



NLP CLASSIFICATION

Using Reddit's API to predict post content for
r/911FOX and **r/NCIS**

DSI -Project 3
Suthatta Dontriros

Background



Title: 9-1-1

Genre: Action drama

Character: Police officers,
Firefighters
Dispatchers

Original release: 2018 –; present

API yield: 923 posts



Title: NCIS

Genre: Action drama

Character:
Naval Criminal Investigative Service

Original release: 2003 – present

API yield: 996 posts

Problem Statement

How to increase the accuracy of post separation and reduce the wrong post on subreddit?

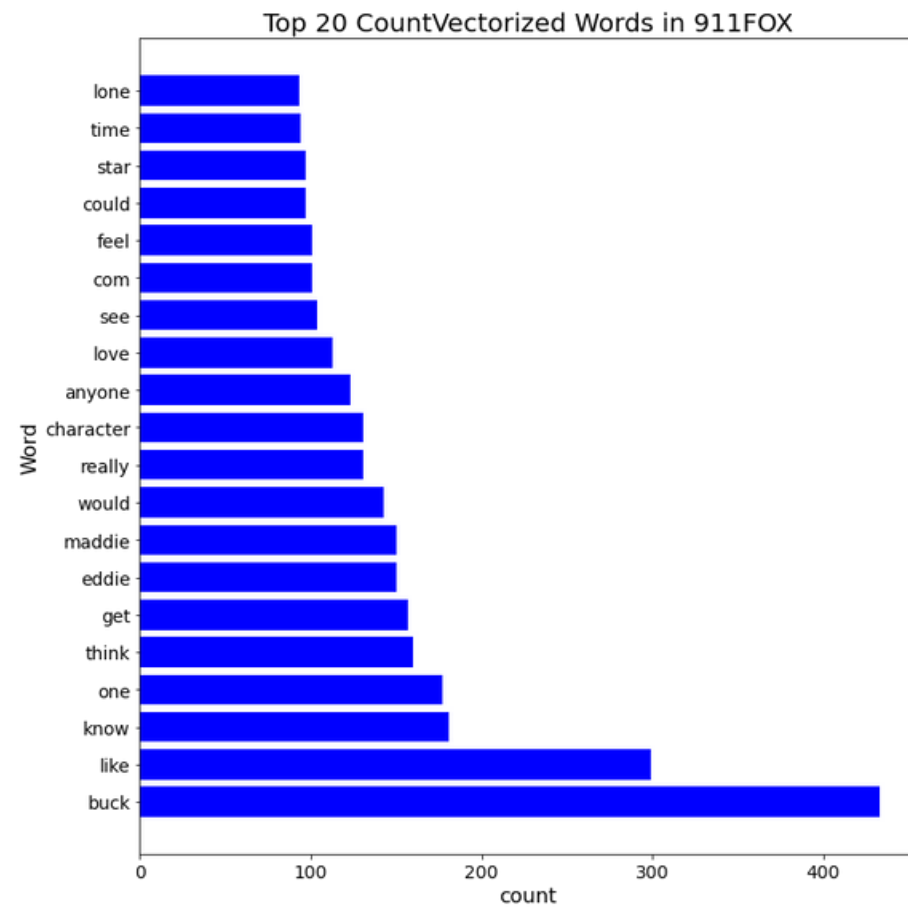
I work as an IT consultant who has a project with Reddit. The problem that Reddit found is the similar movies named 9-1-1 and NCIS have subreddit, but some users do not know the name of the subreddit that make them cannot post the new post or create the post in the wrong subreddit. Therefore, Reddit wants to try a model to collect the post from the user and select the subreddit for the user and increase more efficiency to manage the wrong post in subreddits.

EDA

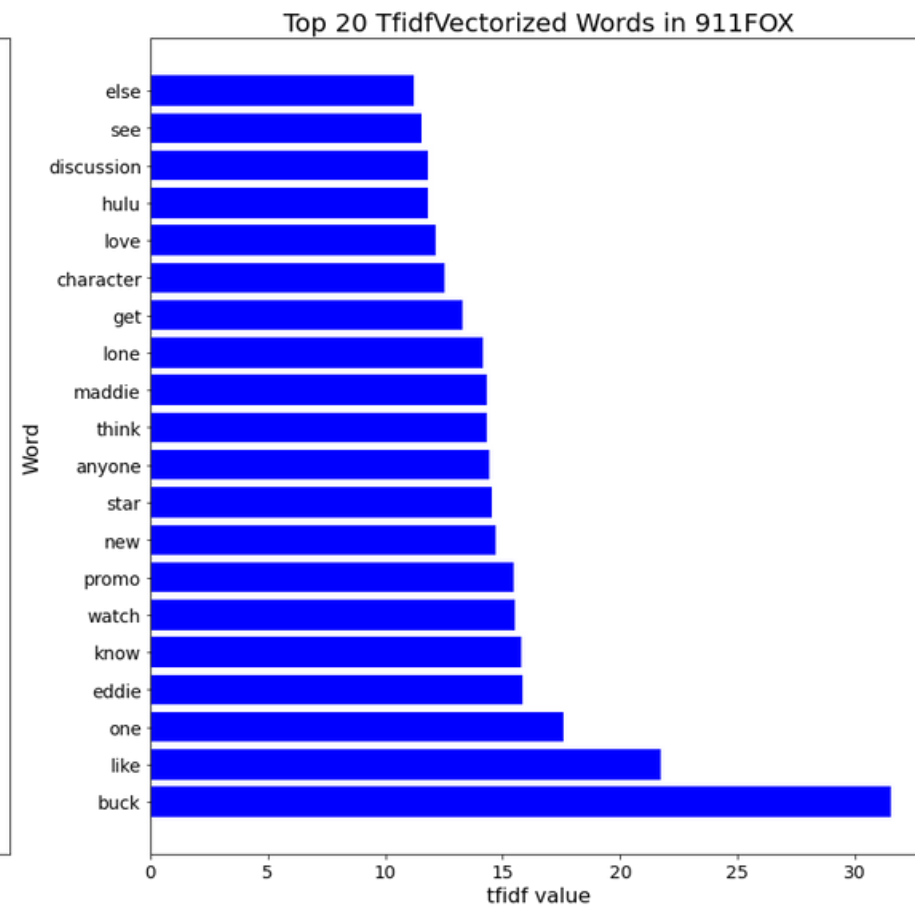


911FOX

CountVectorized

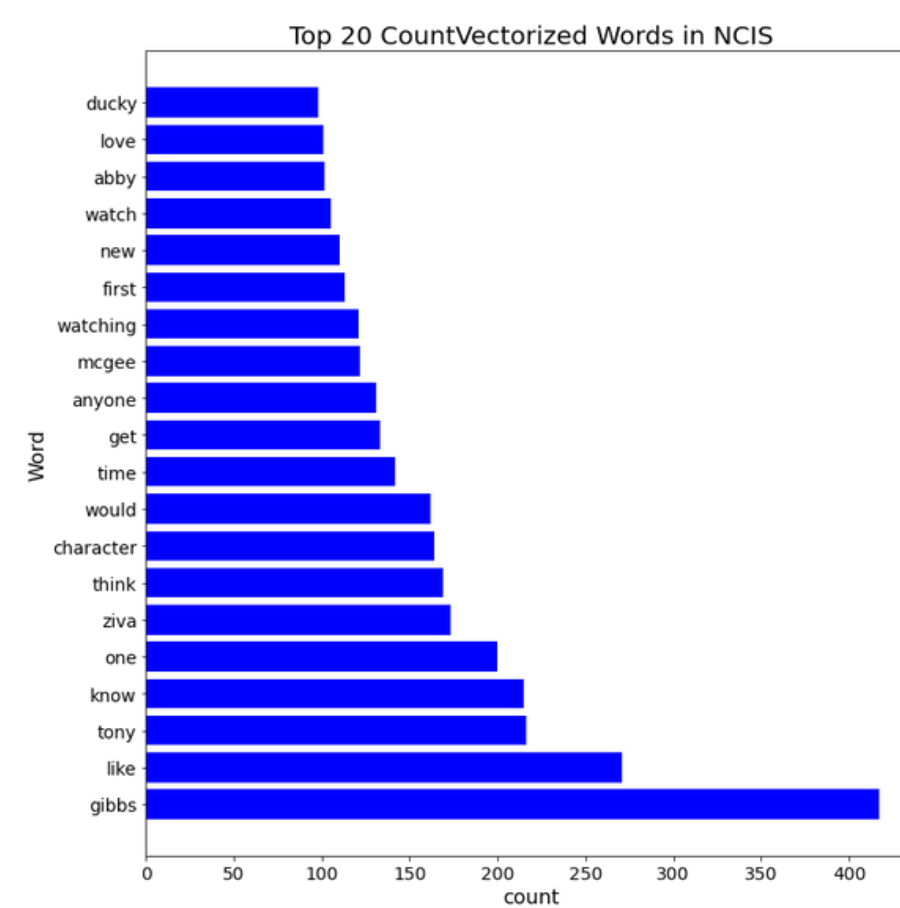


TF-IDFVectorized

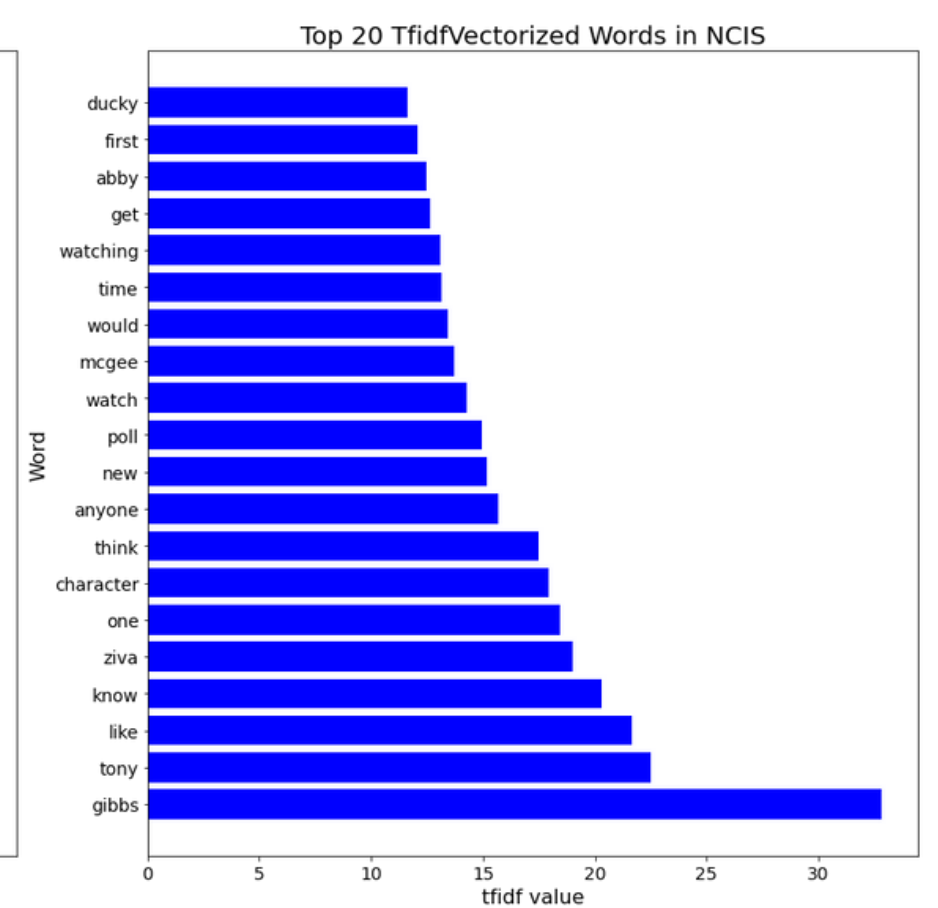


NCIS

CountVectorized



TF-IDFVectorized



EDA: Wordcloud




911FOX

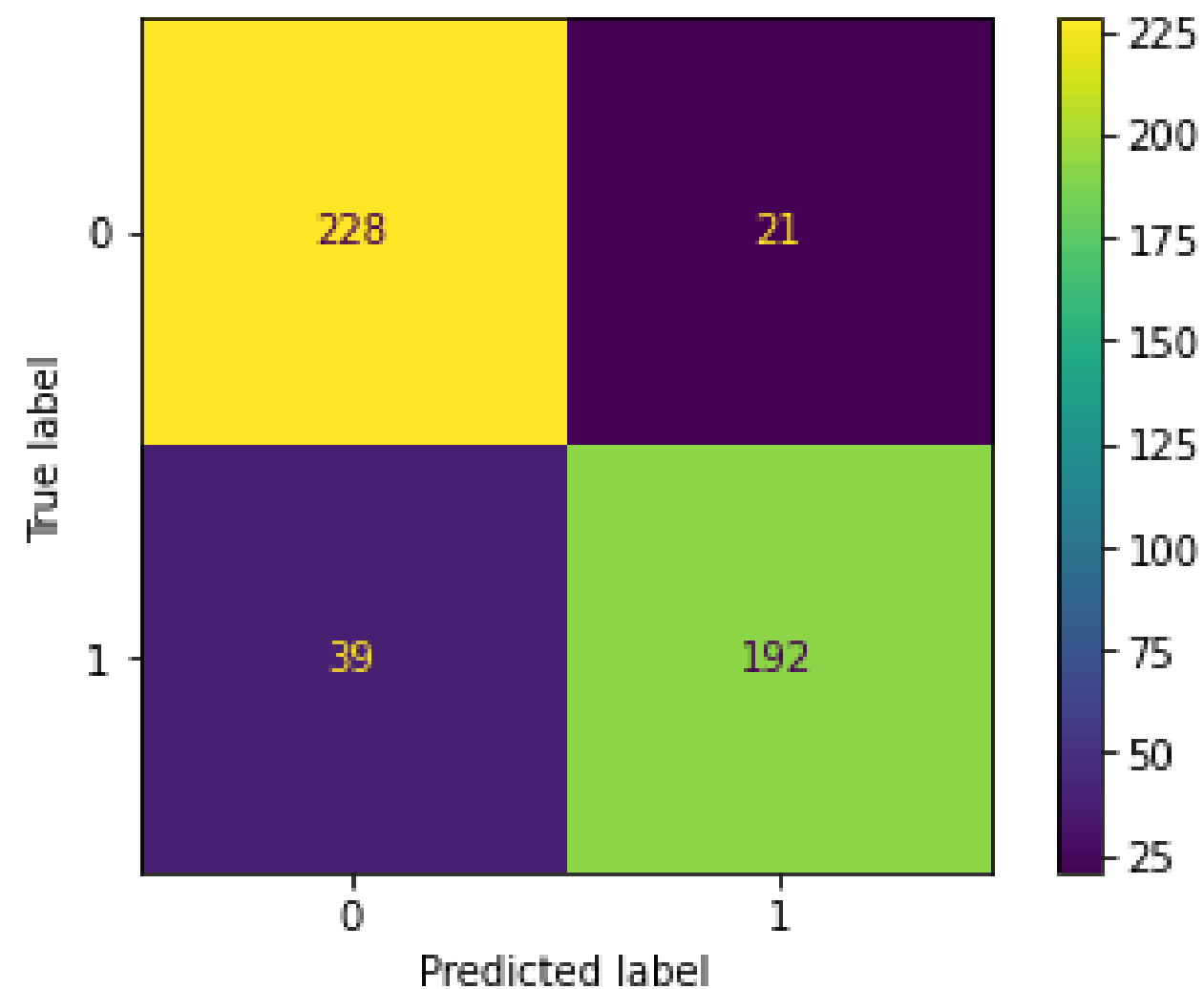


NCIS

Model

| Model | | Accuracy on training data | Accuracy on unseen data |
|-----------------|--------------------------|---------------------------|---|
| Transformation | Classification | | |
| CountVectorizer | Logistic Regression | 97.9% | 84.8% |
| | Naive Bayes | 94.3% | 86.7% |
| | Random Forest Classifier | 99.6% | 86.4% |
| TF-IDF | Logistic Regression | 97.0% | 86.9% |
| | Naive Bayes | 96.3% | 87.5%  |
| | Random Forest Classifier | 99.9% | 85.2% |

Model Evaluation



Accuracy: 87.5%

Misclassification rate: 12.5%

Sensitivity: 83.1%

Specificity: 91.6%

Precision: 90.1%

Important word: 9-1-1



buck



like



one

4. star

5. lone

6. maddie

7. eddie

8. watch

9. hulu

10. anyone

Conclusions & Recommendations



- The best model in this work is the model by Naive Bayes with TF-IDF Transformation with 87.5% of accuracy.
- The top words show the popular characters are 'Buck', 'Maddie' and 'Eddie' and the popular season is 911: Lone Star, and the popular platform the user post on 911FOX Reddit is the Hulu website.
- When we have the model to predict the subreddit for each post, Reddit can improve the searching system to receive the input from the user as the specific word and recommend the relevant subreddit.
- When Reddit has more accuracy of the post in each subreddit, they can make the business deal with another website. In this case, we know the 911FOX has the post about Hulu and the Hulu is the top 10 words which have high coefficient so Reddit can make a deal with the Hulu website to increase the value of Reddit and the number of users.