# Predicting Myers-Briggs Type Indicator with Text Classification

**Team: Ian Knight, Rayne Hernandez** *CS 224N: Natural Language Processing with Deep Learning*

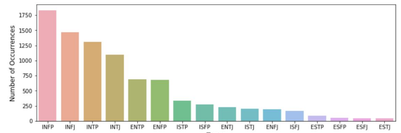
## Problem Description

- The **Myers-Briggs Type Indicator (MBTI)** is a psychological model of human personality with **16 possible personality types**
- Personality types are a combination of **4 different binary categorizations** (e.g., Introverted vs. Extroverted)

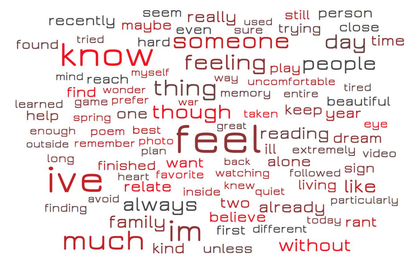
- Using **~430,000 social media posts** from the web forum **PersonalityCafe.com**, we trained a text classifier capable of **sorting people** into their personality type based on their social media posts

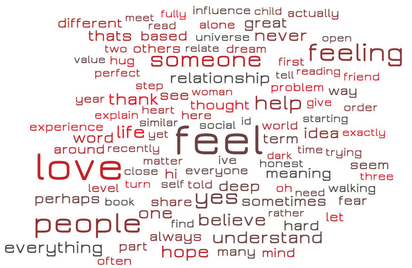
## Challenges

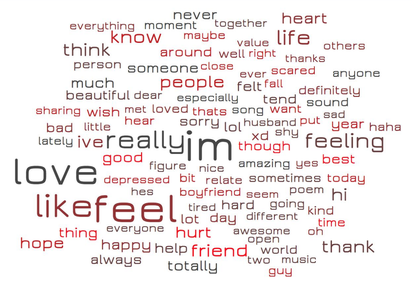
The **data** we acquired from PersonalityCafe tended to be **skewed towards certain personality types**, which forced us to artificially balance our data

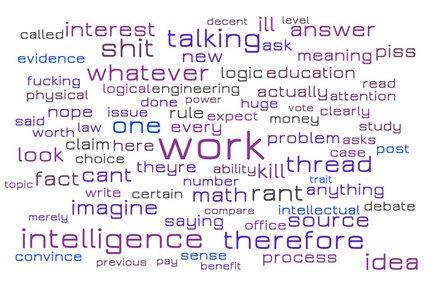


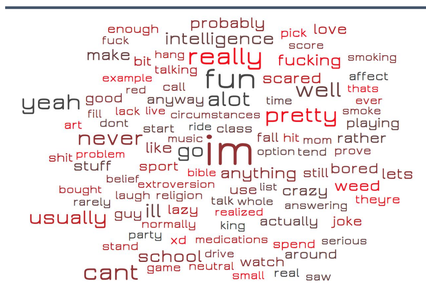

## Extrapolation: Word Clouds



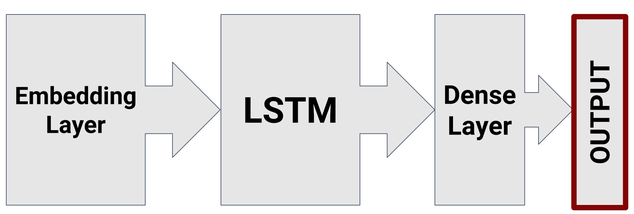



## Approach

### Preprocessing

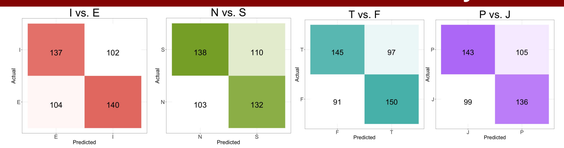
- Proportionality
- Selective Word Removal
- Lemmatization
- Tokenization
- Padding

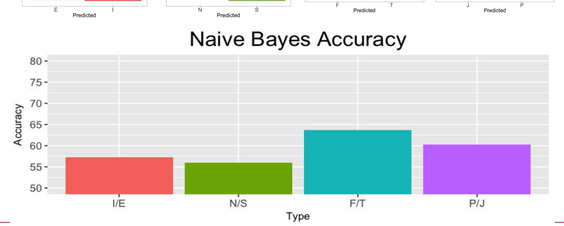
### Model



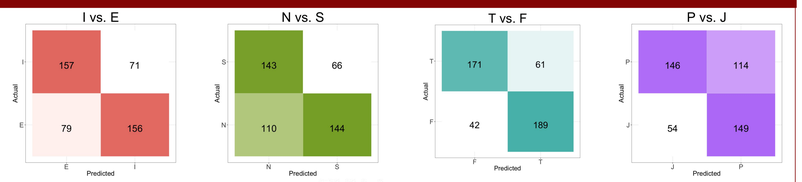

## Results

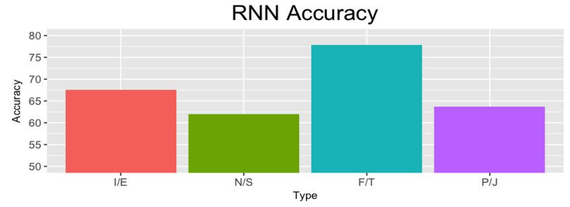
### Baseline: Multinomial Naive Bayes




### Recurrent Neural Network




## Extrapolation: Donald Trump's Twitter Account

- As a real life test case of the hypothetical capability for greater abstraction with increased quantity of text data, we decided to predict the MBTI of Donald Trump using 30,000 of his tweets
- The result matched what MBTI experts claim to be Trump's type.
- ESTP is the archetype known as "the Entrepreneur"

```
IE
0.5533546805381775
NS
0.5134227871894836
FT
0.5419855117797852
PJ
0.47002798318862915

Final prediction: ESTP
```

## Conclusion

- Our work suggests that there is an **underlying linguistic component** to MBTI that is **not fully captured** by means of traditional assessment methodology
- Our model offers fast MBTI classification of massive amounts of publicly available text online
- our model represents a trade-off of between two aspects: we achieve lower rates of perfect classification in exchange for higher rates of approximately correct classification (i.e. "good" classification). This latter aspect is a positive feature that multiclass models fail to achieve

## Acknowledgements

We would like to thank Professor Richard Socher for being our advisor.