Data set:

A dataset is a group of data with similar characteristics, also known as a collection of examples, that is used to train and test algorithms. The single row of data in the dataset is referred to as an instance.

To conduct various actions, the model must first comprehend them, which may be accomplished by providing data to the algorithms. Before providing the data, these datasets must first be separated into two categories: training data and test data. Models are trained using training data, while model performance is evaluated using test data. On the other hand, just a few research included validation data, which is used to validate model after it has been trained. The most prevalent test splits were 80:20 or 60:20:20, which were employed in the majority of research.

To ensure the model's interpretation of data, the training dataset must be passed into the algorithm first, followed by the validation dataset/testing dataset. The speed with which a model learns and develops is determined by the quantity of data supplied to it; the more data provided, the faster the model learns and improves.

Data types:

Market data:

Market data is a word that refers to financial information that is required for research, analysis, and trading. Market data may also be utilised to obtain past prices, they are an important aspect of technical analysis and can help you plan your future trades. It includes data such as the opening and closing prices, high and low prices, trade volumes, and so on.

Text data: social media 🡪 twitter data

Information which are collected from social media texts , news, and online searches, is referred to as text data. Using this text data, sentiment analysis can be performed and the generated sentiment factor (e.g., positive, neural, or negative) can be used by the model for prediction. This datatype is not as usual as market data, because this datatype give more important information which cannot be found in market data.

Trader’s information:

Market data that is commonly used does not give any additional information that might aid in increasing the model's performance. Investor's trading data contains more valuable information, such as the number of transactions, purchasing volume, and selling volume on the target stock, etc. The accuracy of the prediction model can be improved by combining investor trading information with market data.

The Few raw data are difficult to get and considerably more difficult to process. Each data set has its own set of challenges. A huge quantity of data is required to successfully train a complex model. Most studies utilise market data as it gives a huge volume of data when compared to other data sources.

Data Length:

The data length required is determined by the problem's complexity and the algorithm used. Deep learning algorithms that are flexible and high-variance modify their predictions dependent on the input data they are given for training. As a result, its training will need a large amount of data. When it comes to data duration, a short period of time may not be enough to provide useful findings, and there is a considerable chance of overfitting. Apart from that, data availability and cost are the most important aspects to consider when determining the data length.

Clustering:

When the number of related stocks is too small, the useful information obtained by the model is not enough. When there are too many related stocks, more noise will be introduced and the noise will affect the judgment of the model. C represents the number of user clusters. When C is small, the differentiation between clusters is small, and their trading patterns are relatively similar, so there are less differences can be found by our model. When C is too large, some clusters are not representative. The transaction behavior of these users may be special cases, and more outlying data may be introduced