Sentiment Trading Project Summary

LSESU Data Science Society

Outline:
The thoughts of market participants often move prices and there's a wealth of information contained in the millions of tweets that are sent everyday. We will use Natural Language Processing techniques to generate trading signals from tweets, then backtest a strategy which uses these signals.

Details:
We scraped tweets using twint and obtained daily close price data for the S&P 500 (^GSPC) from yfinance.

We used nltk to clean tweets, in particular for lemmatization. Other cleaning steps included converting text to English, removing URLs, mentions, cashtags, punctuation, removing stopwords (words ≤ 3 characters long) and converting the time the tweet was sent to a datetime object. Finally, we kept only those tweets with at least one like and one retweet in order to improve tweet quality.

We used TensorFlow, sklearn and a pre-built PyTorch implementation for modelling.
TensorFlow allowed us to use a BERT model which had been pre-trained on a large corpus of text to generate 'tweet embeddings' for each of our tweets, i.e. to convert each tweet to 512-dimensional vector. Then we fit a k-means model to these tweet embeddings and manually defined the labels after inspecting a sample.
Sklearn provided the tools to apply the SVM algorithm. Here we needed labelled data, so we manually labelled ~200 tweets as bullish, neutral or bearish and applied the algorithm to this dataset.
ULM-FiT, like BERT, was pre-trained on a large corpus, but then fine-tuned on the publicly-available Financial PhraseBank dataset to improve its performance on finance-related text.
The pre-built PyTorch implementation was FinBERT, which is essentially the same as the ULM-FiT model but using a different initial pre-trained model with different word embeddings.

Once we had labelled each tweet with sentiment, we grouped by day and assigned a sentiment score for each day, by summing over the sentiment value (+1 for bullish, 0 for neutral and -1 for bearish). We then backtested a strategy which bought when the sentiment score was > 5 and sold when the sentiment score was < -5. Whenever the signal changed, the existing position was liquidated and an opposite position was entered.