CryptoCurrency vs StockMarket

Classification Model with Natural Language Processing

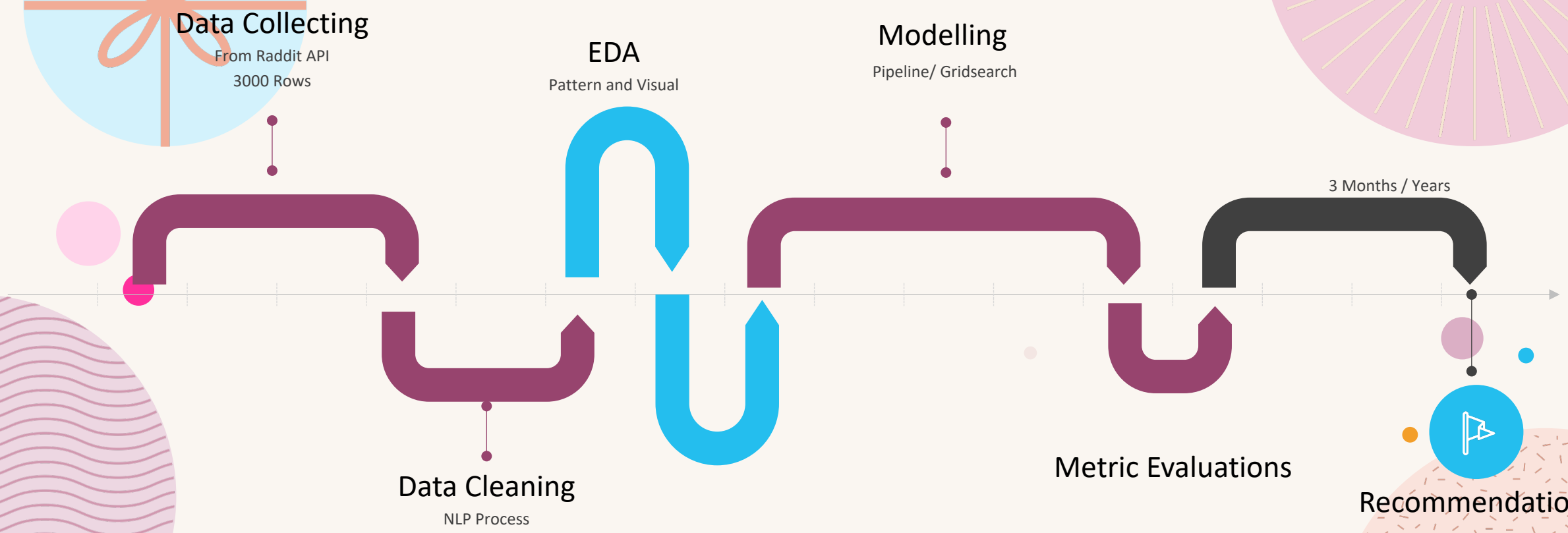
By Jay Li

By Jay Li

Overview

- Online text has become a very important datasource to analyze and provide business insights.
- In this project, data will be retrieved from Reddit API. Specifically subreddit r/CryptoCurrency and r/StockMarket.
- Help CryptoCurrencies Trading Software company to post their advertisement to target threads.

Workflow


Data Collecting
From Raddit API
3000 Rows
EDA
Pattern and Visual
Modelling
Pipeline/ Gridsearch
3 Months / Years
Data Cleaning
NLP Process
Metric Evaluations
Recommendations

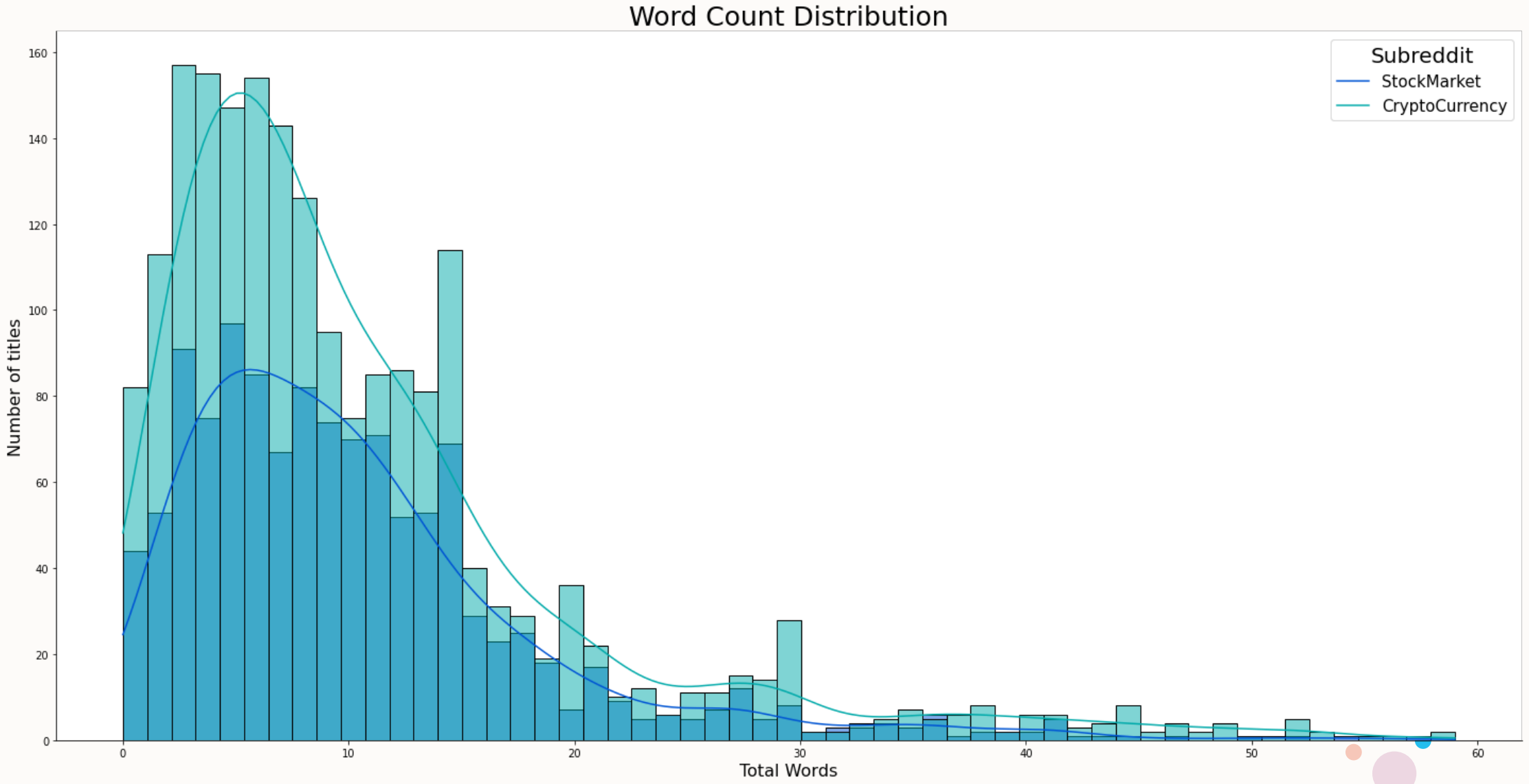
Data Collecting

- Data are collected directly from the existing Raddit.com API.
- Data collected in the same scale for subreddit topics. (r/StockMarket and r/CryptoCurrency)

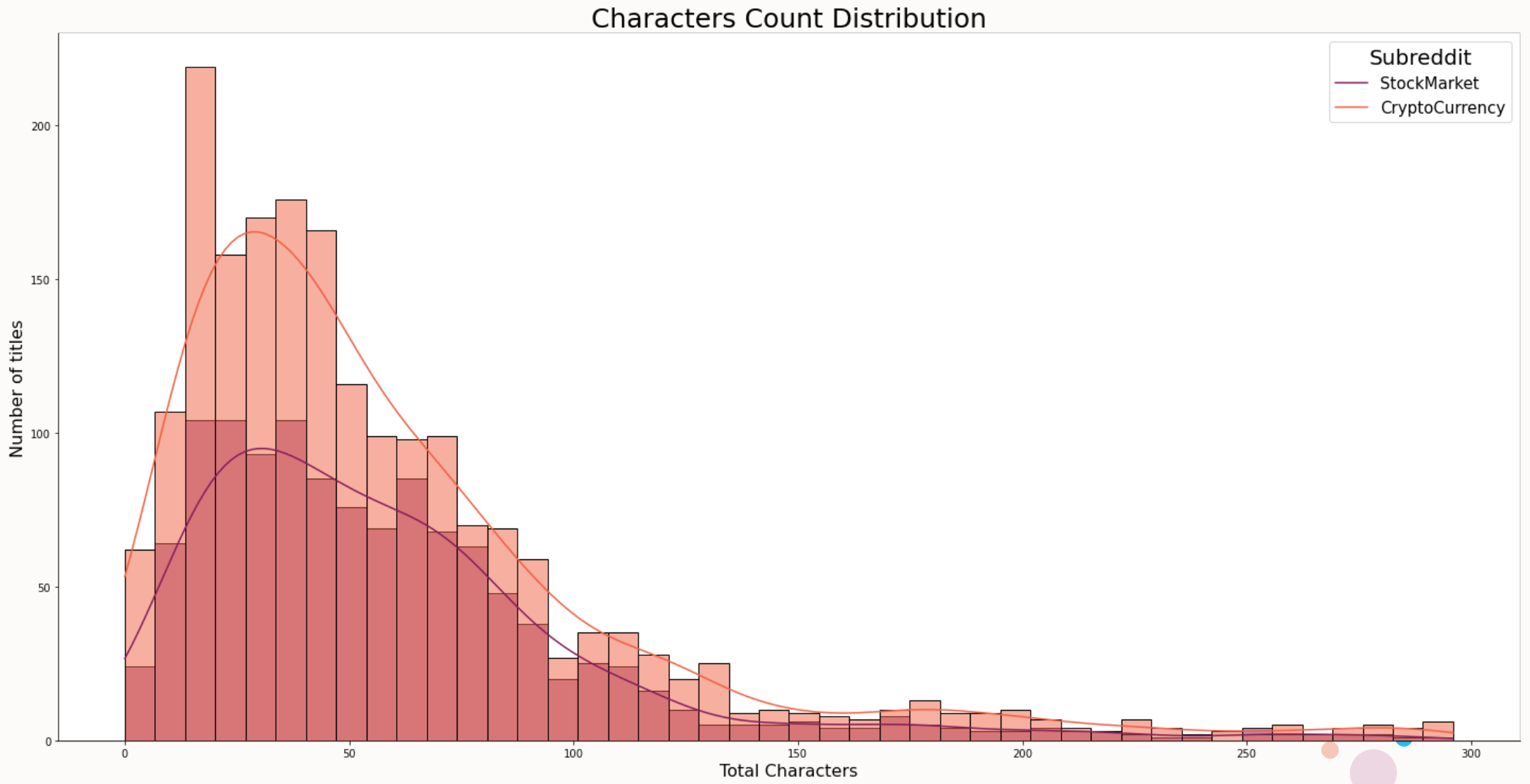
Data Cleaning

- NaN Data
- Emojis: 😀 😃 😄 😁 😆 😅 😂 🤣 🥲 ☺️ 😊 😇
- Puntuations : .,?/”!:’
- Contractions: we’ll, we will
- Lower cases: STOCK, stock
- Remove Hyperlinks: https://www.nyse.com/index

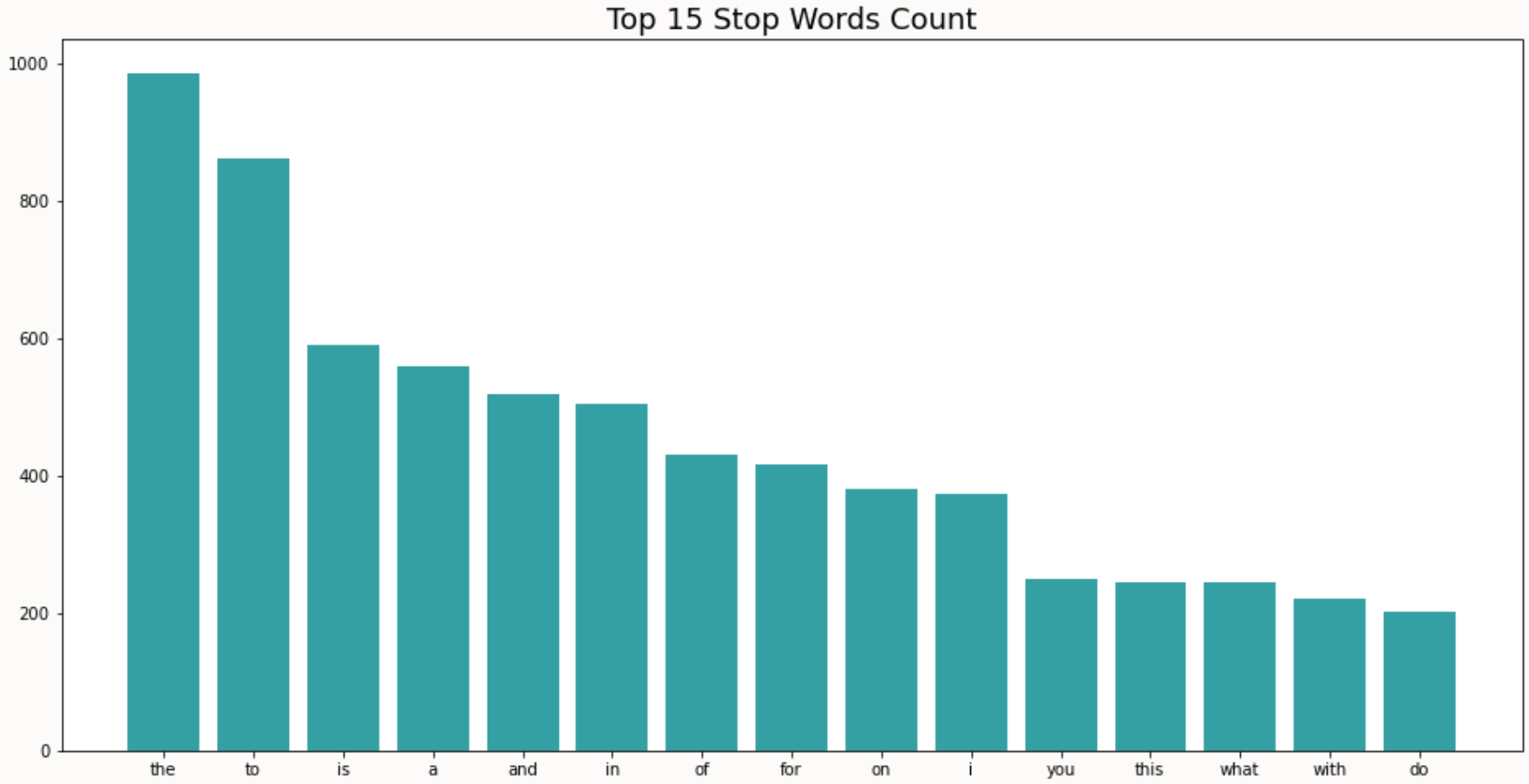
EDA: Word Count



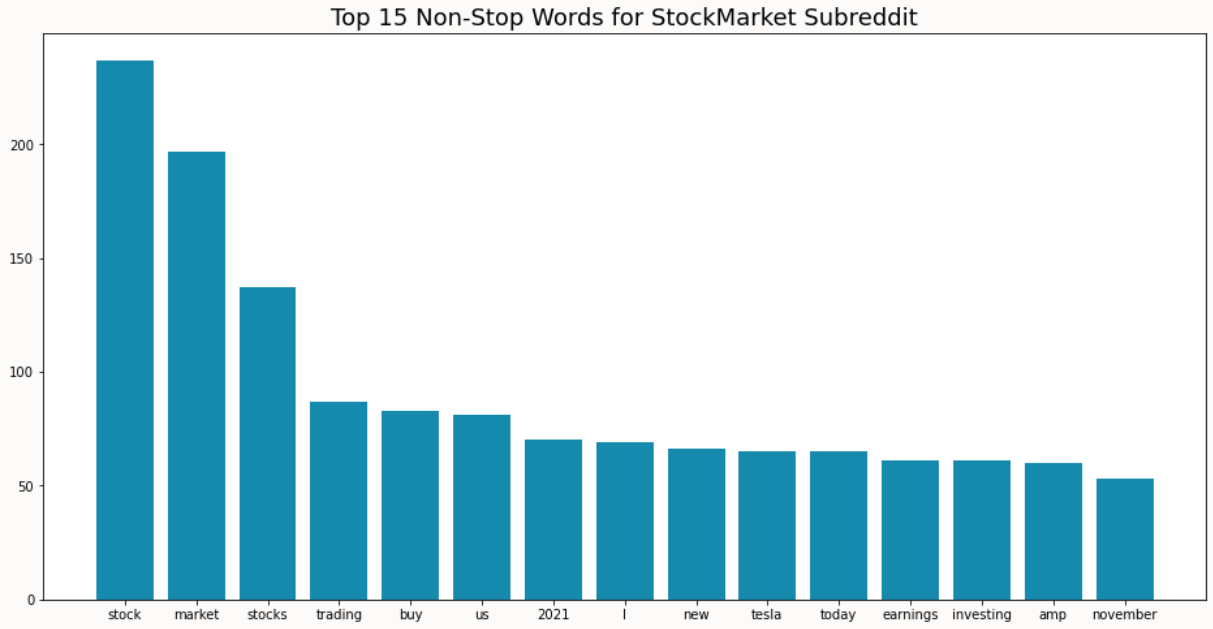
EDA: Character Count

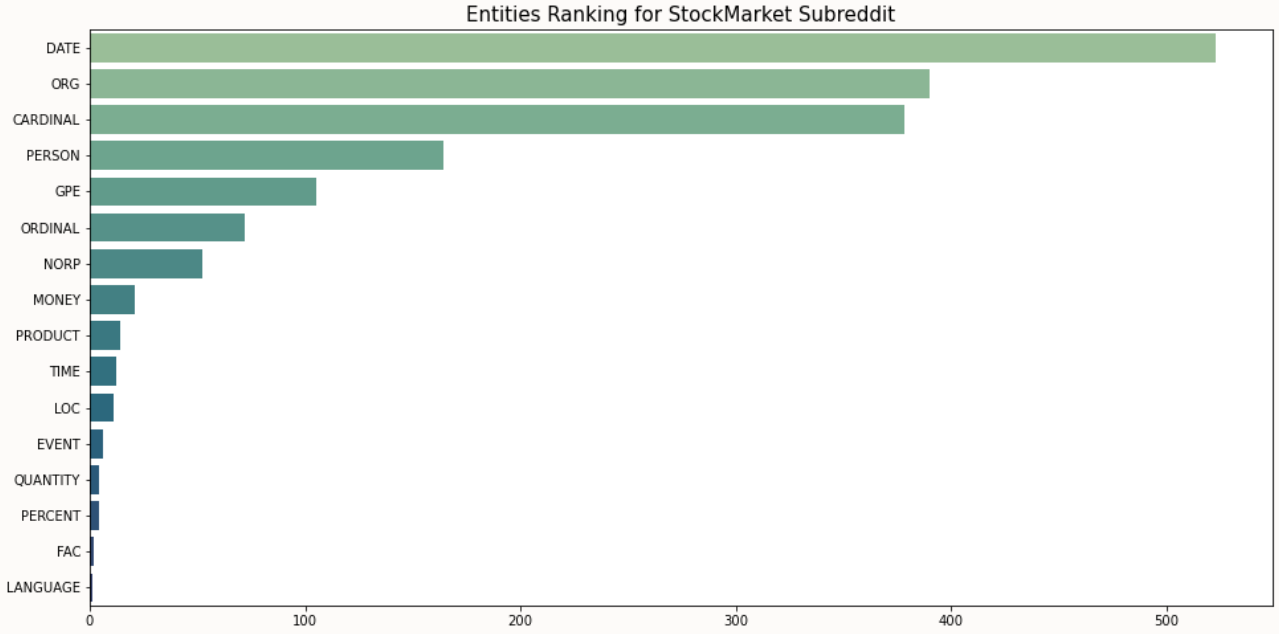


EDA: Stopwords Ranking

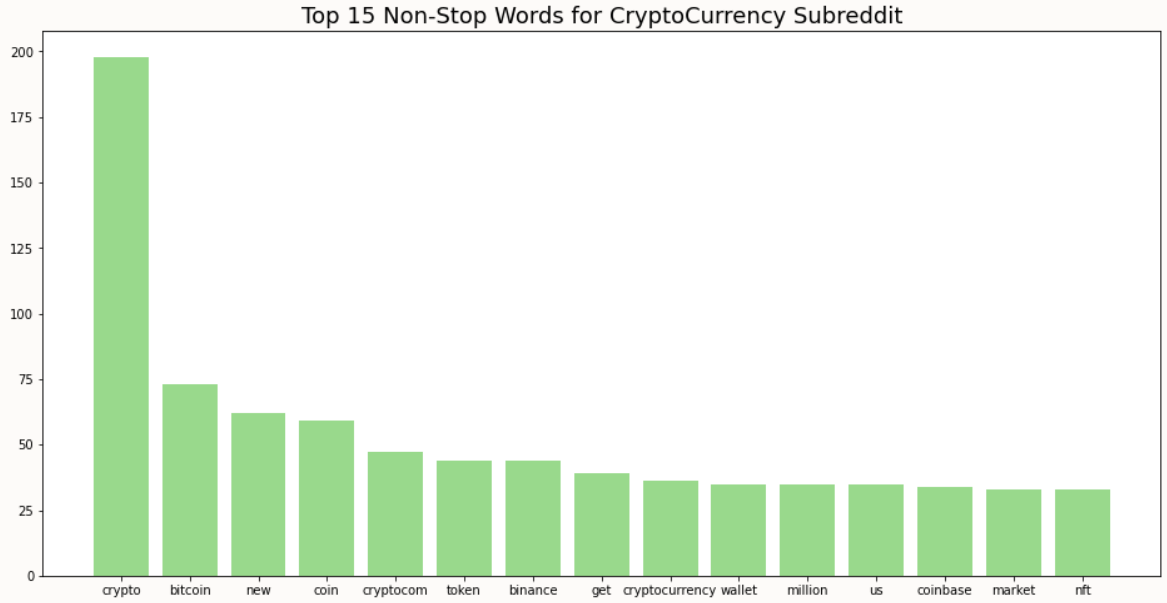


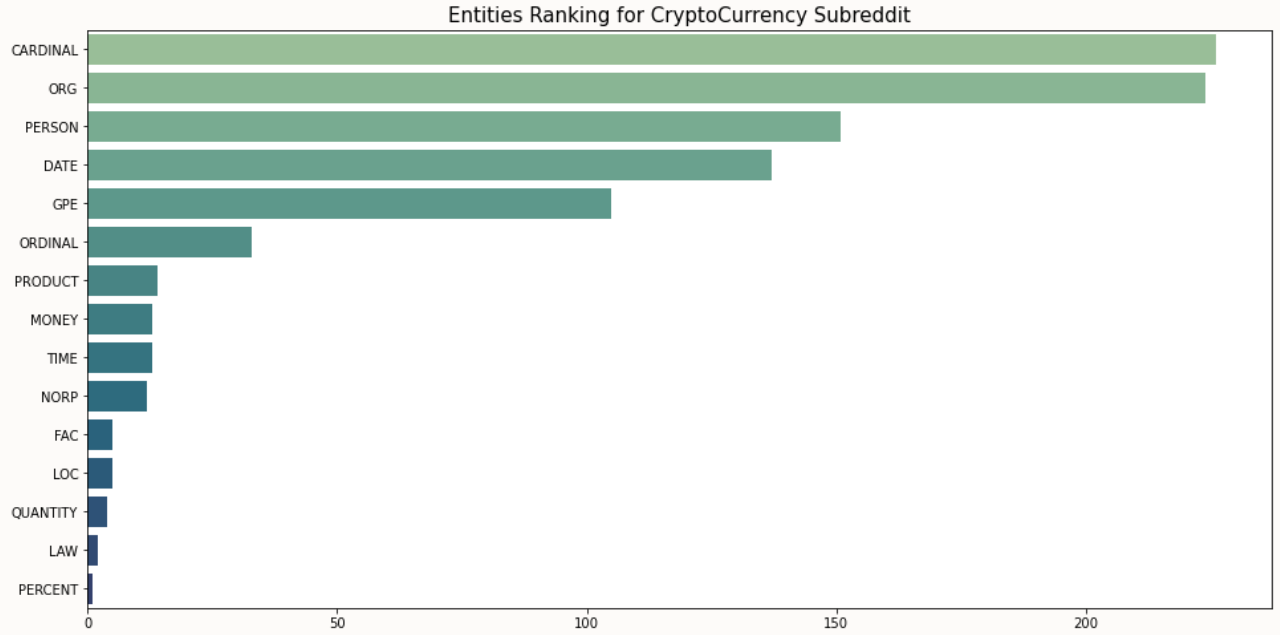
EDA: Non-Stop Words



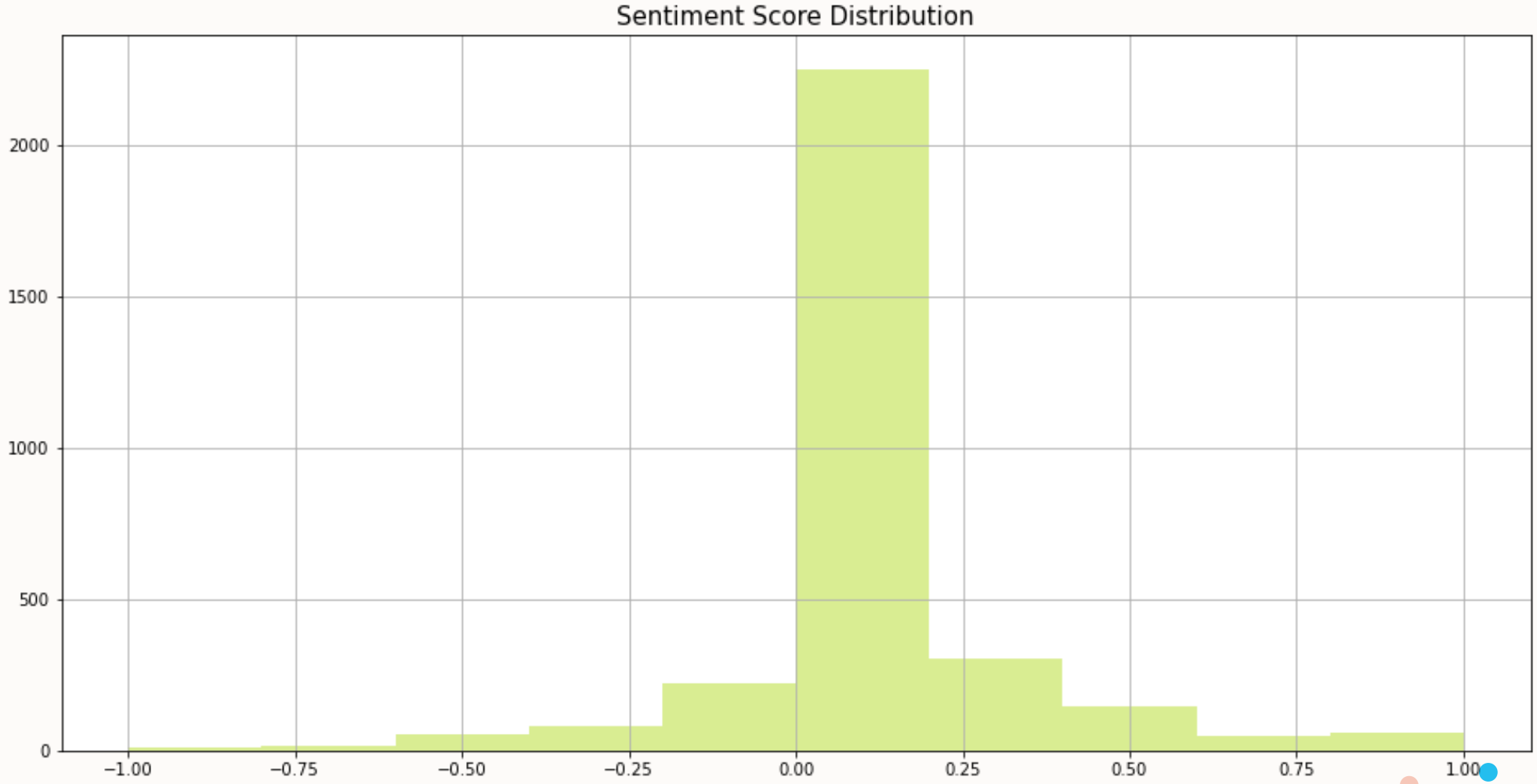


EDA: Non-Stop Words

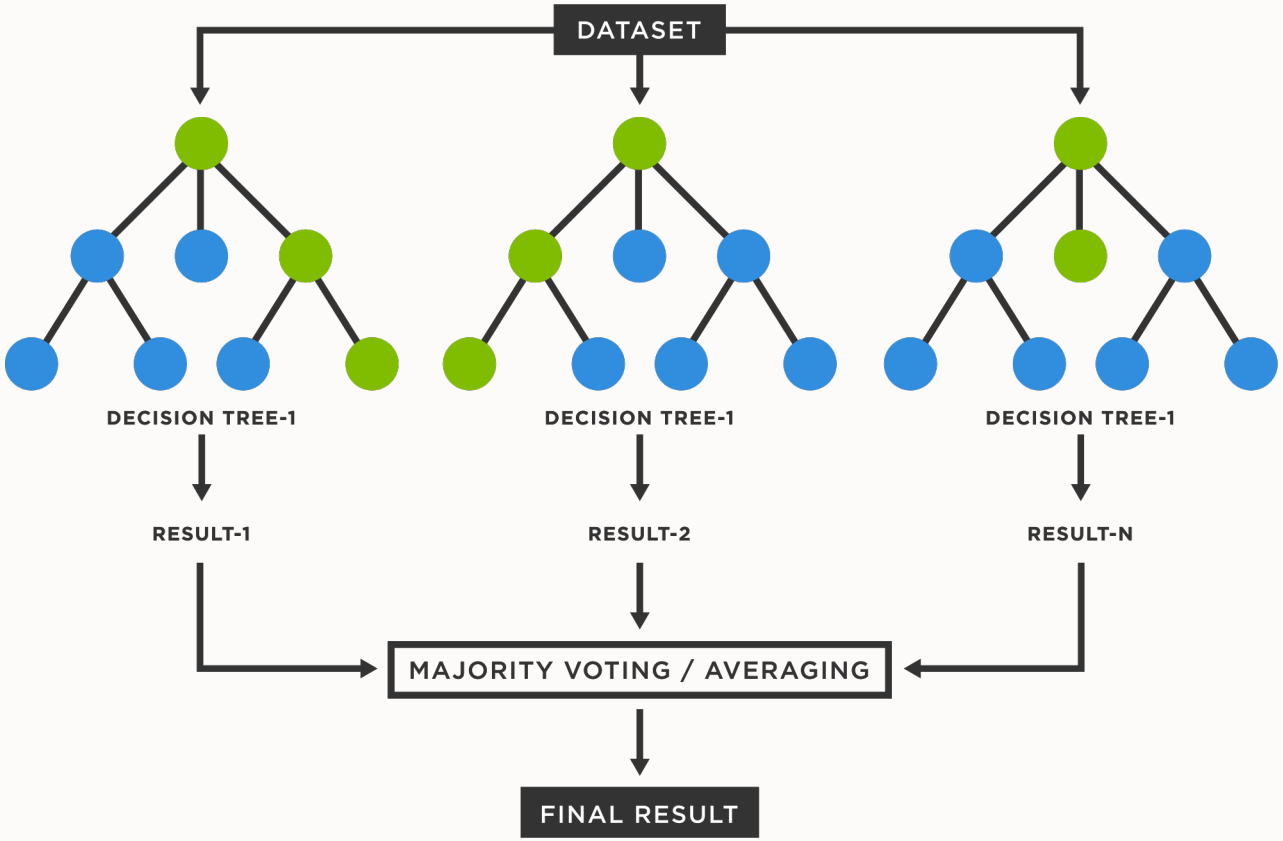




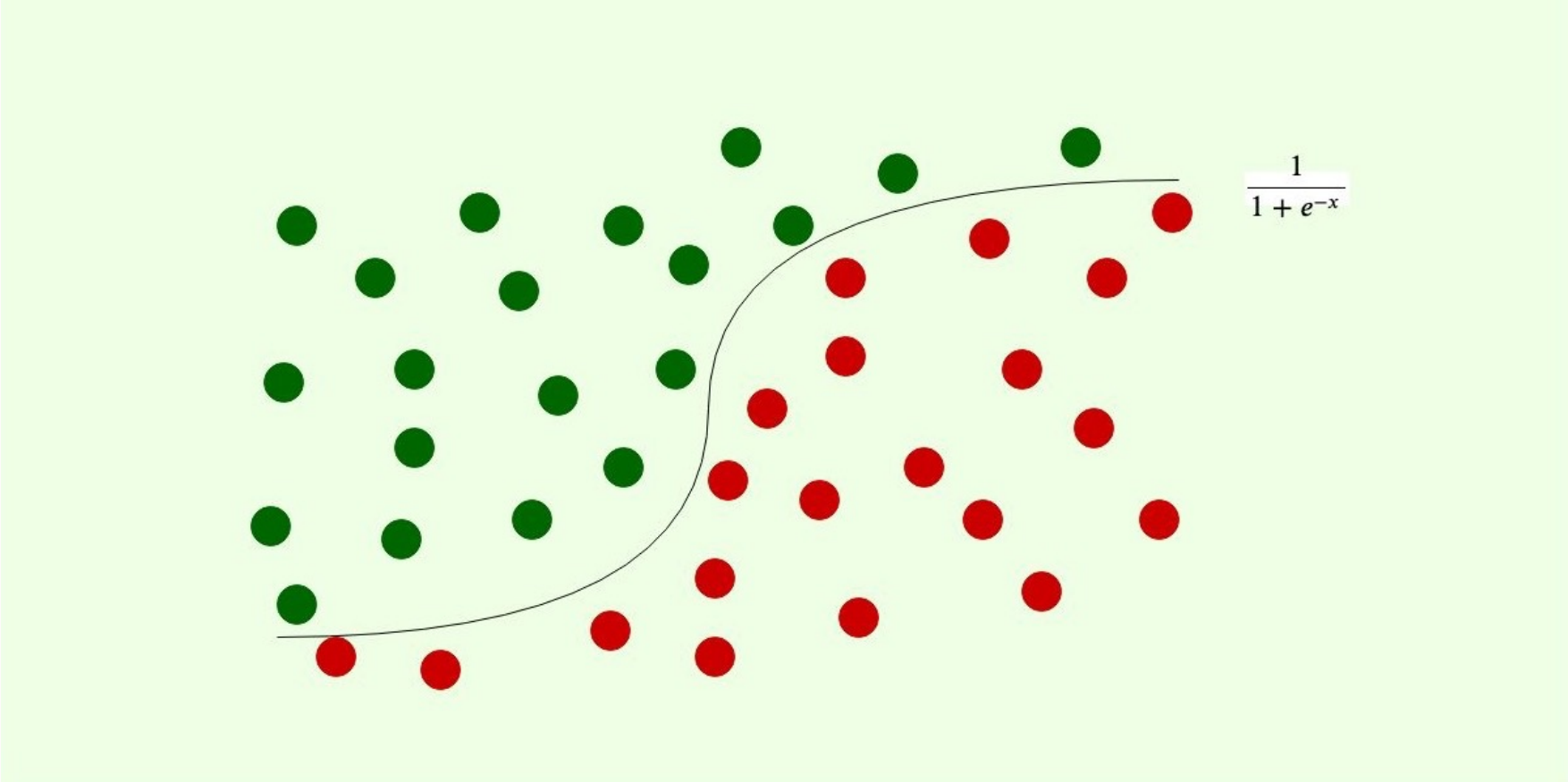
EDA: Sentiment Analysis



Modeling: Random Forest



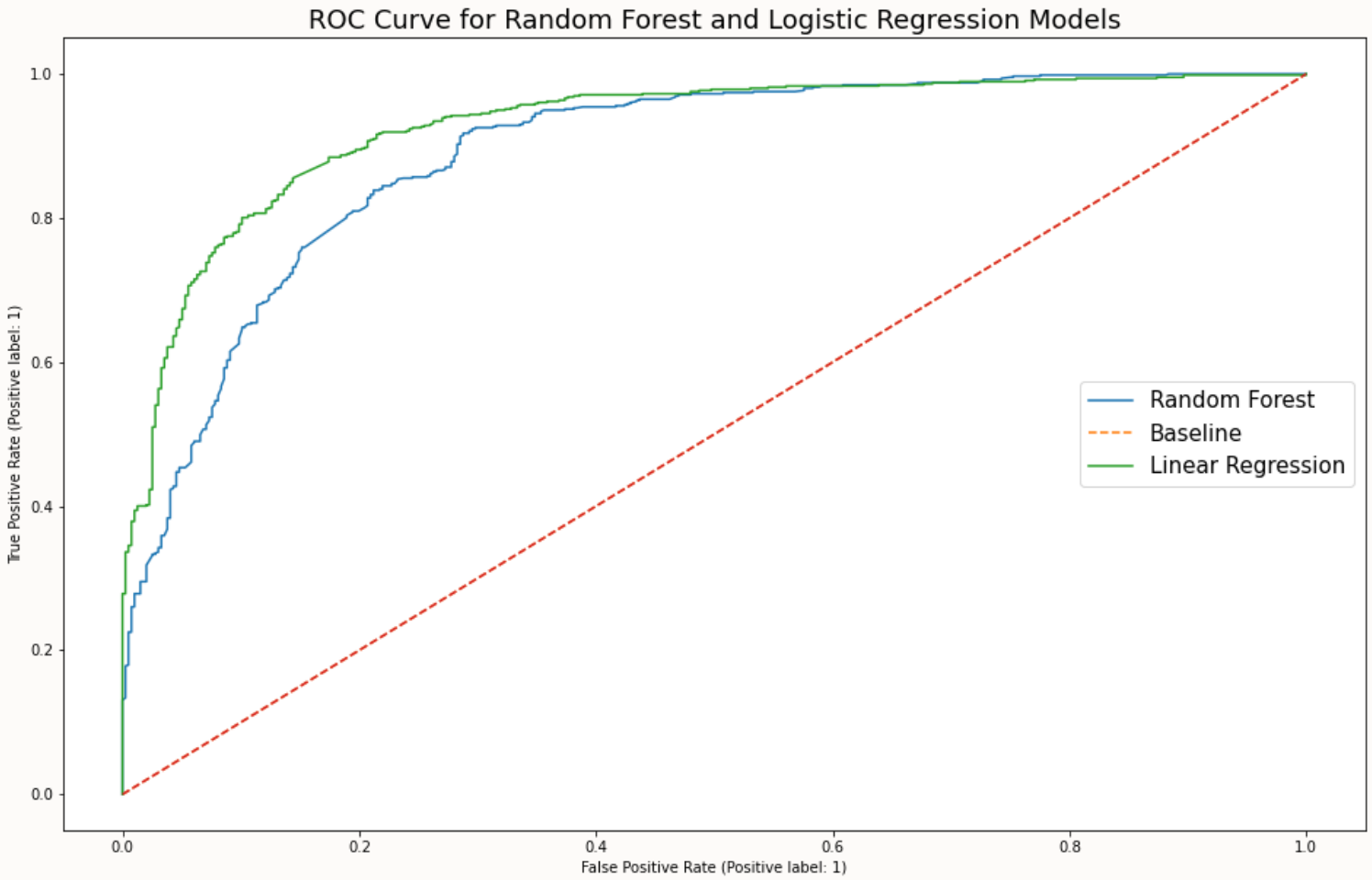
Modeling: Logistic Regression



Modeling: Evaluations.

	Random Forest	Logistic Regression
Accuracy Score	0.840	0.850
Sensitivity Score	0.920	0.945
Precision Score	0.839	0.835
Specificity	0.707	0.690

Modeling: ROC Curve



Conclusion

By Comparing to the baseline accuracy of 62.4%, Both models works better.

By comparing the accuracy score, Logistic Regression Model works better.

By comparing the sensitivity score, Logistic Regression Model works better.


Q & A