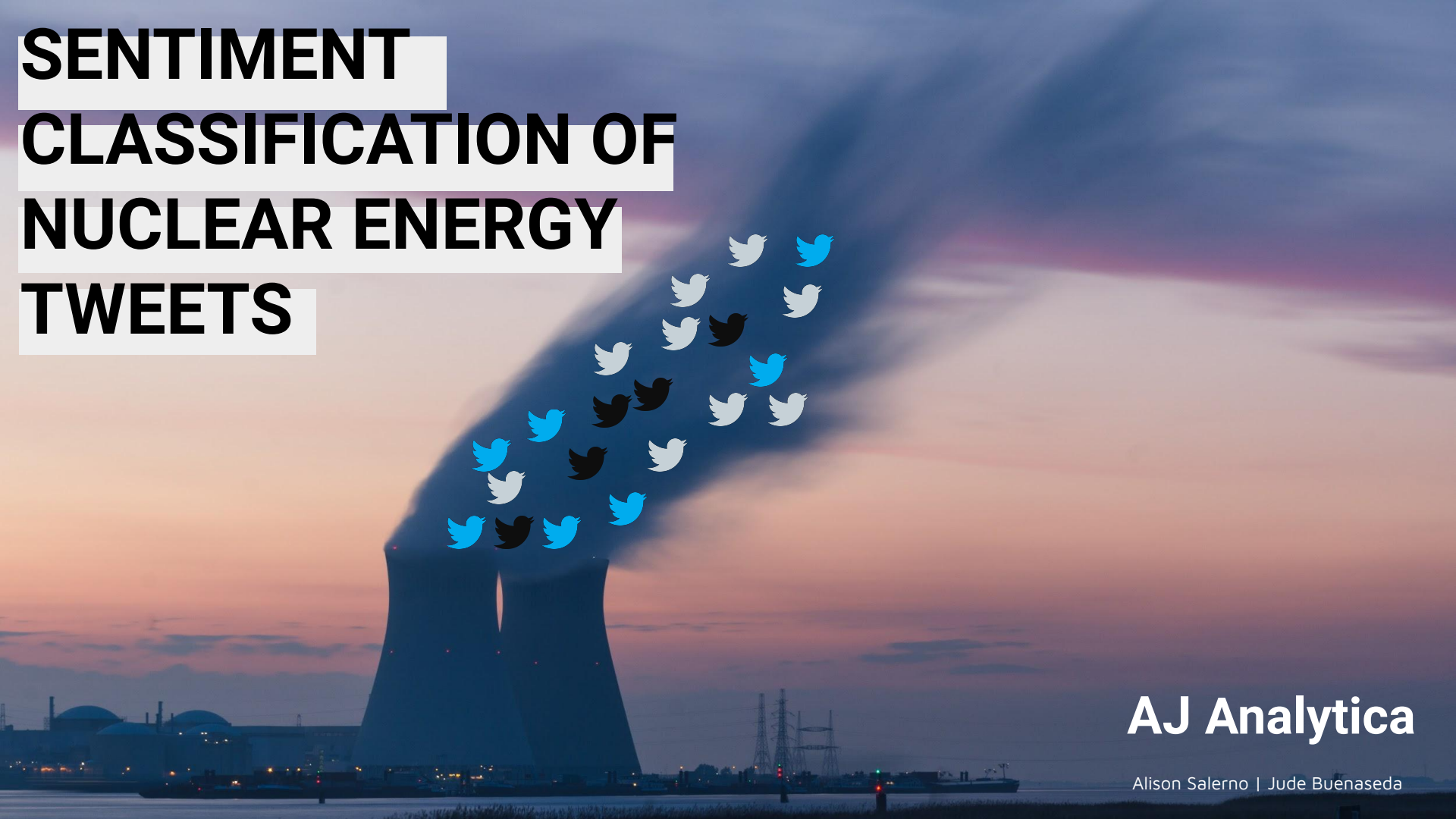
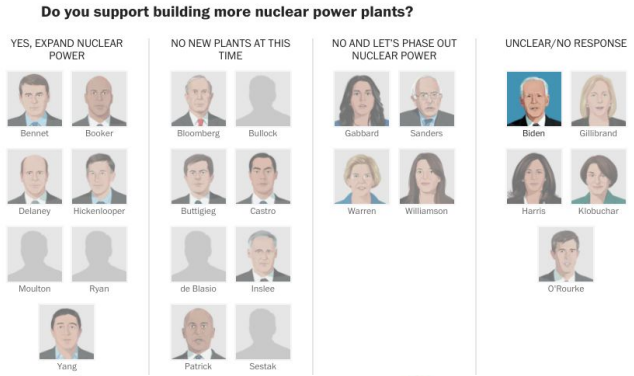
SENTIMENT CLASSIFICATION OF NUCLEAR ENERGY TWEETS

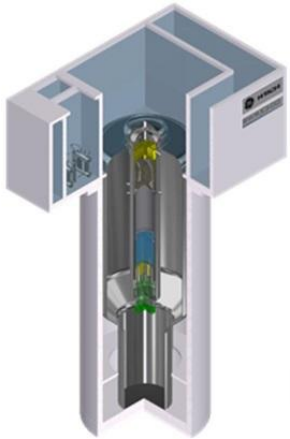


AJ Analytica

Alison Salerno | Jude Buenaseda

INTRODUCTION





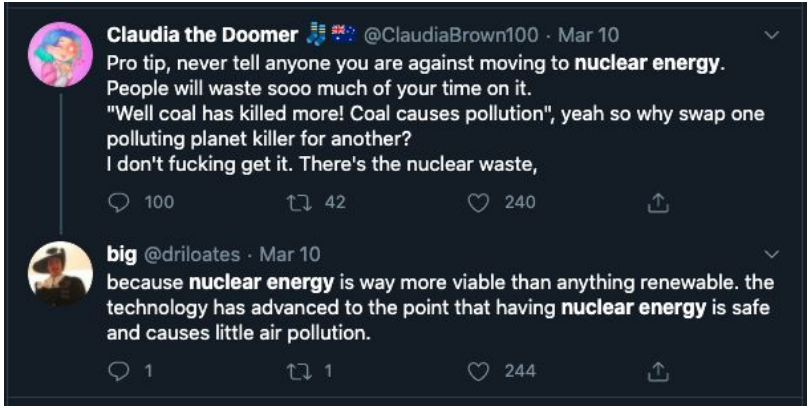
HITACHI

BUSINESS CASE: Identify key words associated with varying sentiment for greater messaging awareness

TASK: Classify the sentiment of a certain tweet around nuclear energy and extract value

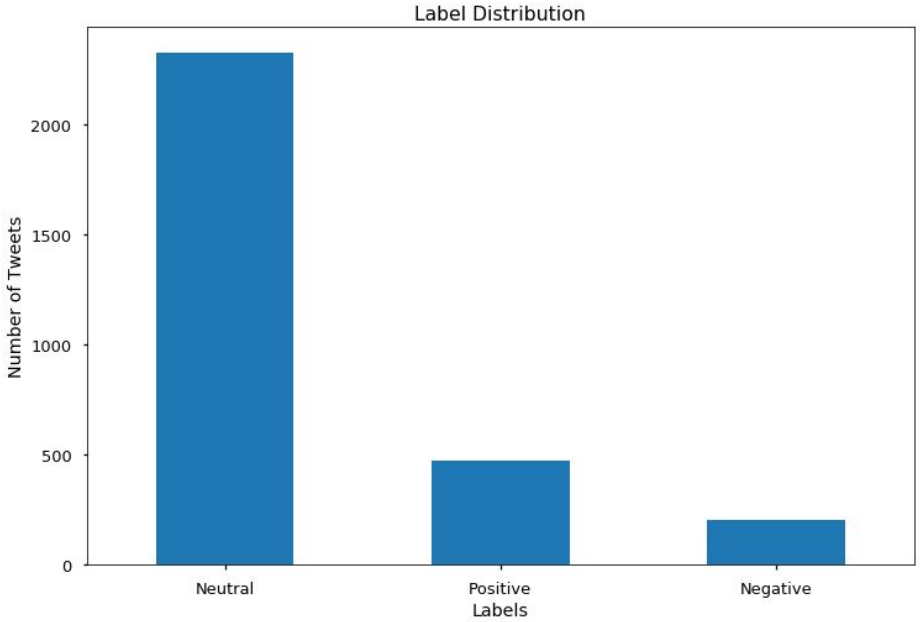
DATA

FEATURES



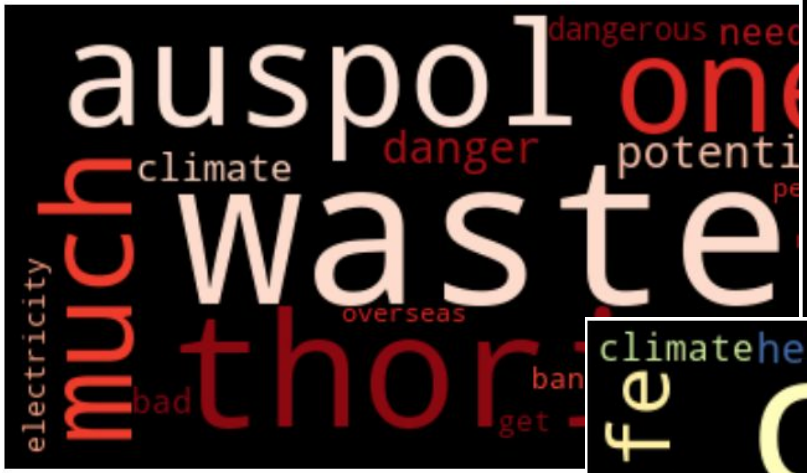
3,000 tweets under the search term of nuclear energy

TARGET CLASSES

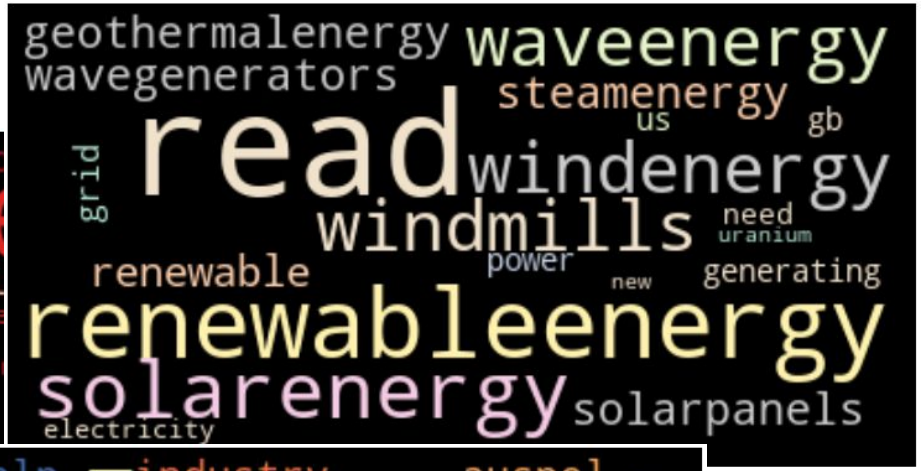


WORDCLOUD

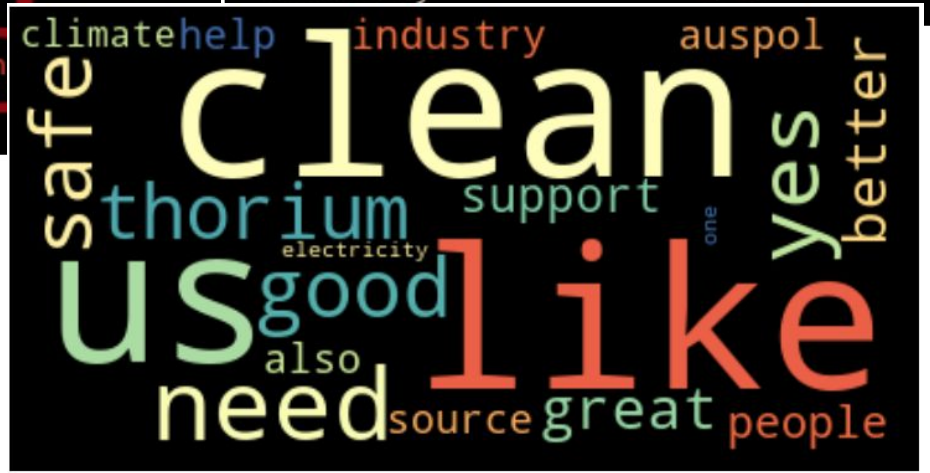
NEGATIVE



NEUTRAL



POSITIVE



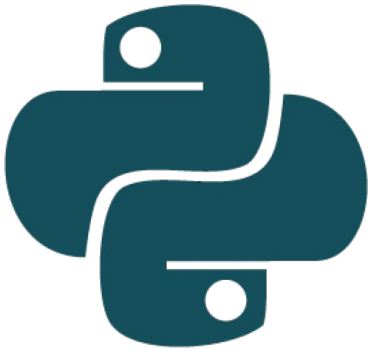
FEATURE ENGINEERING

- English stop words were removed
- 'Nuclear Energy' , 'nuclearenergy', 'energy' were also removed
- Punctuation, links and non-alphabetical characters were also scrubbed from the tweets
- Each tweet was vectorized using TF-IDF

```
tf_idf_test_df = pd.DataFrame(tfidf_data_test.toarray(), columns=tfidf.vocabulary_.keys())
second_doc = tf_idf_test_df.loc[3]
second_doc.idxmax(axis=1)


second_doc['billgates']
```

```
0.3148415973979402
```



NLTK

MODELLING

77.7%

Dummy Classifier

- Baseline Model
- Strategy: 'most_freq'

78%

Naive Bayes Classifier

- First true text classifier model
- Does not account for class imbalance

78.2%

Random Forest Classifier

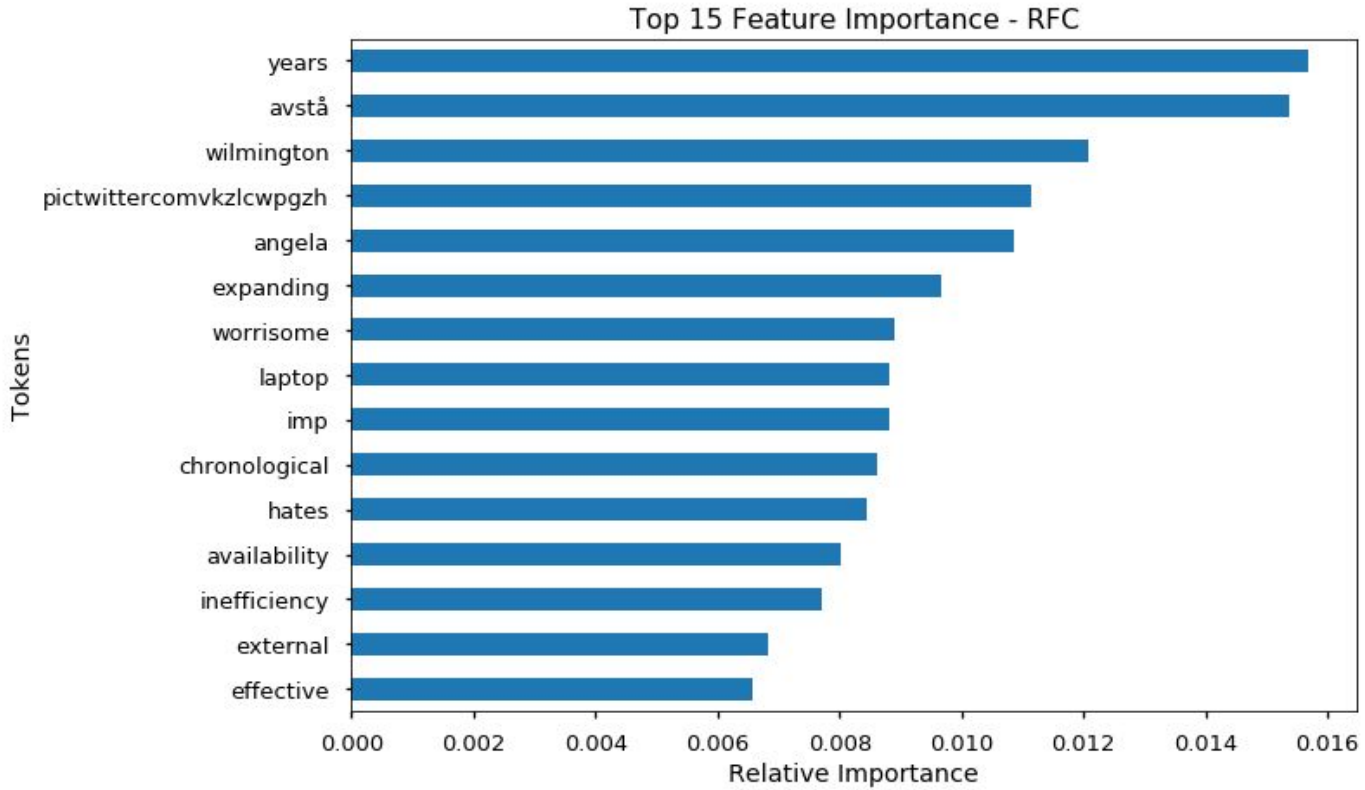
- N_estimators = 300
- Balanced weight class
- Slightly overfit with 99% accuracy in training data

82.1%

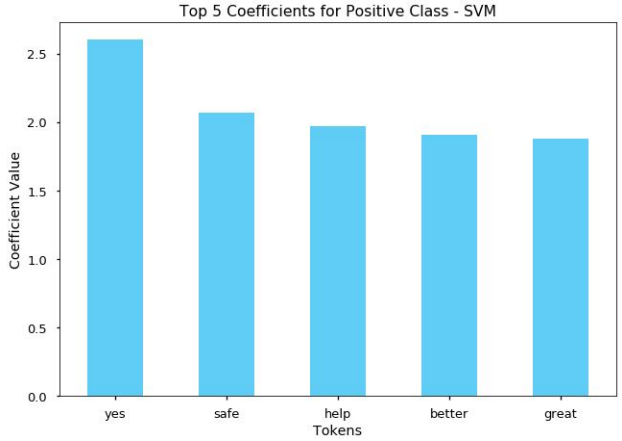
Support Vector Machine

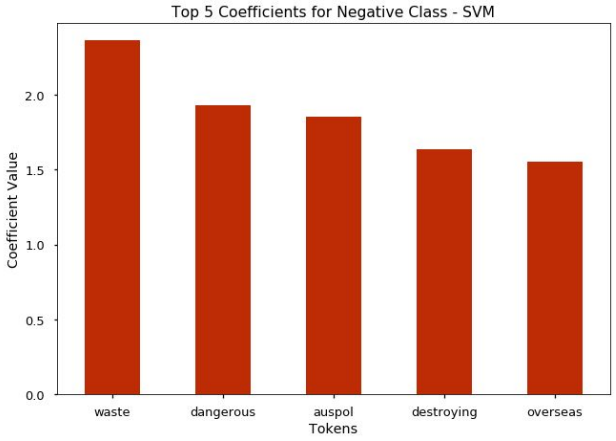
- Linear model of SVM
- Balanced weight class
- Accuracy, Precision and Recall were all the highest

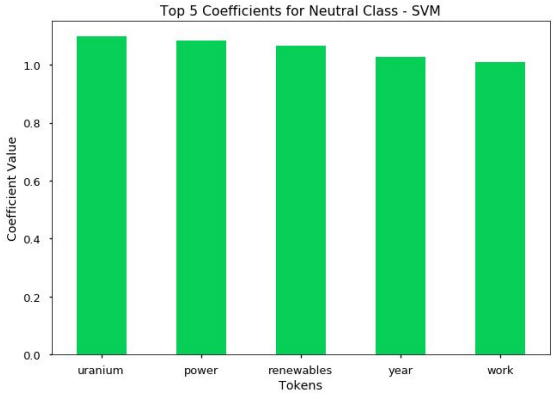
Random Forest Feature Importance



SVM Coefficients by Class



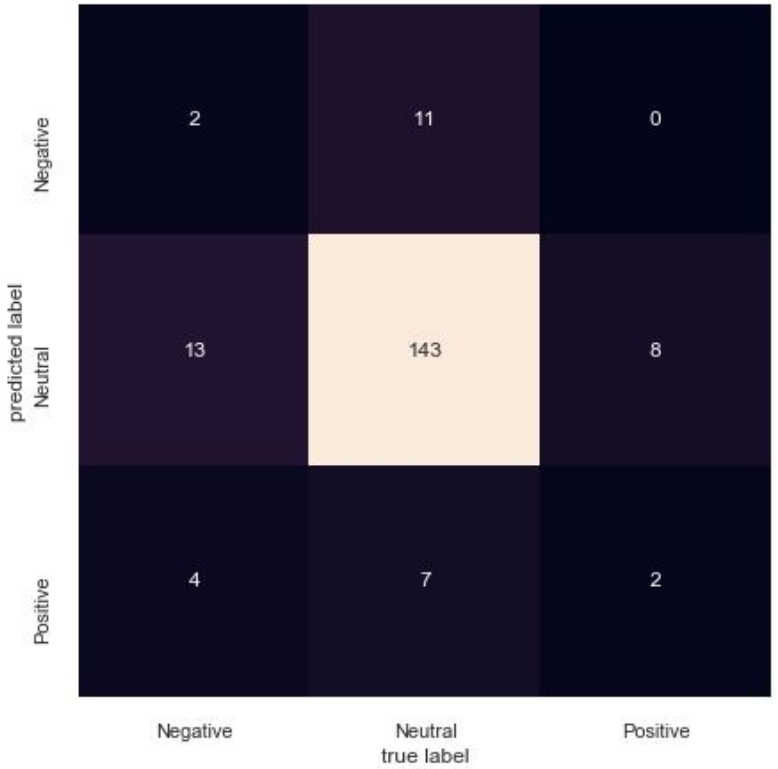




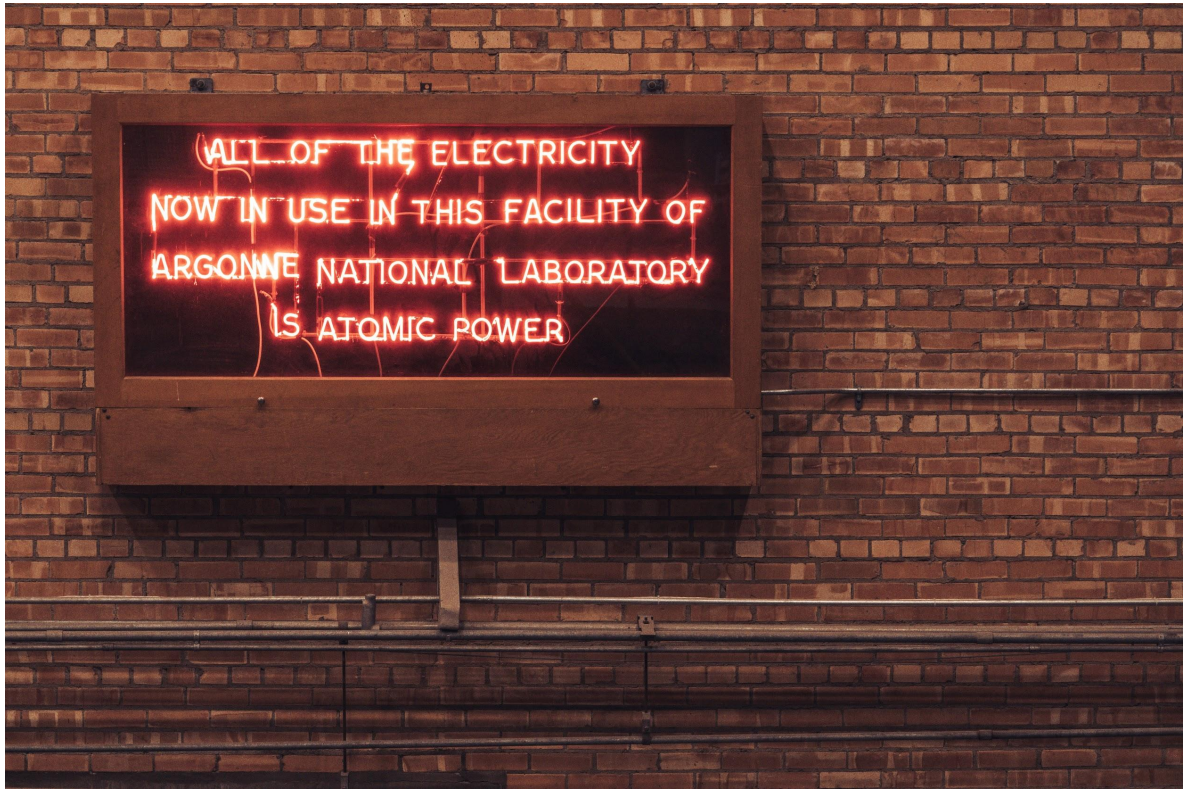
SURVEY COMPARISON

Testing our model with dataset from CROWDFLOWER

- 190 tweets on nuclear energy and sentiment was labeled via a survey taken by real people
- Compared how our label predictions measured against public opinion
- This validated the performance of our model in a real world application with a 77% accuracy



CONCLUSION



- “Waste” was the top key word for negative sentiment
- “Safe” was the top key word for positive sentiment
- “Renewables” was the top key word for neutral sentiment

NEXT STEPS

- Add additional features to our model such as numerical or categorical features:
  - User follower count, length of a tweet, blue check mark (Y/N)

- Explore deep NLP and neural net models:
  - Word embeddings
  - Word2vec
  - Topic Modeling
  - RNN