Learning summary: Financial sentiment AI

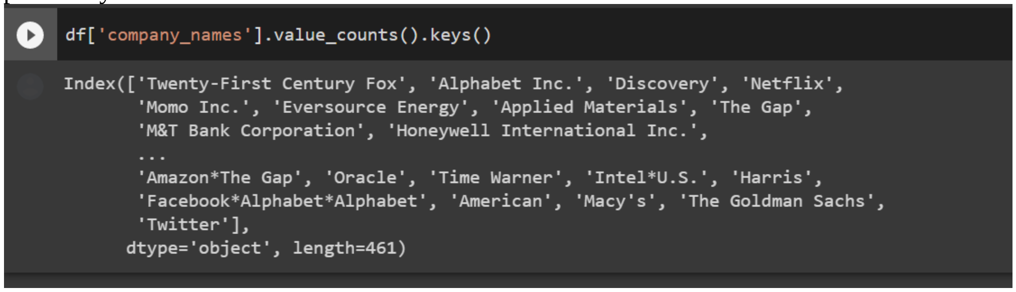
Manun Chauhaan[1*], Rhin Choe[1], and Emma Mui[2]
[1]King George V School, Hong Kong
[2] Canadian international School, Hong Kong
*Contact: chauhm1@kgv.hk

Abstract— Financial sentiment is often derived from economic conjecture and analysis. However, stocks themselves are simply a market, determined by the fundamental concepts of supply and demand. We propose a sentiment analysis based on multimodality which is based on hearsay directly related to the consumer. We accomplish this using Tweet data from selected accounts and text analytics. The step builds the foundation for many further steps which may yield a more accentuated behavioural economic model. The first of which is a generative head.

I. Introduction

Behavioural economics has emerged as a subdomain in response to the much criticised assumptions behind general economic modelling. Our approach is another solution to the problem of economic assumptions, we implement sentiment analysis on hearsay originating form tweets. We hope in this approach to bridge the gap between economic analysis and real-life individual opinions. Our approach used an extension on the Twink API and the yahoo finance API to find it's lingual and economic datasets. It then runs text analysis to label specific tweets to companies in question and run sentiment analysis. Sentiment analysis itself is run using a KMeans unsupervised clustering algorithm with number of clusters set to 2. Then any new data can be encoded into the language model ("all-MiniLM-L6-V2" from Sentence Transformers library, and then evaluated against the KMeans model.

Dataset of tweets consisted of 28000 tweets from 2018 spanning over 2 weeks, wherein 461 companies were identified as important by the t5 model

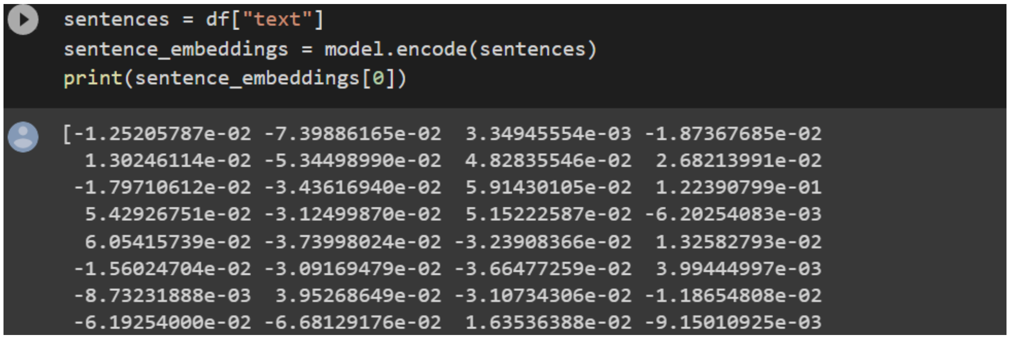


```
df['company_names'].value_counts().keys()
```

```
Index(['Twenty-First Century Fox', 'Alphabet Inc.', 'Discovery', 'Netflix',
       'Momo Inc.', 'Eversource Energy', 'Applied Materials', 'The Gap',
       'M&T Bank Corporation', 'Honeywell International Inc.',
       ...
       'Amazon*The Gap', 'Oracle', 'Time Warner', 'Intel*U.S.', 'Harris',
       'Facebook*Alphabet*Alphabet', 'American', 'Macy's', 'The Goldman Sachs',
       'Twitter'],
      dtype='object', length=461)
```

Snippet of code showing companies mentioned by name.

The tweets were used to finetune the language model and attain encoder last hidden state values in 386 dimensions as shown below



```
sentences = df["text"]
sentence_embeddings = model.encode(sentences)
print(sentence_embeddings[0])
```

```
[-1.25205787e-02 -7.39886165e-02  3.34945554e-03 -1.87367685e-02
  1.30246114e-02 -5.34498990e-02  4.82835546e-02  2.68213991e-02
 -1.79710612e-02 -3.43616940e-02  5.91430105e-02  1.22390799e-01
  5.42926751e-02 -3.12499870e-02  5.15222587e-02 -6.20254083e-03
  6.05415739e-02 -3.73998024e-02 -3.23908366e-02  1.32582793e-02
 -1.56024704e-02 -3.09169479e-02 -3.66477259e-02  3.99444997e-03
 -8.73231888e-03  3.95268649e-02 -3.10734306e-02 -1.18654808e-02
 -6.19254000e-02 -6.68129176e-02  1.63536388e-02 -9.15010925e-03
```

Finally the dataset was labelled using Yahoo finance, to match with the market state at the time.
And KMeans was implemented allowing quick sentiment analysis.
***insert evaluation matrix