Using NLP to Predict Posts from Reddit

Dennis Tran

Problem Statement

A Multimedia Company who owns several Youtube channels specializing on reposting narrated Reddit posts seeks to automate its post collection.

Of interest are two subreddits, r/NoSleep and r/IDontWorkHereLady, known for user submitted stories.

Our team's task is to create a machine learning model that can determine where a post came from.

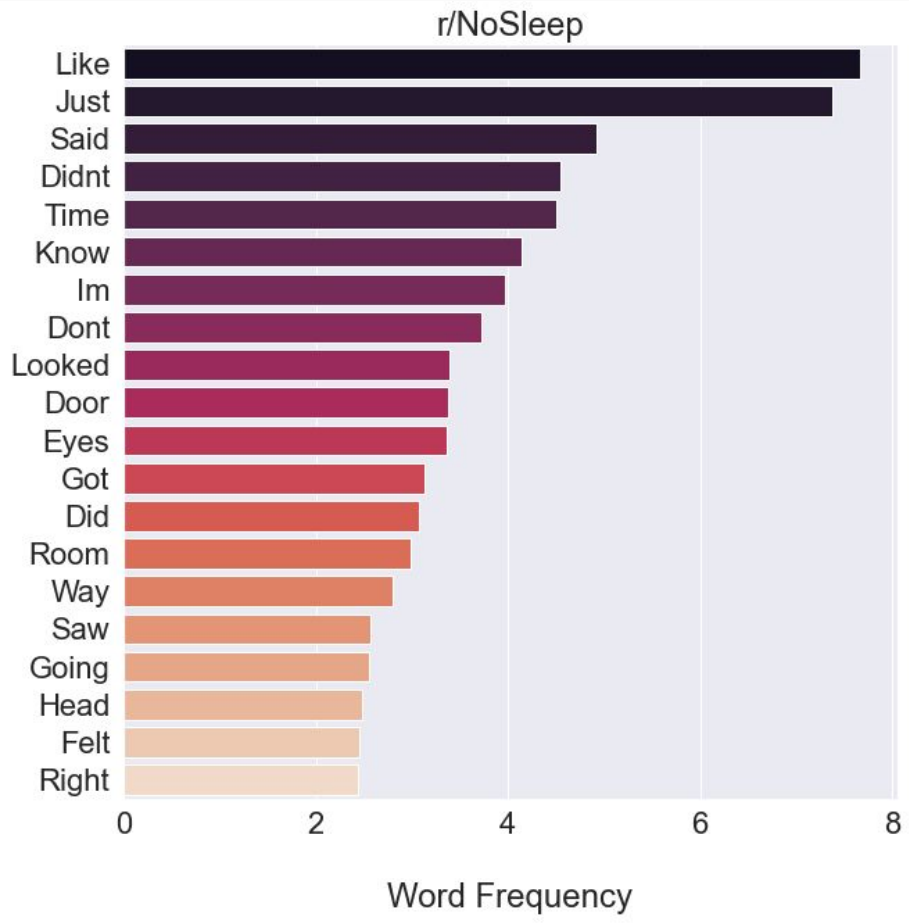
The Process

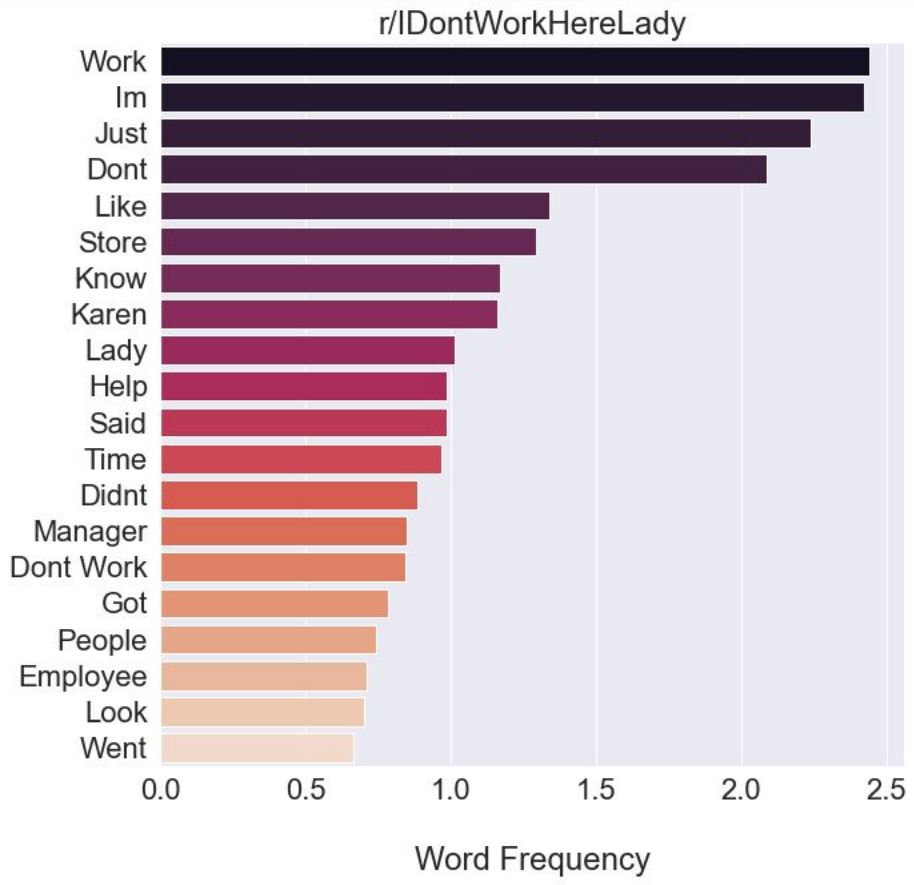
We scraped our data with Pushshift

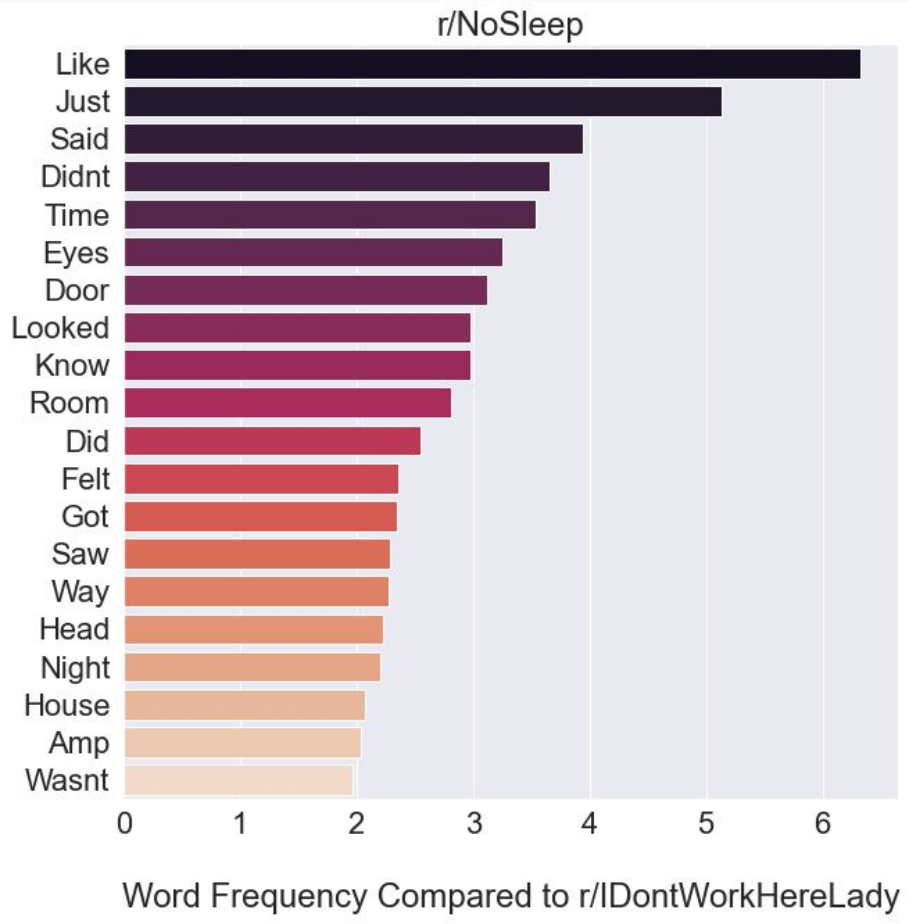
Cleaned as we scraped and after

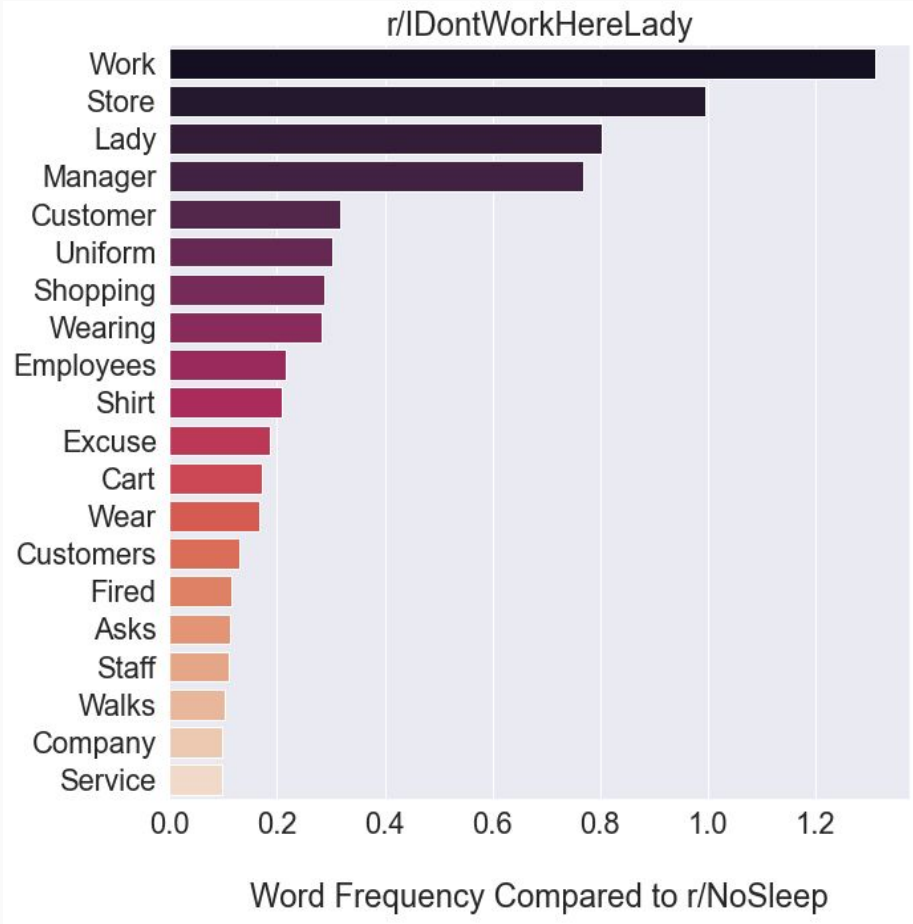
Tokenized then lemmatized

Modeled


r/NoSleep
Like
Just
Said
Didnt
Time
Know
Im
Dont
Looked
Door
Eyes
Got
Did
Room
Way
Saw
Going
Head
Felt
Right
0
2
4
6
8
Word Frequency


r/IDontWorkHereLady
Work
Im
Just
Dont
Like
Store
Know
Karen
Lady
Help
Said
Time
Didnt
Manager
Dont Work
Got
People
Employee
Look
Went
0.0
0.5
1.0
1.5
2.0
2.5
Word Frequency


r/NoSleep
Like
Just
Said
Didnt
Time
Eyes
Door
Looked
Know
Room
Did
Felt
Got
Saw
Way
Head
Night
House
Amp
Wasnt
0
1
2
3
4
5
6
Word Frequency Compared to r/IDontWorkHereLady


r/IDontWorkHereLady
Work
Store
Lady
Manager
Customer
Uniform
Shopping
Wearing
Employees
Shirt
Excuse
Cart
Wear
Customers
Fired
Asks
Staff
Walks
Company
Service
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Word Frequency Compared to r/NoSleep

Modeling

Tested for accuracy

Baseline 50%

Used Ridge and Lasso Regression, Multinomial Naive Bayes, and Random Forest

All models except Naive Bayes overfit

However, we expected much more severe overfitting than what resulted

	Training Score	Testing Score
Logistic_Regression	1.000000	0.976471
Rand_Naive_Bayes	0.993464	0.988235
Grid_Naive_Bayes	0.993464	0.988235
Lasso	1.000000	0.976471
Modified_LogReg	1.000000	0.968627
TFIDF_Random_Forest	1.000000	0.988235
CVec_Random_Forest	1.000000	0.956863

Conclusion

The Multinomial Naive Bayes performed the best, with an accuracy score of 98.8%.

May be in part due to the different nature of the subreddits despite both containing written stories.

Further testing and tuning could be done on the Bayes model, but we must consider if the cost is worth it.