Progress Report

FinBERT

Muhammad Hamza and Christian Berger

Overview

- Recap
- Methodology
- Prospective Plans

Recap

- Carried out Statistical correlation between news and KPI
- Carried out predictive analysis

Methodology

- Gathering Data
- Data processing
- Building a transformer

Gathering Data

- WRDS
- Kaggle Dataset
- Yahoo Finance
- Financial Phrasebank

Gathering Data

- For our transformer we are using:
  - Kaggle dataset to get news headlines for companies on specific dates
  - Yahoo finance: to get KPI across a number of years
  - Financial Phrasebank: labelled financial news dataset to evaluate our models performance on.

Yahoo Finance

- Yahoo Finance:
  - This will be used to pre train our custom transformer model.
  - We will split the data into pre training and fine tuning sections
  - Pre training data will be used for masked KPI prediction
  - Fine tuning data will be used in conjunction with news headlines to predict sentiment.

Financial Phrasebank

- This is the traditional dataset that most FinBert models report their accuracy on.
- We will report our final scores on this.
- One issue: It does not identify dates or stock prices so we cannot provide context for this dataset.

Custom Transformer

- Objective: Create meaningful embeddings of KPIs that provide some context to ground the news headlines.

Architectures

Vanilla

The entire input is flattened into a single vector.

MLP model with self attention layers

1x1 Convolution

The input consists of channels such that each channel represents a separate KPI

A series of 1x1 convolutions to “mix” the channel information

MLP model with self attention layers

Squeeze and Excite

Apply self attention across channels.

Use squeeze and excite blocks in between the self attention blocks

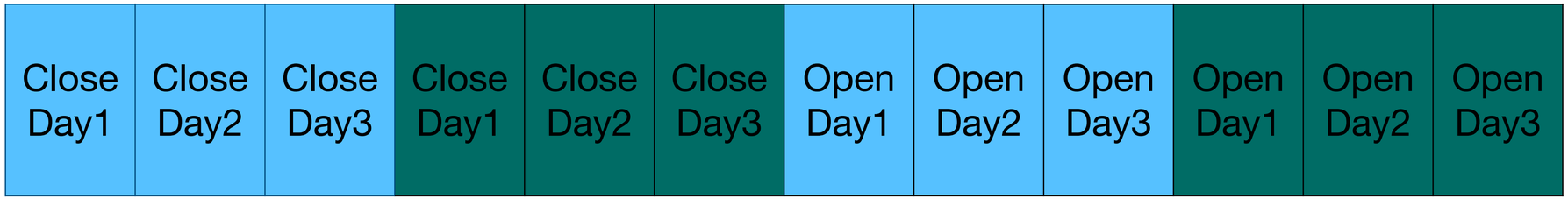
Training Data

- We have three different degrees of freedom:
  - Dates
  - Companies
  - KPI

Training Data

- Our current approach is:
  - Use a window of 5 days
  - Use four related companies.
  - Include all KPIs in a flattened vector (vanilla architecure)

Example Input



- Colours represent companies
- Randomly mask these values using a probability of 0.15

Training Scheme

- First step pre train the custom transformer to predict masked values
- The using LLM sentence embeddings with stock embeddings predict sentence sentiment.
- The labels for these sentences are derived from FinBERT.
- Finally, as a last step use scarce ground truth labels to fine tune the entire model.

Engineering Issues

- Headlines may be mostly neutral
- Company separator tokens
- KPI separator tokens