Machine Learning Approaches to Sentiment Analysis

**Background**
- Brookfield Public Securities: financial institution that invests in global alternative assets.
  - Teams are comprised of investment analysts
  - Main goal is to digest news and make buy or sell recommendations regarding a stock
- Problem: Difficult to stay on top of news and historical trends of hundreds of companies in a respective universe.
  - Analysts are very good at analyzing the current news around a company, but are limited in information retention.
- Objective: Build a sentiment analysis tool that can classify financial news articles based on polarity (positive, negative, neutral).
  - Enable high-level, macro news digest at scale

**Data**
- Labeled financial news data aggregated from two sources:
  - Financial Phrase Bank (FPB)
  - 2017 Semantic Workshop on Semantic Evaluation (SemEval)

**Results**
- Tokenization:
  - TF-IDF outperformed BoW in every model setup
  - Tokenizing at the Bigram level resulted in the lowest fold accuracy across all models, however, this may be due to the limited training corpus used.
- Classifier:
  - Linear SVC and XGBoost had highest test accuracy
  - Confusion Matrix displays difficulty in accurately predicting negative sentiment for Linear SVC

**Methods**
- **Preprocess**
  - Stop Words
  - Stemming
- **Feature Extraction**
  - BoW
  - TF-IDF
- **Modeling**
  - Ensemble ML CV/Bootstrap

**Application**
- **Overview**: Once the optimal sentiment analysis model was selected, measures were taken to integrate this model into a tool that can be used by the investment team.
- **Workflow**:
  - News articles are queried from Newscatcher API.
  - Data is cleaned, processed, and transformed in Python.
  - ML sentiment model outputs a predicted sentiment class for each article.
  - Sentiment labels and article metadata are stored in a MySQL database.
  - Results are displayed in an interactive Tableau dashboard.

**Conclusion**
- TF-IDF appears to be the optimal method to transform text data; no statistical evidence to conclude that one classifier outperformed the others.
- **Next Steps**:
  - Apply deep learning methods such as LSTM and BERT, and compare results with using word embedding and word2vec.
  - Expand analysis for aspect-based sentiment analysis for long text.
  - Analyze how stock prices fluctuate with changes in sentiment.

**EDA: Word Clouds**
- Negative Articles
- Positive Articles