Scraping Subreddits

Hitha Yeccaluri
DSI-R, Project 3

Can we filter fake stories from real ones?

What are we comparing?

r/nosleep

“Nosleep is a place for redditors to share their scary personal experiences. Please read our guidelines in the sidebar/"about" section before proceeding.”

r/LetsNotMeet

“A place to read spine-tingling, unusual, terrifyingly true stories about people you never want to meet again.”

Are they really as similar as they sound?

Data Collection and Cleaning

Scraped 100 posts per subreddit per month from January 2017 to July 2021

Lost a lot of data points, but the heavy regulation is good for training models

Ended up with 2146 posts from r/LetsNotMeet and 3865 from r/nosleep

Dataframe after collection
(11000 submissions, 91 features)

```
In [4]: sub.shape
Out[4]: (11000, 91)
```

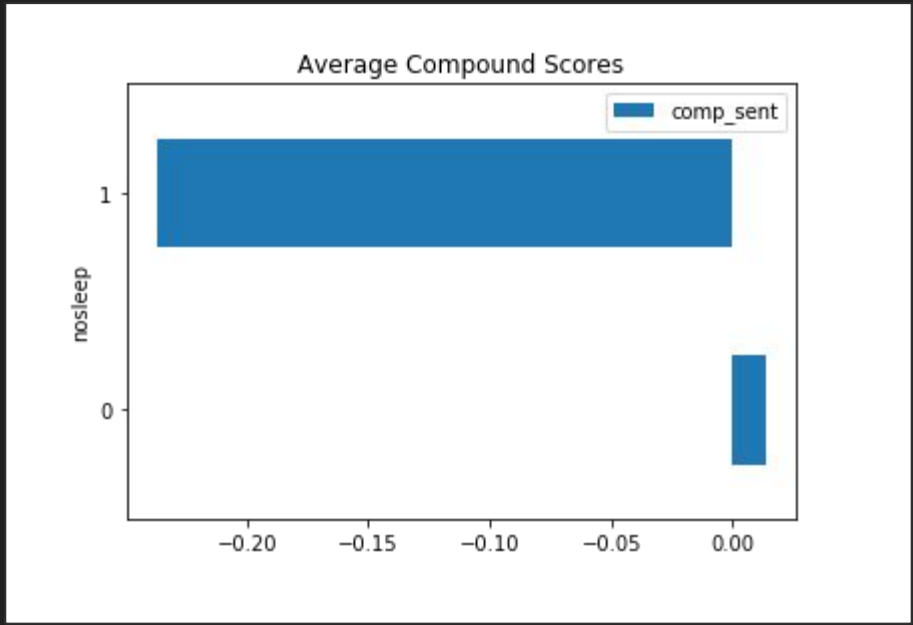
Dataframe after dropping [removed] and [deleted] posts
(6011 submissions, 5 features)

```
In [15]: sub.shape
Out[15]: (6011, 5)
```

Polarity

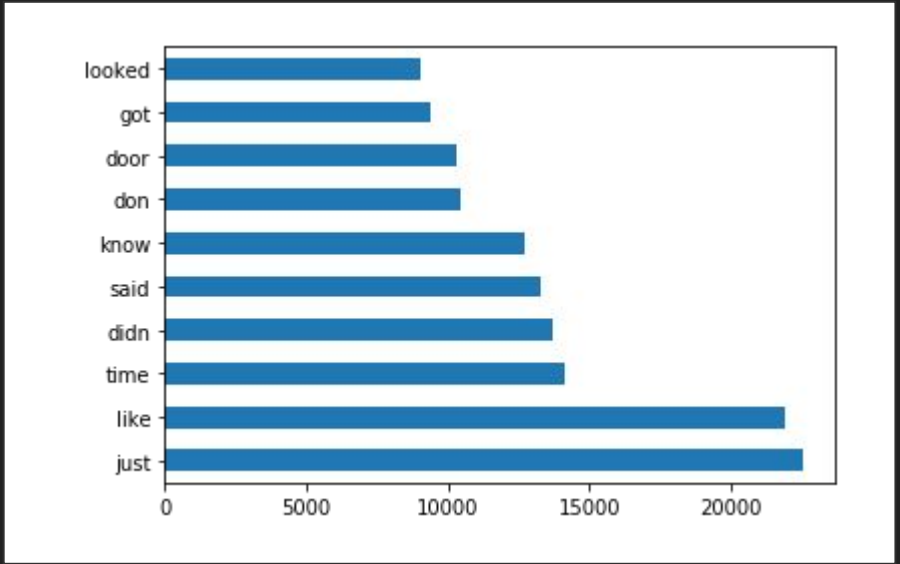
nosleep had a significantly stronger overall sentiment, lot of individual stories reached -0.9 compound scores

LetsNotMeet had a slight positive trend, a lot of individual stories were almost neutral



Initial Word Overlap

Top ten words to appear across both subreddits (before major cleaning); these words would be of no help in differentiating between subreddits



Baseline to beat: 64.3%

	Training Score	Testing Score	Overfit
Random Forest Classifier (w/ CountVectorizer)	97.1%	90.1%	7%
Multinomial Naive-Bayes (w/ TF-IDF)	93.1%	92.1%	1%
Extra Trees Classifier (w/ CountVectorizer)	99.7%	90.02%	9.68%
K Nearest Neighbors (w/ CountVectorizer)	75.7%	66.6%	9.1%

Accuracy and F1-Scores

Accuracy was the main metric of interest, simply because it is easier for people to immediately interpret

F1 scores are important to note here, though--we want to know how many fake stories we are correctly classifying as fake.

	Accuracy	F1-Score
RFC	90.3%	92.7%
MNB	92.1%	93.9%
ETC	90.01%	92.6%
KNN	66.6%	68.4%

Conclusions

Most fake stories can be identified, but some are going to slip through the cracks.

Human intervention is still needed, but the NMB model can be helpful in the initial checks.