**Data types.**

Arguably the most important part of any ML model is what data to include and how it should be represented.

Developing informative features from raw unstructured data is a crucial step before training a ML model.

Financial data is generally divided into four essential types: fundamental data, market data, analytics, and alternative data.

Fundamental data encompasses data that can be sampled from regulatory filings, e.g., assets, liabilities, earnings, etc. It is easily accessible, and therefore one would suspect most indicators in fundamental data are already discovered and exploited.

It is important to keep in mind that some fundamental data is never published at the end of the reporting period. Quite often the data is backfilled or reinstated. Therefore, we need to know when the data was made available for the backtest to accurately reflect the current market conditions.

Market data consists of data related to trading activity from exchanges, e.g., price, implied volatility, volume, etc.

Market participants might leave a “footprint” in market data that another trader can use to predict their related hedging activity and front-run them.

Analytics is processed fundamental, market, and/or alternative data, e.g., analyst recommendations or earnings estimates.

Alternative data is the data that falls outside these other three types e.g., satellite/CCTV images or social-media sentiment.

**Sentiment analysis**.

Sentiment analysis is the field of study that analyzes people’s opinions, sentiments, attitudes, and emotions towards entities such as products, events, and services \cite{liu2012sentiment}.

Human decisions are influenced by the opinions of others, whether it is the products we by, the entertainment we consume, or the pollical opinions we hold.

The digital age has provided unprecedented access to opinionated data.

With the rapid growth of social media in the last decade, sentiment analysis has become an increasingly active research area. Additionally, it has extensive commercial applications.

Sentiment analysis uses natural language processing (NLP) to extract information from text.

Basic tasks include classifying the polarity of text, i.e., whether it is positive, neutral, or negative. Specific features of an entity can also be classified.

The degree of one or more emotions can be classified using one-hot-encoding.

With the recent advancements in AI, the sophistication of sentiment analysis has increased.

The development of advanced deep language models has allowed the analysis of more difficult data domains.

**Applications in trading.**

Markets are influenced by many factors such as geopolitics, economic policy, market psychology, and supply and demand.

Consumer attitudes toward a public company, for example, should affect the share price of that company. Similarly, the price of crude oil should be affected by a conflict in the middle east.

Digital media can be used to obtain qualitative information of this kind.

With recent advancements in AI and vast amount of data available, it is possible to analyze digital media time series and produce a market sentiment based on the content.

The only limit to discovering these relationships is creativity.

The massive amount of data humans generate on a variety of topics can be leverage into a powerful market forecasting model using machine learning.

A bit of research has been conducted on this topic in the past couple decades.

In 2010, Zhang and Skiena \cite{zhang2010trading} developed a market-neutral trading strategy based on sentiment analysis on NYSE stocks.

**Microblogging.**

Microblogging data has the potential informative value to forecasting prices of financial markets.

Dataminr is an API for real-time event and risk detection. It uses the firehose API from Twitter and AI technology to filter relevant information.

…