FinBert Framework For Market Moving Twitter News Detection

Adrian Bialonczyk Hasan Khan Jason Kim

Abstract

We investigate the ability of natural language models to detect market moving tweets. Twitter is a leading source of information for both general and investment specific news. Academia and industry have studied frameworks for text analysis tasks, such as sentiment classification, on this data type as well as broad news publications. Many of these studies identify meaningful ways to categorize data as positive or negative, but fail to prove which data is in fact market moving in a short time interval. We propose a framework of 8 key financial events in a list of 410,000 Twitter users to construct potentially market moving information. We then detail a state of the art FinBert natural language model that has achieved high performance in related financial tasks. This approach may be meaningful in predicting which tweets are market moving, and can potentially be tested on additional text data sources, such as news, in the future. We further believe this framework may generate meaningful results for other asset classes and trading frequencies.

Keywords: Text Analysis, FinBert, Algorithmic Trading

2022

1. Introduction

We are researching and developing a text analysis framework that can be implemented on twitter to potentially identify market moving events in the next minute. The framework may be transferable to multiple time frequencies and other text sources, such as news, in the future. In this research proposal we focus specifically on twitter data with the market impact of T+1 minute, as we believe this is the most challenging, but also least noisy timeframe. The key elements to this research include identifying financial events with market relevance, determining features correlated to future returns, and architecturing a text model that can make sense of noisy data. The data pipeline for extracting this data is important given the short time frame of our research, though it is outside the scope of this proposal. Further, we focus on events specific to US equity markets which according to literature should have both a short and mid-term impact on future prices. This document first provides a review of existing literature regarding Twitter's impact on stock prices, categorizing financial events, and state of the art language models. We then explain the data to be used for the research, introduce a dataframe for the features we will test, and explain important considerations when preparing for modeling this problem. Lastly we detail the architecture and implementation of a state of the art FinBert model that is expected to generate a strong understanding of the text that has the greatest impact on future equity returns.

2. Literature Review

Twitter is a news source for disseminating short snippets of information to the public, with over 290.5 million monthly active users globally, including many of the world's most influential individuals. Company CEOs, Leading Journalists, and News Agencies are sometimes the more active users - such as Elon Musk, Scott Wapner, and the Wall St Journal's official account. Given its character limits and product scope it is often easier to disseminate information via twitter than formal press releases or news articles. Bloomberg (who provides raw tweets via their terminal) found that Twitter is a faster source of breaking financial news than other news providers [1]. Their research found Tweets were published approximately 2 minutes faster than news headlines in many cases - an eternity in the world of trading. For example, on February 28, 2019 at 12:57:06.107 CNBC's Scott Wapner tweeted 'SCOOP: Bill Ackman's Pershing Square has been building a stake in $UTX. Details coming at top of @halftimeReport'. Figure 1 shows the stock price increasing $2 after the tweet, with the short term price fully reflected by the time the first news headline published the story 2 minutes and 6 seconds later.

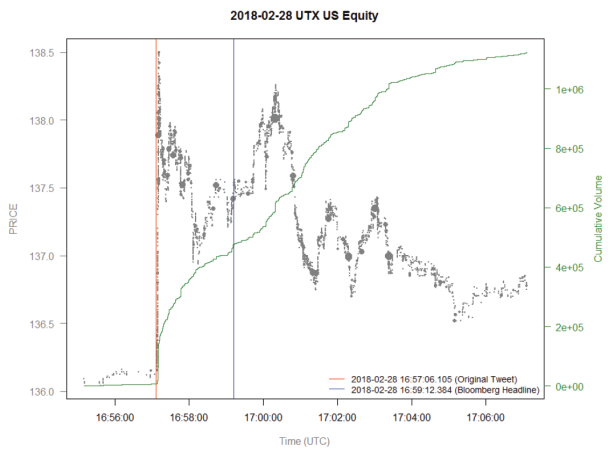


Figure 1. Time of Tweet represented by Red Line and first news release in Blue. Source: Bloomberg, 2018.

Though literature shows that volume and long term price as impacted by news, this research focuses on the advantage of Twitter events being published before these other sources.

There are approximately 500 million tweets published each day. It is important for us to identify which of this may have valuable information for the prices of corporate equity. Boudoukh, 2012 [2] classified 8 major News categories that move stock prices: Acquisition, Analyst Recommendation, Deal, Employment, Financial, Legal, Partnership, and Product. These categories are further classified into sub-categories to create a term hierarchy. Table 1. Below lists each of these related subcategories.

Main Category	Sub-Categories
Acquisition	Merger, Tender
Analyst Recommendations	Activist, Analyst Expectation, Analyst Opinion, Analyst Rating, Analyst Recommendation, Credit or Debt Rating, Fundamental Analysis, Price Target, Rating Agency List, Technical Analysis
Deals	Contract, Agreements, Exclusivity, Licensing, MOU, Negotiation, Pact, Service, Product
Employment	Board, Chair, Executive, Senior Executive, Workforce
Financial	Dividend, Stock Financial, Reports, Forecasts, Metric Change, Investment, Derivatives, Financing
Legal	Bankruptcy, Investigation, Judgement, Lawsuit
Partnerships	Alliance, Joint Venture, Termination
Product	Approval, Discontinuation, Expansion of Line, Issues, New Product, Recall, Submission, Testing, Trial, Updates

Table 1. News categories that may move stock prices.

Upon researching various text analysis frameworks we noted how language development models have developed over time. Initially, we began with long short-term memory, which is a

recurrent neural network that allows the machine to solve sequence prediction problems. However, it was determined that it requires a substantial amount of data for real-world applications, something of challenge given the scope of market moving Twitter events. We next found that transformer models are able to train faster and with more accuracy on text than recurrent networks, thus we investigated these architectures further. The transformer is an attention-based framework for modeling sequential information. It was proposed as a sequence-to-sequence model, therefore including encoder and decoder mechanisms. [1]

To be able to achieve our goal of using Twitter data to detect market moving events using transformers, we discovered a specific transformer called BERT. This model views the problem as a bi-directional one, allowing it to tokenize words into their smallest possible textual representation. The following diagram shows how BERT learns a language:

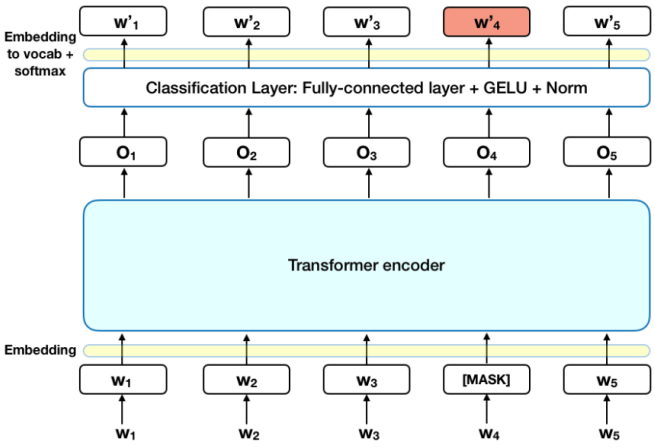


Figure 2. BERT Embeddings and Transformer Architecture.

BERT first embeds the token, then the segment (i.e., which sentence it is in), and, finally, the position of the token itself. Through this, the machine breaks up a string of text into tokens and embeds them in order to use them to understand the language.

Given this advantage, they are able to primarily solve 3 different text problems that pertain to our objective [3]:

Single-input classification tasks: these tasks usually entail classifying a single segment of a text. An example would be sentiment analysis, where the model understands what the text is saying and determines whether it is a positive or negative statement.

Two-input classification tasks: these tasks require the machine to observe two sentences, understand what they mean in relation to each other. For example, it would be useful in detecting if two inputs are paraphrases of each other.

Single-input token labeling task: this task focuses on grouping tokens of a single input to certain labels and labeling them accordingly. This would be beneficial when we want to automate the process of identifying which stock a tweet is about.

The architecture of Bert is especially important for us in understanding the meaning of tweets is relevant to stock market movements given the context factor. Figure 3. shows the benefit of bidirectional compared to sequential modeling for NLP interpretation.

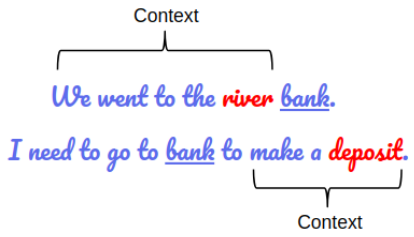


Figure 3. Sequential vs Bidirectional Model Considering Context

This is especially true in many of our categories, with many companies (such as Apple, Meta, and Visa) acting as both public companies and general English language terms. There have been previous works [4] on named entity recognition that emphasizes that the BERT model has capabilities for solving these problems with embeddings using a unified MRC framework. We find Bert is a State of the Art model that can identify the companies being mentioned, and understand the text in our categories which may be significant market moving.

Bert is a pretrained model that solves many sentiment, Q&A, and text classification problems. For our domain with focus especially on the financial domain. Liu et. al adapts the Bert Model to our specific domain, pretraining the data on financial corpus such as Yahoo Finance [4]. The proposed framework performs better than other state of the art methods on sentiment classification for financial information. We view our problem similarly as a sentiment forecasting model by tailoring the problem to labeled data as a magnitude of positive, negative, neutral to the normalized values of future stock returns at time t. We follow the FINBERT framework for model development and fine tuning.

The proposal is prepared as follows: first we discuss data used for initial twitter news impact, then we review feature preparation and extraction that may be useful for us. We follow that with a model development for architecture Bert in the context of twitter stock market data. Lastly we provide a potential framework for the data pipeline with conclusions on the research.

3. Data

Given the number of tweets published daily we begin the data collection process by filtering for individuals whose information is mostly likely to be distributed by the larger Twitter audience. We do so by identifying 410,000 Verified Twitter users. These accounts include journalists, business executives, investors, company pages, and various other titles. To proceed with a dataset that is useful for our initial research we further filter the usernames based on accounts located in the US or in the English language. This results in approximately 250,000 accounts. We collect the most recent ~2,000 Tweets (or less) from these accounts to create an initial dataset of approximately 125 million data points. We next shift our attention to identifying those pieces of information that may have an impact on the further equity returns for our universe.

To identify meaningful Tweets that contain market moving financial events we first consider three approaches:

1. We identify the companies in the S&P 500 in list their stock tickers ($Ticker) and abbreviated company names (after cleaning for corporate stop words, such as Inc., Corp., and Incorporated). When filtering for these Tweets we notice that the great majority are related to post-event analysis and the impact it had on the stock price. This is a valuable finding, and those potentially useful, we believe it is not the optimal method for detecting the events of significance - a first order problem in creating the dataset for our further model training.

2. We initially expand on the above process by filtering the Tweets identified based on an exact term match of profile keywords found in a username's description. We use exact terms related to both market practitioners and news disseminators: financial reporter, financial journalist, portfolio manager, investment manager, and equity analyst. The findings related to those in step 1 above: that all the tweets are relevant for market analysis, we are not detecting the events most commonly associated with future market movements. These Tweets are most often reactionary.

3. It is finally decided that the best approach for identifying market moving Twitter events is a data driven approach. We collect the one minute stock returns for S&P 500 companies during both market open and after hours. We then filter the dataset for great one minute percentage changes with a bar count during the minute of greater than 10. It is assumed that this price reaction with at least moderately high volume is a reaction to a specific market event. We will analyze all Tweets from the verified Twitter users during the preceding one minute to identify which, if any, Tweets contained information that may have led to this price movement.

To build our Twitter dataset we aim to annotate a corpus of 10,000 market moving Tweets. Given the existing literature we anticipate a focus on a few categories referenced in Section 2 above,

with an emphasis on events that occur frequently enough in our universe that we can build a significant enough training dataset for future model understanding. This initial focus is on the corporate events category, including: acquisitions, mergers, buy backs, stock splits, dividends, and index constituents. There were 309 share buybacks in the 2nd quarter 2021 by S&P 500 companies (a record high). We will further investigate the impact of these announcements via Twitter to identify which have an impact on future price movements. Other categories, such as automobile recalls and product defects occur in relative high frequency and may be considered for training data. We note the importance of language models in this problem to understand the contextual differences between 'Tesla Recalls New Fleet' and 'Soccer Player Recalled For National Team'. The approach we take with the Bert Framework is further explained in the following sections.

We note that the approach for data collection and identifying relevant events for Twitter training data can be expanded to news information in different time frequencies, including testing the market impact on higher frequency time series data. Further, these categories can be expanded to cover macroeconomic rather than equity specific events.

4. Feature Space

The BERT language modeling framework has built in features that allow for tokenization, stop words, and embeddings that is further discussed in the following sections. In this section we propose domain specific features from Twitter that may have correlations to future price movements. These include: username profile information, text contained in the Tweet and novelty of News. We note that a user's prior account history may have some significance for further consideration, though given the filter for verified users, we save this for future works. A user's profile contains valuable information in both the number of followers they have (a measure for how quickly information is spread, thus a measure for influence) as well as the description in their profile, which may indicate their perspective on news sources. We anticipate building a feature space that combines categorical, dummy, real values, and text. The feature expected to have the greatest impact on model performance is the raw text of a user's tweet. This is further explained in the Bert modeling section. We also consider two values for classifying whether the text is breaking news. Two measures, topicality and novelty, measure if a particular news topic has been published in the past, and also if the topic has been published by other sources. We find these two measures especially relevant for news sources, though we may consider their feature calculations in Twitter text analysis.

$$Nov(a_t) = \sum_{0<\Delta t \leq \tau} SIM(a_t, a_{t-\Delta t}).$$

$$Top(a_{t,k}) = \sum_{j \neq k, j \in K} SIM(a_{t,k}, a_{t,j}).$$

Novelty measures whether a particular topic has been published before. Topicality measure if a

particular news v.s. what is published by other agencies. In the two equations, At represents the frequency of word counts for a specific news topic and SIM is a measure for Cosine similarity between either the same source's prior releases or those of a third party.

The target variable for our training data is future price return at time T+1. For our label we assume T to be in minutes. Further – we normalize this value to be between -1 to +1 with a normal distribution for both positive and negative price returns.

4.1 Feature Importance

It is important to understand which Twitter specific features in our problem have greater statistical significance on future prices. Though an initial impulse is to research historical evidence from Twitter sentiment publications, the information that identifies causal relationships for stock prices varies from those literatures. We consider an example that applies a tree based model to provide insight about feature importance. We will likely first use a random forest to classify data using bootstrapping, a process of running multiple sub-samples for the data, combining the results into one. Figure 4 shows the forest's design.

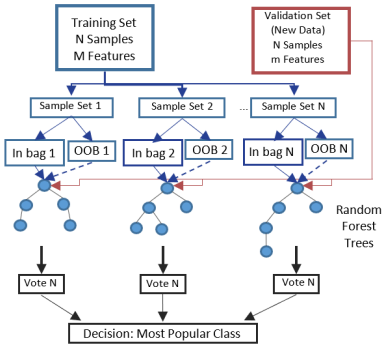


Figure 4: Random Forest For Feature Significance

The idea of feature importance is to measure how the feature impacts a model's performance. If we believe a feature is important, then if we assign a randomness with the feature, it will degrade the overall model performance. Therefore, for a trained random forest, we can permute the test set to calculate the prediction accuracy and compare it with the original accuracy. Permutation ensures that various tests on unique sections of the data yield similar results.

When analyzing feature importance, extraction, and dimensionality reduction we consider pre-trained frameworks using Python libraries such as Spacy. Since we are using BERT as our text analysis model, we inherently use term frequency–inverse document frequency, a technique considered in the model's design.

In the case that the random forest method does not yield fruitful results, we may resort to exploring other correlation techniques and choose the best one upon examining the results.

Given that the idea of extracting important features is relatively unexplored in financial Twitter data, we are still open to experimenting with other techniques.

5. Model Preparations

We introduce three processes for testing an algorithm on our dataset: in sample, validation, and out of sample. In sample is used to train a model to know which weights lead to maximum accuracy given an initial piece of data. Validation is used to analyze performance, then go back and optimize parameters of the model. The model is then rerun insample with the new settings, before looking at the validation results again. The third chunk of data is saved for a final test to see how the model performs on new information after being optimized. This is what the portfolio manager will analyze to determine whether an algorithm is worthy of moving to real time.

In traditional machine learning algorithms, the goal is to minimize a loss function, such as mean squared error, to attain optimal results. In the financial setting, further consideration is provided to profit compared with risk metrics. In this framework we propose a simple classification accuracy for next period stock return as our accuracy metric, and also view machine learning specific performance values. We discuss this, and the importance of not overfitting the model, in the next two sections.

5.1 Setting The Objective Function

Setting your model's objective is essential to view insights from multiple angles. In sample and cross validation can be aimed at either minimizing mean squared error or maximizing the accuracy of your prediction, the out of sample run while focus on maximizing direction of prediction accuracy above a baseline of 0.50. Should this be successful a future objective function can be tested to assess the profitability of a theoretical trading strategy that enters a position in the company for a specified time period.

5.2 Avoid Overfitting

A popular technique to avoid overfitting is to use cross validation. This is done by splitting our in-sample data into two pieces: in-sample training set and in-sample test set. In the standard cross validation technique, called -fold, we partition the data into k subsets, called folds. Then, we iteratively train the algorithm on folds while using the remaining fold as the test set (called the "holdout fold"), as shown in Figure 5.

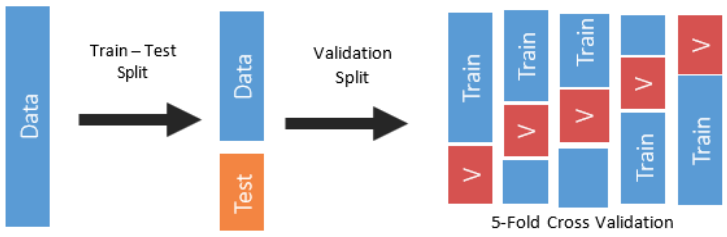


Figure 5. Data set split, training, validation and test.

In the financial world, data is sensitive to time series properties. To solve this problem, it is best to use the first set for in-sample training, and the th set for validation.

Another technique to avoid overfitting is called early stopping. As Figure 6 shows, when learning an algorithm iteratively, we can analyze the performance after each iteration. New iterations improve the model until it reaches an inflection point. At that time, the model's ability to generalize weakens, and it begins to overfit the training data.

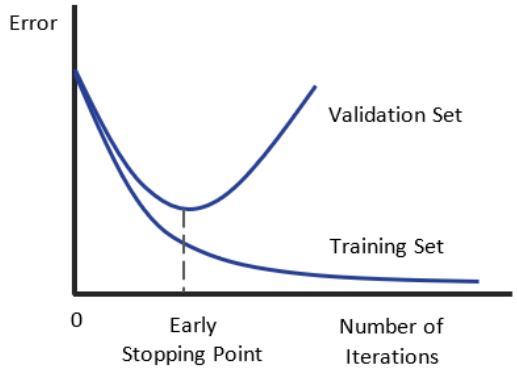


Figure 6: Early Stopping

Early stopping refers to stopping the training process before the algorithm begins overfitting. This process is commonly used in deep learning due to the massive amounts of data.

Now that we have techniques to reduce the chance of overfitting, it's important to quantify how strong the risk is. Bailey and Borwein (2014) defined the probability of backtest overfitting (PBO) and developed a framework (called combinatorially symmetric cross validation) to assess the reliability of a backtest. They concluded that their approach produces accurate estimates of the probability that a particular backtest is overfit. We borrow similar ideas to test our results. For our purpose, we plan to use 5 epochs since only the last layer of the FINBERT model is trained. More than 5 epochs may lead to more bias, and less epochs may lead to overfitting.

6. Modeling

Machine learning algorithms provide better forecasting accuracy than human or traditional statistical approaches can achieve. In a well known image recognition problem from Dodge, S. and Karam, L. (2017), humans and neural networks were both tasked with detecting objects. Both of their error rates are analyzed, with the human's being 5.1%, and the network only 3.6%.

This defeat of the human by a computer is becoming common ground, with similar happenings in chess and Go board games. Since the discovery of Graphic Processing Units by Nvidia, the algorithms are able to learn faster, and solve large scale problems. Figure 7 below shows the effect that greater computational power has on prediction accuracy.

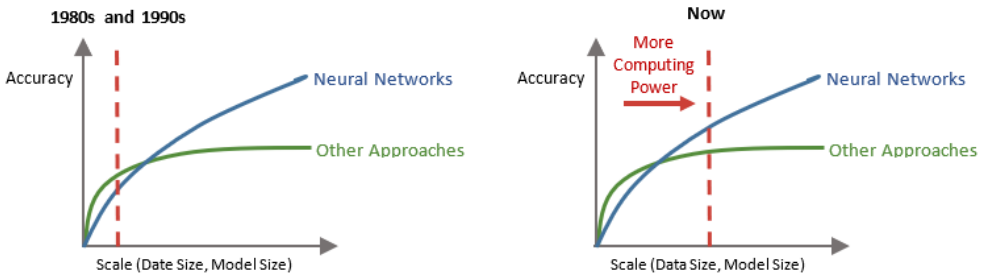
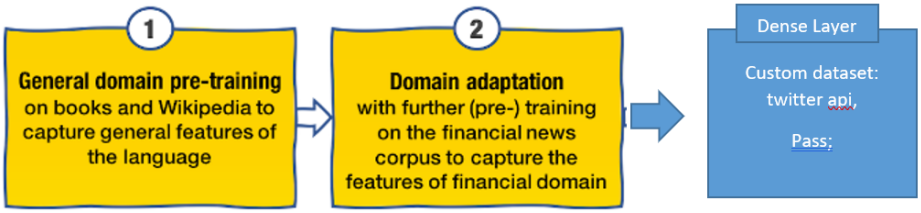


Figure 7. Relationship between data scale and model performance. Source: Jeff Dean, Google Brain.

6.1 Model Initialization

BERT-large was trained on 64 TPU chips for four days. We train this model further on the specific financial domain to improve its knowledge of this terminology. In step 1 below the model training on Wikipedia related data, then adapts to finance, and finally proceeds with our Twitter Dataset in the dense layer (explained below).



We can use transfer learning to prevent catastrophic forgetting, a common problem in neural networks where the machine forgets previously learned information. This brings us to a three-step process: slanted triangular learning rates, discriminative fine-tuning, and gradual unfreezing.

6.1.1 Slanted Triangular Learning Rates:

Slanted triangular learning rate applies a learning rate schedule in the shape of a slanted triangular, that is, learning rate first linearly increases up to some point and after that point linearly decreases.

6.1.2 Discriminative Fine-Tuning:

Discriminative fine-tuning is using lower learning rates for lower layers on the network. Assume our learning rate at layer l is α. Then for discrimination rate of θ we calculate the learning rate for layer $l - 1$ as $\alpha_{l-1} = \theta\alpha_l$. The assumption behind this method is that the lower layers. Represent the deep-level language information, while the upper ones include information for actual classification task. Therefore we fine-tune them differently.

6.1.3 Gradual Unfreezing:

With gradual freezing, we start training with all layers but the classifier layer as frozen. During training we gradually unfreeze all of the layers starting from the highest one, so that the lower level features become the least fine-tuned ones. Hence, during the initial stages of training it is prevented for the model to ”forget” low-level language information that it learned from pre-training.

Some initial values we may test for fine tuning the model are: Dropout probability p = 0.1, Learning rate=2e – 5, Warm-up proportion = 0.2, Batch size=64, Epoch = 4 to 6

6.2 Model Training

Given the model architecture we have designed is trained on a specific adaptation of the BERT framework to financial corpus (and our specific Twitter dataset) we recognize it as the FinBert model. We will train the model to see its predictive power on theTwitter testing dataset. A few metrics we will view for analyzing accuracy are classification (positive/negative) accuracy and F1 Score. Accuracy has been measured in a related model configuration (though not on stock magnitude) on a Financial PhraseBank dataset, the model yielded both an accuracy and F1 Score above 0.90. Given the noise in financial data we do not expect such results and set a baseline prediction accuracy of 0.50 percent. The performance metrics are available in Tensorflow. Should a successful model with above baseline performance be generated then we will next analyze the financial performance of a basic trading strategy to see if it is profitable after accounting for transaction fees.

The data pipeline for this twitter research is outside the scope of this work. We do provide a simple offline and online architecture that may work for a general framework of the modeling running.

7. Conclusion

Twitter is a popular source of news for retail and institutional investors. Information is often published on Twitter before news agencies. We propose a framework for categorizing and modeling Twitter information for identifying potentially marketing movement events. The framework we introduce focuses on corporate finance announcements from verified Twitter users. We frame the problem as a machine learning text analysis task, using a state of the art language model. The model is an evolution from BERT, called FinBert, that is trained on a domain specific corpus before we implement it for the Twitter dataset. We believe this framework, when implemented, may yield a meaningful accuracy that can be used by investment practitioners for the trading strategies. We further believe a derivative of this framework can be applied on other news sources, securities, and trading frequencies.

References

1. Huang (2018) Bloomberg-curated Twitter feed
2. Boudoukh, Feldman (2012) Which News Moves Stock Prices? A Textual Analysis
3. Lin, Nogueira, Yates (2010) Pretrained Transformers for Text Ranking: BERT and Beyond
4. Gluon (2022) https://nlp.gluon.ai/examples/sentence_embedding/bert.html
5. Liu, Huang, Huang, Li, Zhao (2020) FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining
6. Li, Feng, Meng, Han, Wu and Li (2020) A Unified MRC Framework for Named Entity Recognition