# STOCK SENTIMENT PREDICTION USING NEWS HEADLINES

Benjamin Lee, Brian Seo,
Shufan Feng, Mackenzie Shen,
Meini Fan

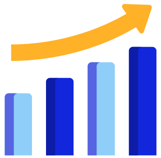
# Table of Contents

# 01
# GOALS

# What do we aim to achieve?

## ANALYSIS

Assign sentiment to news headlines



## PREDICTION

Use the sentiment data as input into a naive stock prediction of a australian ETF

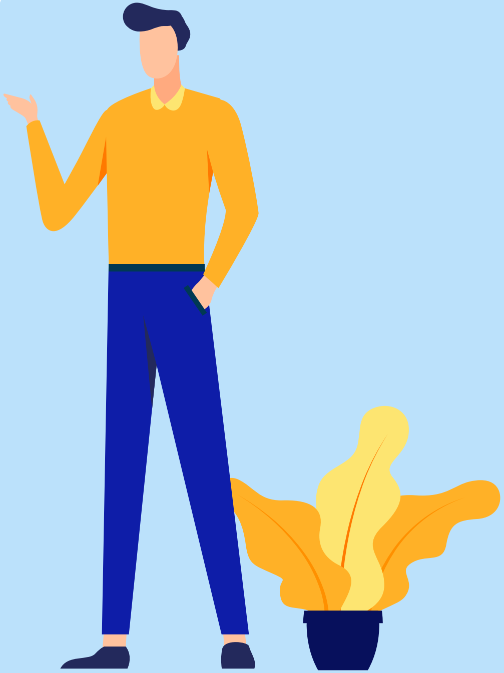

# 02
# EDA & DATA CLEANING

# Our Data

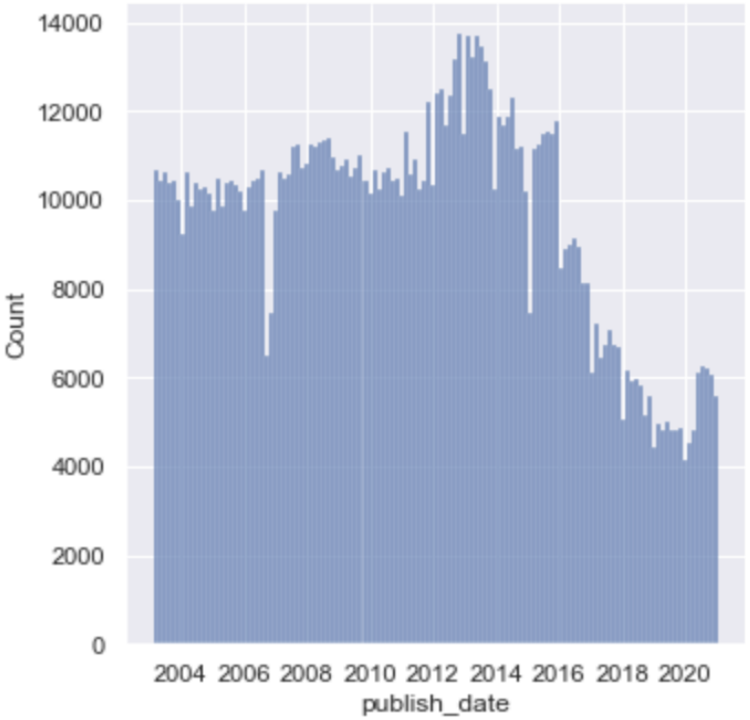
- We found this dataset from Kaggle.
- This dataset displays the headlines of news articles and their publish dates.

| | publish_date | headline_text |
|---|---|---|
| **0** | 2003-02-19 | aba decides against community broadcasting lic... |
| **1** | 2003-02-19 | act fire witnesses must be aware of defamation |
| **2** | 2003-02-19 | a g calls for infrastructure protection summit |
| **3** | 2003-02-19 | air nz staff in aust strike for pay rise |
| **4** | 2003-02-19 | air nz strike to affect australian travellers |

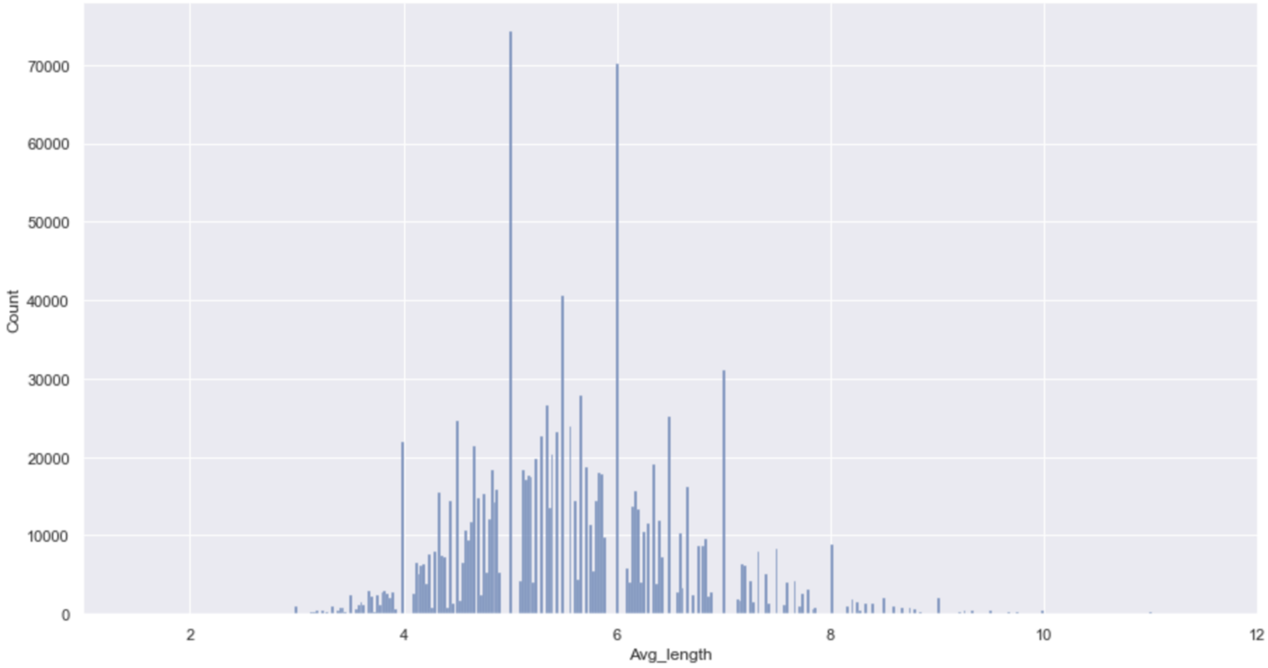
Source: https://www.kaggle.com/chandanarprasad/million-headlines-nlp-exploration/data



# Exploratory Data Analysis

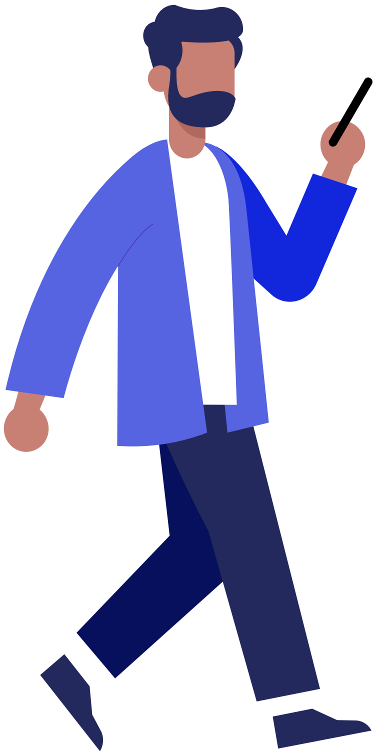



# Exploratory Data Analysis




Number of news articles according to length of headlines

# Data Cleaning



**The data is already pretty clean, so we just did light data cleaning**

**Pre-processing**

- Remove hashtags, emoji, etc

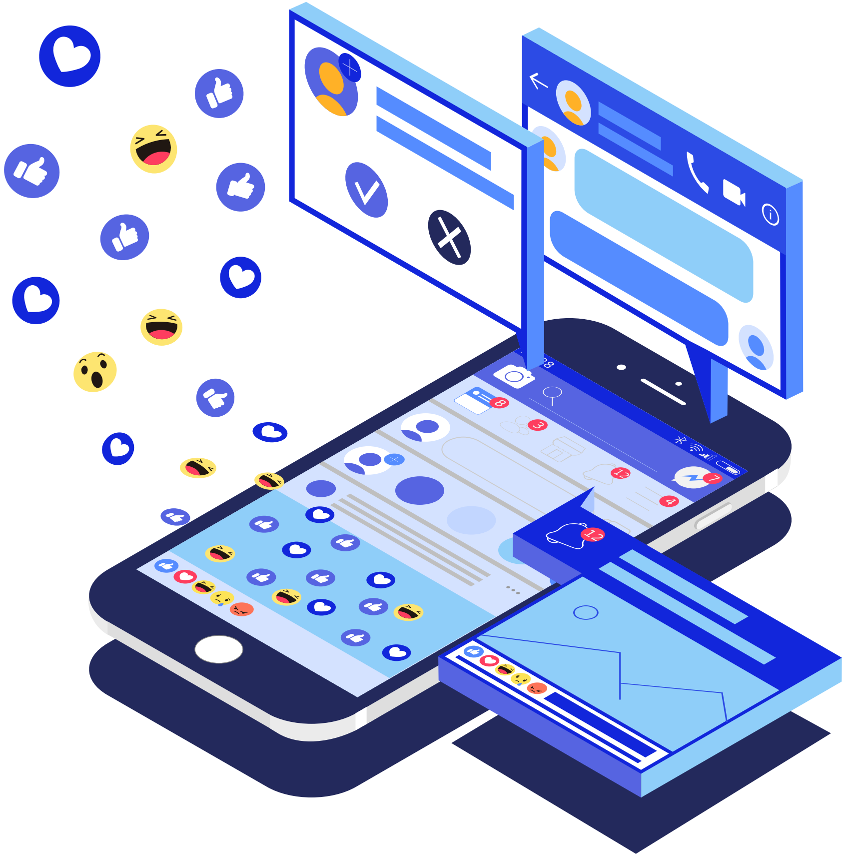
**Regex**

- Remove punctuations

# 03
# ASSUMPTION

# Assumption for Modeling Stock Price

- We will be using our daily sentiment values to aid in predictions for **an australian companies ETF**, in combination with other technical indicators a daily trader would use.
- We assume our company **makes trades daily.** We will not shift our sentiment data back a day which does make the assumption that all news come out before trading session for a day.
- To evaluate the usefulness of the daily sentiment we will use **sharpe ratio** to see if our portfolio actually benefits from the inclusion of daily sentiment in our predictive models.



# 04
# MODELING

# Modeling

## PART 1: Assign Sentiments to News Headlines

- *Two approaches*
  - **Bert**
    - Implemented Bert Sentiment Classification and generate **sentiment scores**
  - **K-means**
    - Clustered words into **two groups**, positive and negative
    - Generated **sentiment scores** of headlines

## PART 2: Predict Stock Price with Sentiment Values

- *Three steps*
  - Created **financial variables**
  - Trained baseline models using **linear regression** either with or without sentiment scores
  - Chose our final model, **Neutral Net**, with sentiment scores generated by **K-means**

# Part 1: BERT

## BERT: Bidirectional Encoder Representations from Transformers

- BERT is designed to pre-train deep **bidirectional** representations from unlabeled text by jointly conditioning on both left and right context in all layers
- We therefore used BERT to calculate sentiment scores
    - For example, we ran **Sentimental Analysis** in BERT to generate a label and score for our document as shown below.

```
classifier('cemeteries miss out on funds')
```

```
[{'label': 'NEGATIVE', 'score': 0.9995111227035522}]
```

# Part 1: K-means Clustering

## Step 1

Performed K-means to group words into two clusters

## Step 2

Assign reasonable sentiment labels to the two clusters as either positive or negative

## Step 3

Generate the sentiments of a given vector by taking the average of the word sentiment scores

## Step 4

Calculated normalized daily sentiment. Now we are ready to go for stock prediction

# Part 2: Create Financial Variables

**vwretd**

Value-Weighted Return (includes distributions)

**wretx**

Value-Weighted Return (excluding dividends)

**ewretd**

Equal-Weighted Return (includes distributions)

**ewretx**

Equal-Weighted Return (excluding dividends)

**sprtrn**

Return on S&P Composite Index

**Sharpe Ratio**

A metric to compare returns of different portfolios

# Part 2: Create Financial Variables

## Sharpe Ratio

- When possible, it is generally good to model based on the actual business metric of interest, in our case that is sharpe ratio
- To compare if our sentiment methods produce higher returns, we will use the Sharpe ratio. Sharpe ratio is a metric to compare portfolios or different strategies by seeing how much excess return was achieved over the given volatility.
- A higher Sharpe Ratio represents the portfolio will generate a higher return, while a lower Sharpe Ratio indicates a lower return

$$\text{Sharpe Ratio} = \frac{r_P - r_F}{\sigma_P}$$

Where:

rP is the return on the portfolio.

rF is the risk-free rate of return

σP is the portfolio standard deviation

# Part 2: Train Baseline Models

**Linear Regression w/o sentiment rates**

Sharpe Ratio: 1.091534

**Linear Regression w/ K-means sentiment scores**

Sharpe Ratio: 1.092180

**Linear Regression w/ BERT sentiment scores**

Sharpe Ratio: 1.085634

**After comparing Sharpe Ratios using K-means and BERT, we will be using the K-means sentiment scores for our final model because it has a higher Sharpe Ratio.**

# Part 2: Our Final Model

## Neural Net

- ***<u>Hyperparameters:</u>***
  - 'activation': 'relu'
  - 'alpha': 0.1,
  - 'hidden_layer_sizes': 3,
  - 'learning_rate': 'constant',
  - 'learning_rate_init': 0.01,
  - 'max_iter': 300,
  - 'solver': 'lbfgs'

- ***<u>Best Score (Sharpe Ratio):</u>***
  - 1.361721978248645

# 05 CONCLUSIONS

# Conclusions

- With the addition of **K-means sentiment**, our **Sharpe Ratio increased by ~0.0592%** when using Linear Regression
    - Kmeans Linear Regresion: 1.0921
    - Technical Indicators alone LR: 1.0915
    - ***<u>This shows our K-means Sentimental Analysis is useful</u>***
- We trained **Neural Net** and achieved a **1.36** Sharpe Ratio and the **Sharpe Ratio increased by ~24.75%** from the baseline Lineare Regression
- **This means that adjusted for risk we can expect 24.7% greater returns on our portfolio following the buying recomendations of our final model ove rour baseline model.**

# 06 FURTHER IMPROVEMENTS

# Further Improvements

We would like to train wider and more complex models, such as multi layer NN’s, and tree based architectures

Run all headlines through Mordecai, then predict only with data relevant to Australia

Mordecai was too slow to run all headlines through so not used for prediction, only used for additional intelligence

Functionalize things and make it scalable for streaming data

Find hourly headlines data to do intra-day trading

# THANK YOU!

# References


- https://stackoverflow.com/questions/54888490/gensim-word2vec-print-log-loss
- https://github.com/openeventdata/mordecai
- https://yahoofinance.com
- https://wrds-www.wharton.upenn.edu/
- https://www.ishares.com/us/products/239607/ishares-msci-australia-etf

# Appendix 1

## Mordecai

- Mordecai is a geospatial library that maps references in unstructured free text to ISO geographic information from.
- It extracts the place names from a piece of English-language text, resolves them to the correct place, and return their coordinates and structured geographic information