Abstract

A pipeline is explored in acquiring press statements from various companies’ websites. Then classified into multiple classes using various classification algorithms. Baseline methods such as naïve bayes perform sufficiently but are improved by deep learning techniques.

1.0 Introduction (Quantitative finance, event study)

Corporate press statements serve a vital function as the primary means for companies to provide current and future investors with the information they need, enabling them to make informed decisions. Press statements can range from future acquisitions, new products or services, and the most common quarterly reports. The only mandated type for public companies is the quarterly reports as they provide key updates on the financial state of the company. Although not mandatory, most companies release statements of many types, to increase transparency and allow the market to operate efficiently.

With time in the market, it quickly becomes apparent that press statements do have a significant effect on stock prices. Volatility groups around statements, especially quarterly reports that the market has not priced in [1]. Therefore, understanding the relationship between press statements and share price is beneficial for investors and financial analysts. As well as having applications in quantitative finance by including the use of such relationships in financial models. With a large number of public companies listed around the world the volume of press statements released is far beyond manual classifications. The development of an accurate classification system needs to look through each statement.

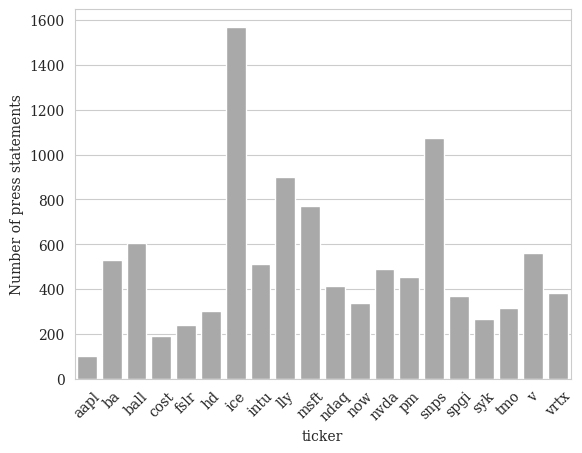
Event-based trading models have been shown to post attractive returns, which has only increased interest in applying natural language processing to corporate statements [2]. In finance there is much debate on the viability of these models. With the efficient market hypothesis stating that current security price reflects all the information available.[3]

2.0 Data collection (How? Choices made? Web scraping?)

A novel dataset is aggregated through the process of web scrapping. Press statements are extracted from a company’s investor relations section where they release the most important statements. And where possible the companies press statements are to obtain only the financial or investor-related statements. This allows the number of statements to be limited with the most important statements kept in.

Web scraping is done using python with the use of the requests library which extracts the html from the website. With parsing done by the beautifulsoup library. More modern websites generate html with the use of JavaScript code. Therefore, selenium is used to load a web driver that can run the code to generate the html code. For each statement the date, URL to full statement, description, title is extracted then stored in a dictionary. These are generally found using id tags within the html code. The description is found in the head of html, unfortunately not every website does this. For some the first paragraph is taken, but attempts have been made to avoid this different press statements can have different layouts. If this happened an error would occur.

Using the yfinance and the dates obtained from the press statements the change in share price can be obtained. One issue faced with obtaining press statement is the date given on the statement likely does not contain a time. As it isn’t known if the press statement is released premarket or post market the resultant price change could be on the date or the day after. To avoid this the change is as from the open price on the press date to the closing price of the next day.



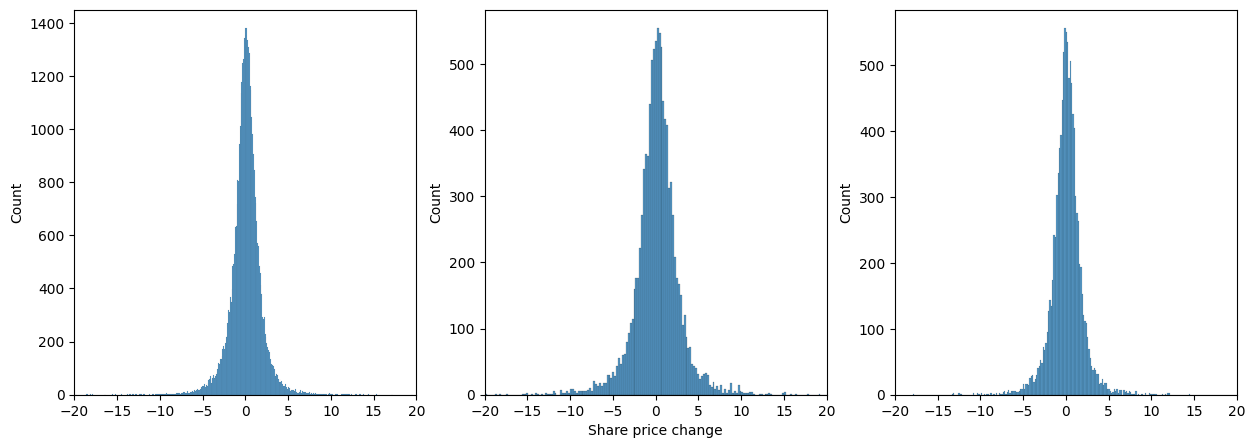
3.0 Hypothesis testing

Despite press statements being fundamental to the financial state of the company. From viewing the share price change over time, it is not immediately clear which days the press statements are released. The share price can also undergo significant volatility without press statements. This is of course the result of numerous other compounding factors within the market. Figure 1 gives an idea of how volatile a company can be without press statements. It may be the case that the impact is no greater than noise within the market. As we two distributions are obtained the first being share price changes on days with press statements and changes without press statement. Statistical analysis then can be performed to find which has more impact, background effects or the press statements.

A graph showing the growth of a stock market

Description automatically generated

*Figure \_\_\_ - Candlestick bar chart of Eli Lilly and Company over a 6-month period. Blue dots represent the release of a press statement.*

However, due to the afore-mentioned problem of not knowing the time of release, share prices change the days after are removed from the days without distribution. Histograms of the data are plotted in figure 1 visually there seems to be differences within the tails of the data. Especially when looking at the resampled comparison (c).

*Figure \_\_\_ - cvbcvgbdcfgdfgdfgdfgdfgdf*

To statistically confirm that the resulting distributions are different two tests are performed. To do this the following null hypothesis can be proposed, *there is not a difference between share price changes on press days and non-press days.* Therefore, the alternative hypothesis is, *press statements have significant effect on shares price compared to non-press days.*

The first test performed is the student’s t-test. This test compares the means of the two groups to identify if the differences are the result of randomness. Also, the similar Welch t-test is performed which assumes that the variance is unequal. This test should be more accurate as visually the variance is different as well as the standard deviation is significantly different. Both t-tests can be said to support the alternative hypothesis as the value is above zero with low values. No significance level was defined but both p-values are below common choices such as 0.05 or 0.01. To provide further evidence each group can be bootstrapped then added to a list. After one thousand samples the list of p-values can be used to calculate a 95th percentile confidence interval. The interval also supports the alternate hypothesis. All calculations are shown in tables 1-3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistic | Value | P-value | Confidence interval LB | Confidence interval UB |
| Mean | 0.167 | N/A | 0.114 | 0.218 |
| Standard deviation | 2.73 | N/A | -0.026 | 0.026 |

*Table 1 – Statistics for press days. For the mean confidence interval, a t distribution is assumed. With 95th percentile confidence intervals for mean and standard deviation.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistic | Value | P-value | Confidence interval LB | Confidence interval UB |
| Mean | 0.0757 | N/A | 0.0576 | 0.0936 |
| Standard deviation | 1.89 | N/A | -0.0180 | 0.0180 |

*Table 2 – Statistics for non-press days.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistic | Value | P-value | Confidence interval LB | Confidence interval UB |
| t-test | 4.00 | 6.25e-05 | 7.81e-11 | 1.03e-01 |
| Welch-test | 3.22 | 1.27e-03 | 1.58e-07 | 1.94e-01 |
| Mann Whitney u | 228148156 | 8.85e-08 | 1.18e-13 | 2.09e-03 |

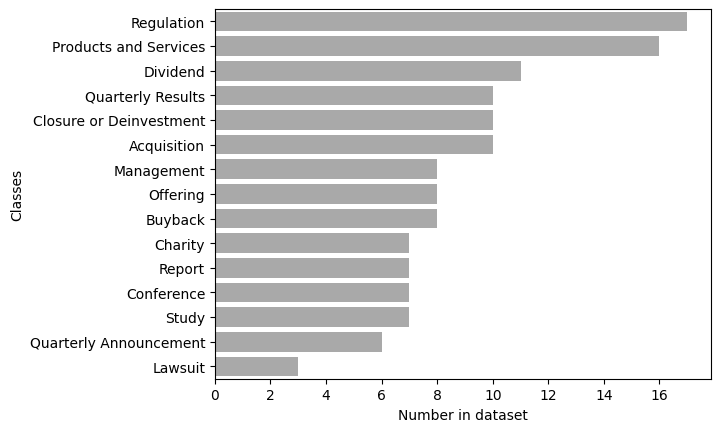
*Table 3 – Statistical tests calculated on the two distributions. 95th Confidence interval for the p-values.*

4.0 Classification of press statements (which class to choose, step classification, NB, Bert)

It can be concluded that press statements do have a significant impact on the share prices. But determining which press statements have the most impact requires classification. As there are no clearly defined rules in the type of press release a company can disclosure choosing the classes becomes difficult. This is mainly as a result of the supervised learning algorithm, as the need for a test and train set requires manual classification. If there are not press statements of a certain causing a class imbalance in induced significantly effecting performance. Classes need to choose regarding the dataset by not having too many for specific cases. It is therefore helpful to group up certain types of. The following classes are used.

**Acquisition, Buyback, Charity, Closure or Disinvestment, Conference, Dividend, Lawsuit, Management, Offering, Products and Services, Quarterly Announcement, Quarterly Results, Regulation, Report, Study**

To then produce the training set \_\_\_ press statements are manually classed by observation. A much larger training set is limited by the time it takes to manually do this the need for a class imbalance. Many classifications algorithm suffer from a class imbalance as they often default to just predicting the most common class.



4.1 Naïve Bayes

The first machine learning method used is naïve bayes this approach is a supervised technique and a popular choice as a baseline algorithm. To train Naïve Bayes, Bayes theorem is employed (equation1) to train and calculate posterior probabilities given occurrences of certain features within the dataset. One assumption with Naïve Bayes is that each class is independent of each other. As a result of this the number of calculations is reduced drastically making the method extremely fast to train.

*Equation 1 – Bayes theorem where A is the class and B is the features within in the press statement.*

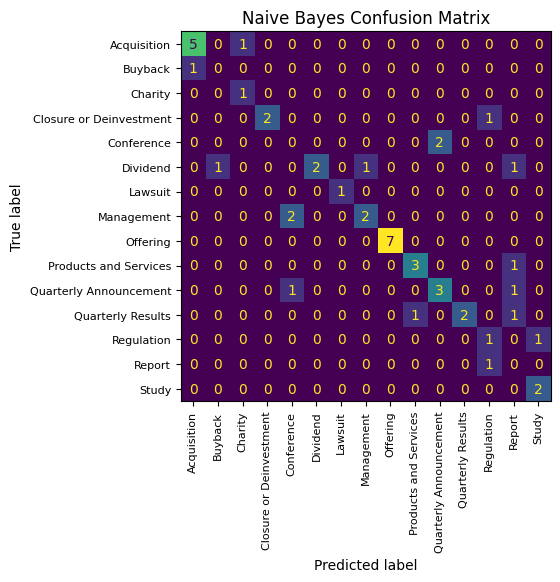
Before using naïve bayes the data must be pre-processed. As the first step stop words are removed these are generally words that are inconsequential to the context of the statements such as it, but, in etc. This will allow the algorithm to focus on words that matter more to each class. NLTK library provides a set of stop words to be used within the code.

Now the press statements need to be vectorized, a simple method is the count vectorizer. A vocabulary of words is built from the dataset containing every instance of every word. Each statement is then separated to words and count, the vectorized statements forms and array with an index corresponding to each word in the vocabulary. The number of occurrences of each word is stored at each index.

To train on the data the probabilities how often is class appear within the dataset (prior probabilities). Then the probabilities of each feature (words) given each class is calculated. Now given a new statement the posterior probabilities can be calculated for each class. The largest is then chosen as the class for that statement after, by converting the index to corresponding class.

Pros and cons

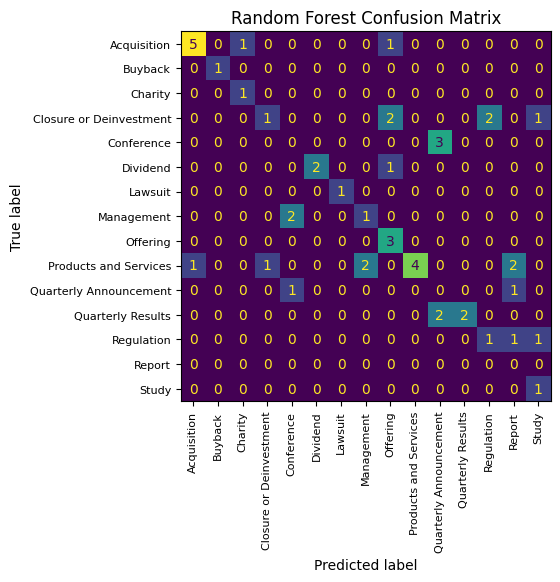
Results



4.2 Decision Tree/Random Forest

Decision trees are another simple alternative to Naïve Bayes. The data is vectorised in the same way using the count vectorizer. In a decision tree a hierarchical structure of nodes is created to split up the input data in way that identifies patterns. At the end of the tree there are leaf nodes that represent each class. The optimal splits of each node are learnt by minimizing an impurity measures such as entropy. Before training a stopping criterion such as maximum depth is used. A trained tree function similar to a large number of layered if statements. Decisions are made by traversing nodes within the tree if conditions are met for example when a specific word is detected.

A random forest an ensembel method of multiple decision trees trained on the data. Each decision tree is trained on a subset of the data by using the bagging method. Trees are devloped independent of each other. The input is then passsed into all the trees at once, after which the ouputs are then aggregated together.



4.3 RNN/LSTM

4.4 BERT (Transfomer)

BERT is a state-of-the-art language model developed by google in 2018, that excels various NLP tasks including text classification. The model was trained on 2.5 billion words contained in Wikipedia articles and 800 million words in Google’s book corpus dataset. Using the dataset BERT learnt a fixed vocabulary containing a total of \_\_\_ words and sub words in the English base version.

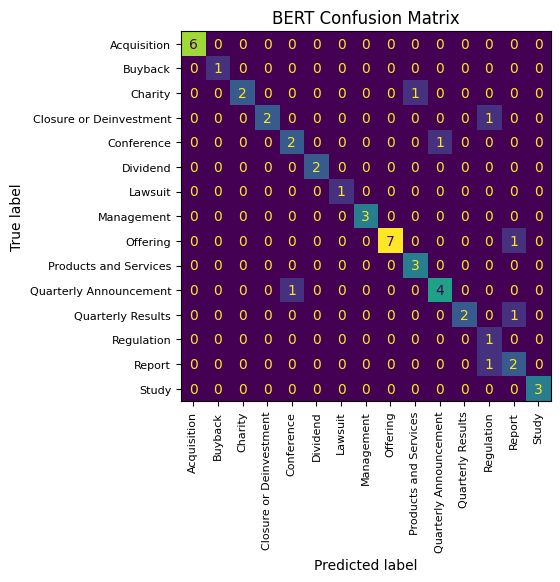
When an input is entered into BERT, the text is first tokenized. This involves first separating the sentence up into individual words. The tokens can then be looked up in BERT vocabulary, if the word does not appear in the vocabulary it is split up into sub words until each sub word is found in the vocabulary. In some cases, even reducing to single characters if needed. After tokenization each token can be looked it fetch its embedding. Where the embeddings are vectors that serve as numerical representation to words and subwords in the vocabulary. This is a process known as word-piece embedding. This process makes up the first layer of BERT. Following this layer there are 12 transformer layers.

The transformer layer revolutionised natural langauge processing by fixing many of the issures faces by previous neural network appoaches. The main reason for this is the self attention mechanism for learning weights. In BERT and others each word has three vectors that are associated with it query, key and value vector. The value vector carries the information for each word. Query vector is the word currently being looked at and the key vector is represents all other words in the sequence. Each one of goes through a linear transfromation that is trained representing a set of weight matrices. The dot product is taken between the query vector and the key vector. The result of this after appling a softmax is the attention scores which represent the similarty between the two vectors. Value vectors are now multplied element-wise by attention scores. The scores act as weights for the value vectors, now each value vector can be summed up to obtain the final representation. This final representation can be converted into a class using a pooling strategy then, a dense layer which output represents the probabilities for each class. Of the probabilities the argmax can be chosen which will then map to a specific class.

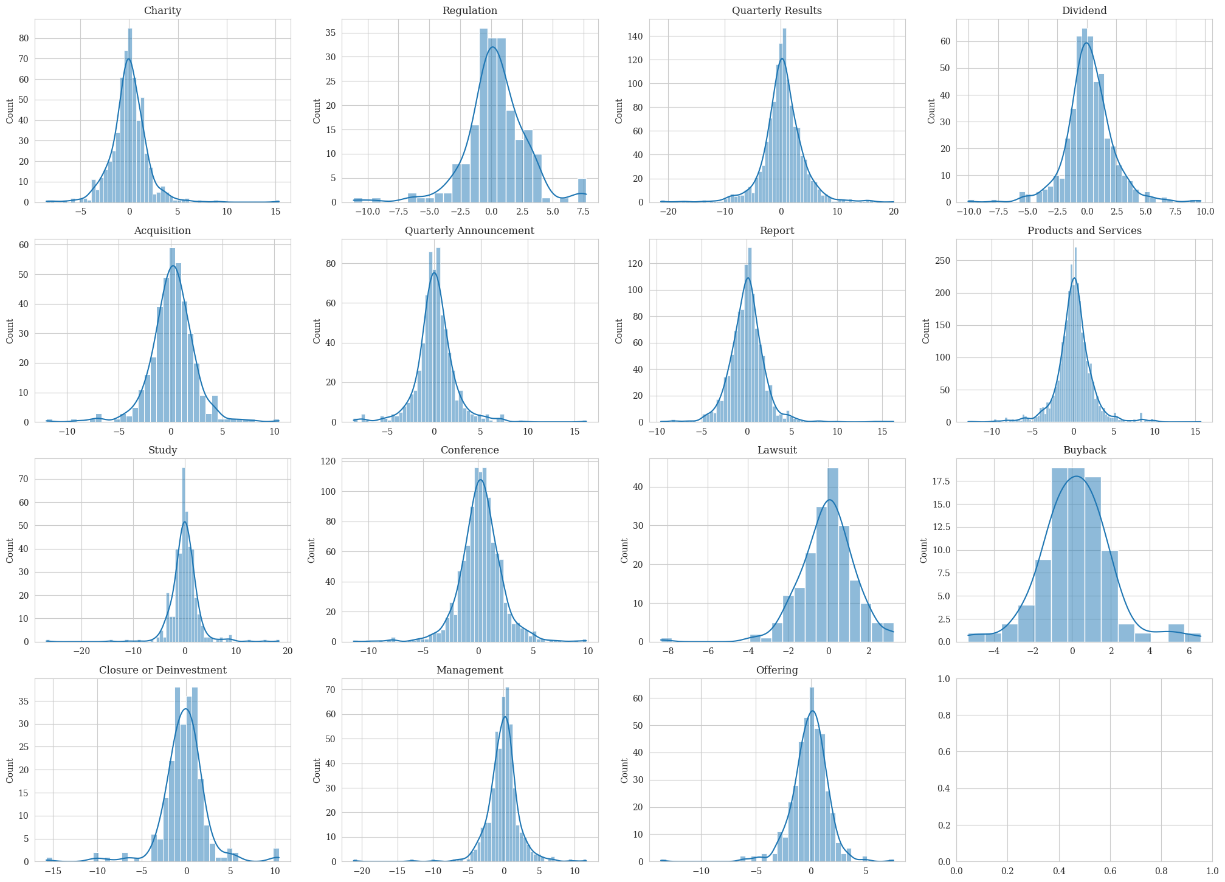
In BERT and others, it is common to use multi-head attention which extends the usage of self-attention. In multi-head attention multiple instances of self-attention are used in multiple heads. Each head is computed in parallel to allow the calculation of alternative outputs. By doing this the mechanism explores more than one context for each word then finds the most probable. This helps with words that change meaning with the presence of another.

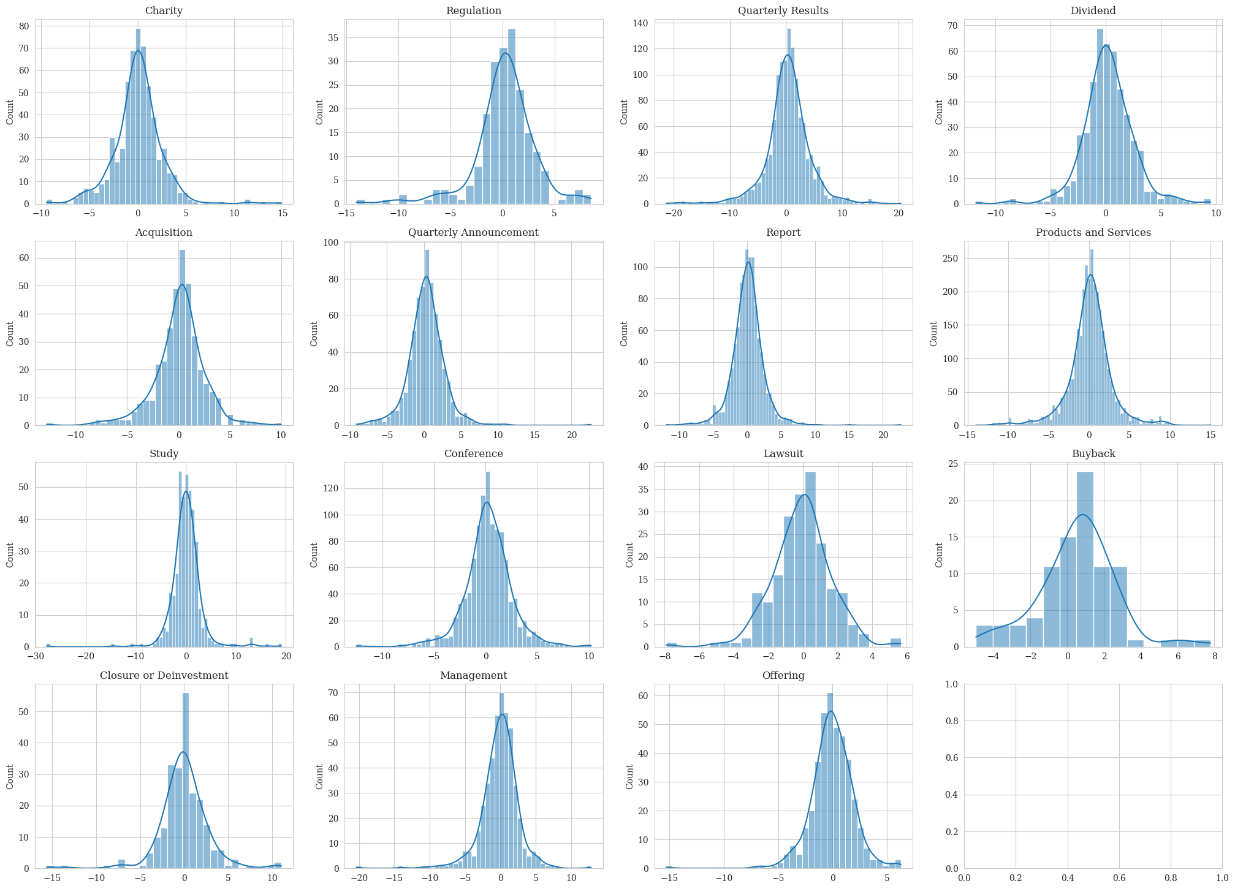
Results

As BERT is already pre-trained the model needs to be only fine-tuned on the data. To do this the model weights in the are updated with the backpropagation algorithm. Categorical cross entropy loss is used to evaluate which statements are the best predictor of each class. The parameters set for the model are, learning rate=7e-5, batch size=4, epochs=15. The optimizer is set as adamw with parameters set to be optimized for transformers by the huggingface library.



Share price change analysis impact of classifications, impact of sector





Regression with Bert (Not going to work)

Conclusion

References

[1] <https://www.researchgate.net/profile/Edwin-Elton/publication/227445172_Expectations_and_Share_Prices/links/555490ce08ae6943a86fd930/Expectations-and-Share-Prices.pdf>

[2] Fama emh

Bert <https://arxiv.org/pdf/1810.04805.pdf>

Self-attention <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>

Multi-Head Attention https://arxiv.org/pdf/2006.16362.pdf