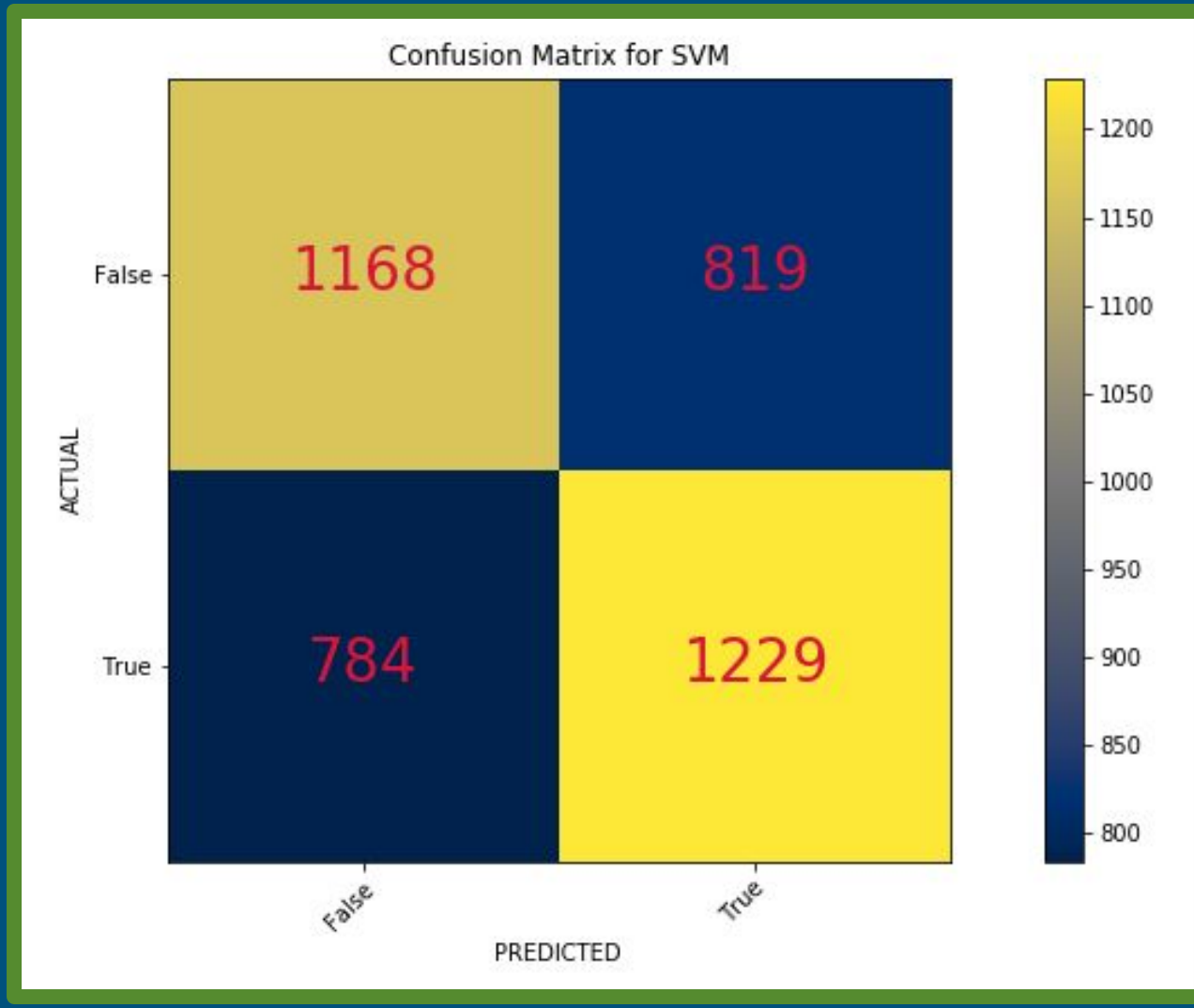
Testing Score: 61%



Results: Support Vector Machine

Training Score: 87%

Testing Score: 59%



Deep Learning Neural Network
(Attempt)

- Sequential Model
- Very simple, only 3 layers.
- Similar results as the classification models.

Improvements & Closing

Potential Improvements

- More models could be used such as XGBoost, KNN, etc.
- Feature engineering such as *ngrams*.
- Other cleaning techniques such as different lemmatization modules.
- More exploration of the neural network.

Closing

- Models performed at least 10% better than the Baseline
- Best performing model was ***Naive Bayes***
- Different subjects with different cities could alter results.