Twitter Sentiment Analysis

By Mohammad Abou-Ghazala

Roadmap

1. Business Problem

2. Data Summary and Preparation

3. Modeling and Analysis

4. Summary and Recommendations

Business Understanding

- Company wants to factor sentiments expressed on social media about our products and services

- Positive sentiments matter more than Non-Positive

- Need binary classifier

Questions for us:

a. **How successful are we at identifying positive tweets?**
   - Are Positive tweets falling through the cracks?

Data Understanding

- Dataset (2013)
  a. 9,000 Tweets
  b. 3 Columns ("Text", "Directed At", "Sentiment Expressed")
- **TARGET**: Positive vs. Not Positive
  a. Data Imbalance:
    - Positive - 33%, Not Positive - 67%
- False Positives more significant than False Negatives

Metrics: Precision Score

High Precision Score =

Low False Positive Rate

Data Preparation

- Consolidated the Data
  a. Turned target column into binary
    - 0 = Not Positive, 1 = Positive

- Splitting the Data into different sets
  a. Allows for multiple testing runs to minimize impact of "flukes"

- Model Experimentation
  a. Testing different predictive models to see **which has the highest Precision score**

Data Modeling & Analysis

Modeling

Critical Questions:

- How successful were we in keeping Not Positive sentiments out of the Positive category?

- Did we allow Positive sentiments to "slip through the cracks"?

Sentiment Analysis

a. Analyzing tweets
b. Identifying relationship between terms/words and sentiment

How successful were our models?

- **Precision** - Did Non Positives pollute our Positive category?

- **Recall** - Did we miss out on any Positives?
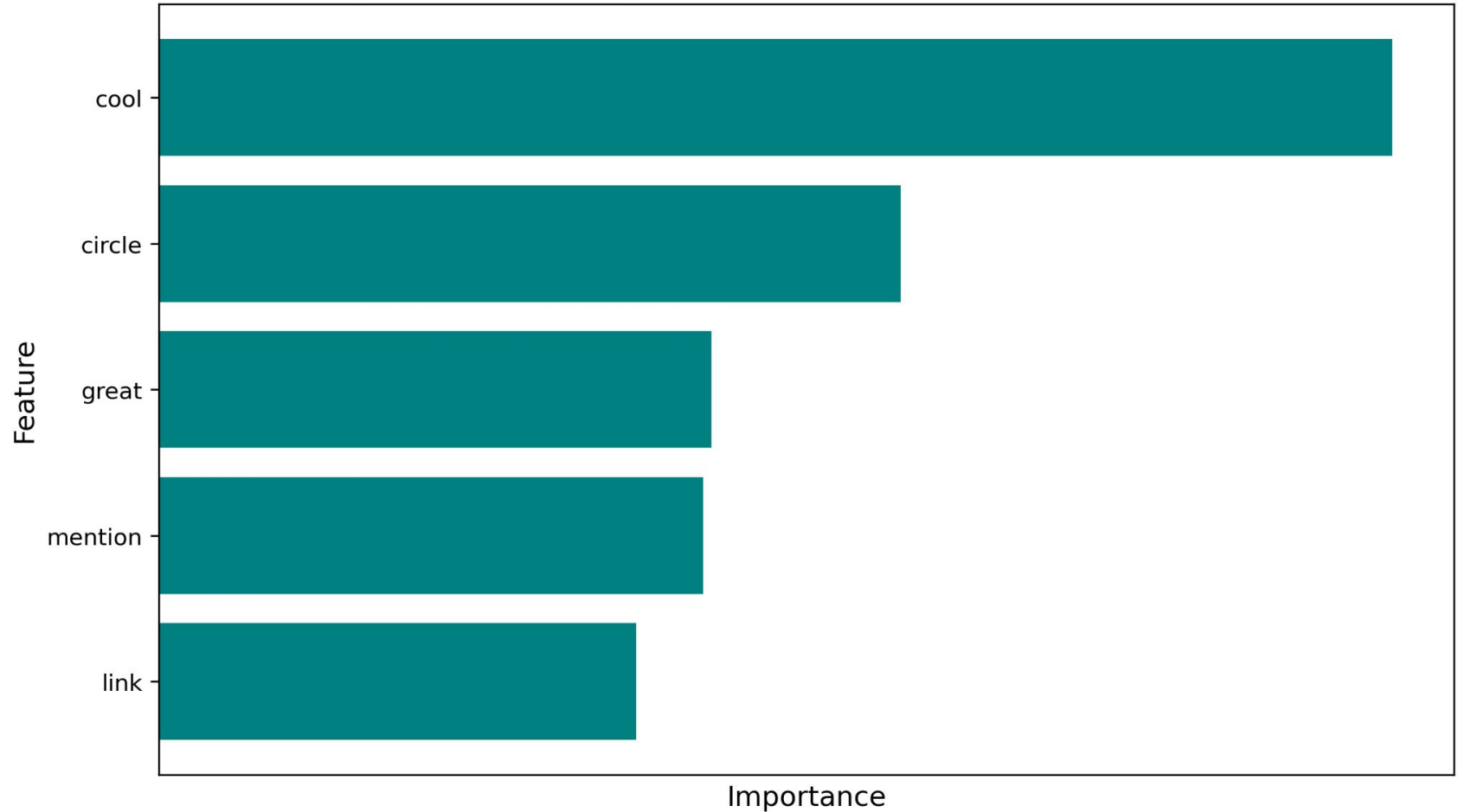
- **F1** - Balancing between Recall+Precision

Model Performance Comparison (Precision, Recall, F1)
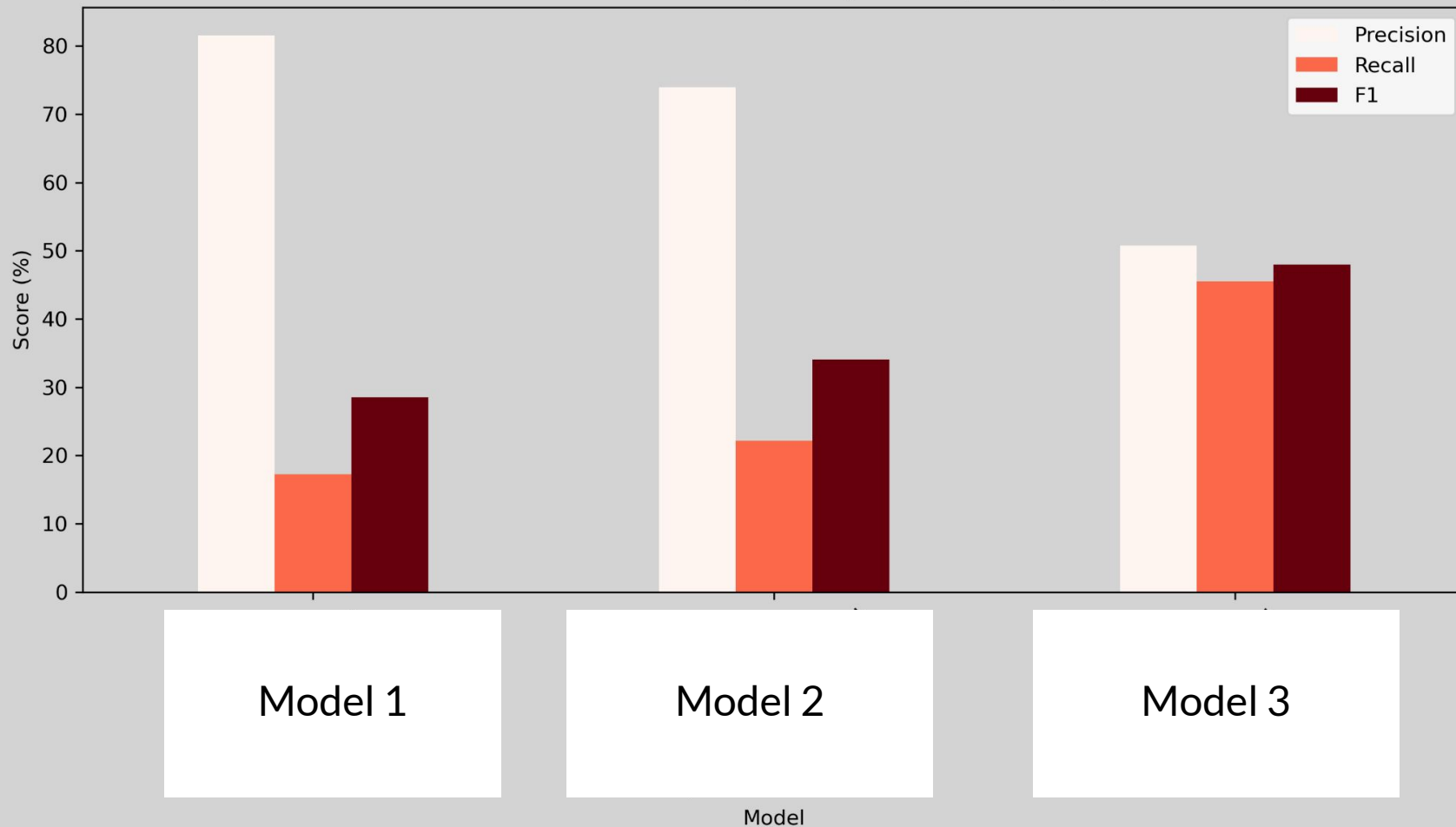
Were we able to identify terms to guide prediction?

Top 5 Feature Importances - Random Forest Classifier

Conclusion

1. Correctly identifying Positive sentiments while keeping out Not Positive Sentiments

2. However, allowing too many Positive sentiments to fall through the cracks

Conclusion

Limitations:

- Dataset: outdated, small, imbalanced

- Weak Hardware

- Time

Next Steps

Gather updated, larger dataset

Ensure data is well-balanced

Invest in superior hardware to make use of stronger models such as BERTweet

Questions?

Mohammad Abou-Ghazala
mohammadaboug@gmail.com
@MohammadAbouGhazala