BlackRock Applied Finance Project: Using Natural Language Processing techniques for Stock Return Predictions

Ming Li Chew, Sahil Puri, Arsh Sood, Adam Wearne

March 27, 2017

Abstract

Our Applied Finance Project aims to develop a framework to determine if financial news headlines have meaningful impact on stock prices. This framework is a novel structure that primarily leverages on existing Natural Language Processing, including Name Entity Recognition,[36] and Global Vector for Word Representation (GloVe) model[56], before combining them with techniques such as k-means clustering[44] and portfolio optimization[28]. The subsequent study on events with predictive abilities could be of interest to institutional investors.

Starting with 1.8 million financial news headlines obtained from the Internet Archive: Wayback Machine[34], we successfully identified several events with meaningful post-event drifts. These events include situations where the equities of a firm are oversold, approval is given to a firm, a deal or agreement is signed as well as when an advisor is hired. The out-of-sample information ratios for these events are between a range of -1.02 and 0.76. The events we identified are by no means exhaustive, signifying the potential of our model.

Keywords: NLP, portfolio optimization, GloVe, k-means clustering, S&P 500, financial news headlines

Contents

1 Introduction

In May 2016, Business Insider published an article entitled *"A Giant Hedge Fund Used Artificial Intelligence to Analyze Fed Minutes"*, [74] highlighting the importance of transforming unstructured data into structured information to uncover investment opportunities. As a counter-point, MIT Technology Review reveals skepticism in artificial intelligence-driven trading strategies [37]. Traditionally, analysts scoured financial news to gather intelligence and information regarding firms in which they are interested. Over time, firms began to pursue sentiment analysis techniques like bag-of-words [2] to generate signals based on sentiments for their trading strategies. However, to what extent do these one-dimensional techniques capture structural relations in news texts? How much information is lost when investment strategies are driven by the sentiment of the articles? How do human investors analyze corporate news?

Conventional approaches of classifying words into a positive or negative category might miss out on the isolated effect of each individual corporate action. Our study attempts to investigate if there are future stock returns that can be attributed to individual corporate actions taken by firms.

1.1 Objectives

1. Use NLP techniques to map organization entity names and categorize corporate actions
A particular challenge in using NLP techniques is the ability to differentiate between different variations of entity names and categorize them accordingly [31]. For example, Goldman Sachs could be referred in the media as Goldman Sachs, Goldman or even GS. These variations will have to be recognized by the model as same entity in order to avoid the risk of unnecessarily losing information.

2. Identify and assess meaningful post-event drifts in stock value
We aim to assess the significance of events on the subsequent movements in stock prices. We plan to look at the sensibility of the event cluster centers to ensure that there is an underlying economic intuition behind the event performance.

3. Combine potential events in an information efficient manner
Once we identify the events which have potential future forecasting power, we plan to combine them into a portfolio using an optimization [28] approach that gives us risk-adjusted returns.

1.2 Significance of Current Study

Our study aims to extend the current literature in an attempt to connect financial news headlines with their impact on post-event stock returns:

1. Timeliness of Processing Unstructured Data
This research is especially timely as financial firms gain increasingly more access to unstructured data, with the hopes of gaining an informational advantage to value financial instruments. According to Gantz and Reinsel (2011) [25] unstructured information "accounts for more than 90% of the digital universe". With advancements in computational power, ease of availability of unstructured data, and recent advances in Natural Language Processing research, unstructured texts play an increasingly important role [74], especially since it is challenging to parse unstructured texts using automated algorithms [75]. By converting these unstructured text data into a more structured format, the market inefficiencies associated with unstructured data could be targeted.

2. Novelty in Methodology

In line with the existing literature, our study offers a comprehensive methodology that addresses the following two main components:

- The parsing of financial news headlines and the process of mapping them to tickers.
- Event study framework that entails post-event returns study and backtests.

Our methodology is original in the way we leverage on latest individual existing methods in the public domain, such as Name Entity Recognition [36], Global Vector for Word Representation (GloVe) model[56] combining them with known techniques such as k-means clustering [44] and portfolio optimization [28] in order to build a practical framework that can be used by institutions.

2 Literature Review

As NLP techniques gain popularity amidst the exponential availability of unstructured data, many research studies have been done using different perspectives and datasets. We studied multiple papers related to NLP and event studies, four of which are highlighted in this section due to their relevance to our study.

2.1 Ding, Zhang, Liu and Duan (2015): "Deep-Learning for Event-Driven Stock Prediction" [18]

One of the research papers that attempted to capture structured relations in news texts includes Ding, Zhang, Liu and Duan (2015) [18] who proposed to extract events as dense vectors before using a convolutional neural network (CNN) with learning event embeddings to predict and model the effect of the events on stock price movements. The rationale for using a CNN lies in its ability to:

- Learn semantic compositions.
- Store only the most useful local features.
- Combine the short-term and long-term effects of events.

Their paper employed an event-based trading strategy that relied on the signal generated by the model. Specifically, a long position would be taken if the model showed that stock prices will be higher the next day. A control test in the form of randomized positions would also be taken to assess the statistical significance of the trading strategy using outputs produced by the CNN.

Learning Outcomes Relevant for Our Study

Ding et al. [18] suggests that:

- Information embedded in events leads to better stock market prediction than only using words individually.
- Structural relations can be modeled well using semantic compositionality.
- While Ding et al. [18] trade on a very short horizon, we look at events which have a medium to long term impact (20-60 days).
- CNNs are more effective than other neural network based prediction models because of their ability to analytically take into account long term effects of events. However, the outputs are difficult to interpret because the methods through which non-linear weights assigned by the CNN are not transparent, hence one cannot be entirely confident of the sensitivity of the stock price predicted with respect to the events.

Even though their paper clearly showed that CNNs are a good method to assign non-linear weights to events in stock market prediction and to take into account short and long-term impacts of events, the biggest takeaway that is relevant for our study is the demonstrated informational value of structured relations within financial news texts. This gives us the motivation to not only determine if we are able to unlock the impact of structured relations but also to verify their predictive potential via events study analysis.

2.2 Herz, Ungar, Eisner and Labys (2003): "Stock Market Prediction Using Natural Language Processing" [31]

A pioneering study linking the use of NLP and stock market predictions, this patent publication aimed to construct a profitable trading strategy based on financial news. This was achieved by using NLP to parse financial news and identify events - specifically, entity names and actions - according to a simple template. Their aim is to predict the chances of appreciation and depreciation in equity value.

The inputs of the trading strategy outlined in this paper included not only the events parsed through financial news texts but also the written opinions, recommendations or articles by analysts or successful investors who have had higher historical streaks of making correct predictions in the stock market.

One interesting aspect worth noting about this study, considering that it is one of the earliest attempts to use artificial intelligence within the field of quantitative trading strategies, is that this study recognized and addressed a number of common challenges faced by NLP techniques even though this paper was written in the infancy stages of NLP application within the finance setting. A few example of the challenges are as follow:

- Categorizing variations of entity names as well as verb phrases. For example, the model should be able to recognize that "Goldman Sachs", "Goldman", "G.S." and "GS" mean the same thing. Furthermore, verb phrases such as "lay off", "fire", "let go", "sack", "downsize" also mean the same thing.
- Recognizing relations such as the acquirer or the "acquiree" in a merger and acquisition announcement.
- Addressing the challenge for the model to assign the correct organization if pronouns such as "it" is used in the new texts.

Learning Outcomes Relevant for Our Study

As one of the earliest papers to parse financial news texts for the purpose of stock market prediction, Herz et al. [31] provided valuable guidance in that regard. The three aforementioned challenges highlight the difficulties that we face in our research as well.

However, even though this patent publication encapsulated the general idea of using financial news data for the purpose of stock market prediction, our study differentiates itself from this publication in the following aspects:

- Herz et al. [31] did not go into details with regard to the specific statistical models or methodologies used to parse the financial headlines and map tickers to name entities. Our study leverages on individual existing methods in the public domain, such as the Name Entity Recognition [36], Global Vector for Word Representation (GloVe) [56] model, k-means clustering [44] and portfolio optimization [28], and combines them into a single framework.
- Herz et al.'s [31] parsing methodology is confined to a template, which classifies events according to the object and the relative or absolute change in the item. For example, the patent publication quoted a news headline "XYZ company announced that ... 20 employees would be laid off" where *employees* is tagged as the item and *20* as the absolute change. Our research parses financial news headlines based on the S, V, O structure and hence would have tagged "laid off" as the

Verb Phrase and "20 employees" as the Object. This marks a significant difference because they represent different methods and perspectives of analyzing financial news.

- While Herz et al.'s [31] cluster their templates based on the predictability of each template, we look at the similarity of our verb phrases based on a language model which has been trained independent of any performance based outcomes.

2.3 Xie, Passonneau and Wu (2013): "Semantic Frames to Predict Stock Price Movement" [77]

Xie et al. [77] are different from other current literature because their goal is not to build profitable trading strategies. Rather, they aimed to improve the understanding of companies through these companies' external media communications. In other words, Xie et al. recognized the informational value inherent within linguistic communication and used the semantic features of financial news as independent variables in order to predict changes in equity values, instead of conducting sentiment analysis. Because our study has an interest in a similar vein, Xie et al. proved to be an insightful reference.

The main contribution of this study lies in its methodology that is a combination of building a semantic framework, bag-of-words and sentiment scores. Specifically, this is achieved by the introduction of a novel tree representation. In this regard, the goal of this paper was two-fold, i.e. to predict binary changes of price and polarity. In order to do so, Xie et al.[77] use the support vector machine (SVM).

With the outcome of the SVM, Xie et al.[77] evaluated the accuracy of their predictions by using the Matthews correlation coefficient.

Learning Outcomes Relevant for Our Study

Although Xie et al. have built a strong technical framework to parse financial news, this research only studied three sectors and focused predominantly on financial accounting information. This may reduce the effectiveness of their proposed model within the context of the stock market because of the huge amount of data lost. Using the MCC alone is a one-dimensional assessment of the model that does not provide finance practitioners with perspectives of the volatility, market-adjusted returns, industry-adjusted returns etc of the signals.

In this regard, our paper will be a combination that aims to achieve an ideal balance of two important parts of such an exercise:

- Parsing of financial texts for event identification.
- Events study analysis testing the statistical impact of the events on stock prices.

2.4 Kothari and Warner (2004): "Econometrics of Event Studies" [38]

Kothari and Warner (2004) [38] claim that "event studies examine the behavior of firms' stock prices around corporate events". They defined abnormal returns as follow:

$$e_{it} = R_{it} - K_{it}$$

where

$$R_{it} = \text{Return of stock at the time of the event, } t$$
$$K_{it} = \text{Expected return of stock at time } t$$

Kothari and Warner (2004) [38] demonstrated a few methods to determine abnormal returns. These include:

- Aggregation of cross-sectional abnormal returns.

$$\text{Average Residual at time } t, AR_t = \frac{1}{N}\sum_{i=1}^{N} e_{it}$$

- Aggregation of abnormal returns across the entire time series.

$$\text{Cumulative Average Residual at time } t_1 \text{through time } t_2, CAR(t_1, t_2) = \sum_{t=t_1}^{t_2} AR_t$$

Kothari and Warner (2004) [38] also highlighted that properties such as specification, sensitivity of assumptions and the strength that the model holds against different opinions will vary between studies on short term basis and long term basis. Specifically, long-horizon event studies of one-to-five years are more challenging because appropriate risk adjustments are needed. This poses problems because any minor errors could result in vastly different abnormal returns. Furthermore, there is no consensus on the best long-term expected return model.

Learning Outcomes Relevant for Our Study

Our study focuses on the short-term behavior of S&P 500 stock returns after corporate actions to determine if there will be any interesting trends or insights in post-event drifts. Because the S&P 500 consists of large cap stocks, the financial news headline dataset we use is biased towards stocks of large cap companies. As a result, comparisons of returns are conducted within the universe of large cap stocks.

To determine if post-event drifts are attributed to chance or mispricing, we use t-statistics to examine the departure of the abnormal returns from the model expected returns as well as the associated standard error. This helps us to confirm the statistical significance of the abnormal returns.

3 Data

3.1 Stock-Level Information

Stock level information was obtained from the WRDS [62] Database. Specifically, we used two main databases in WRDS:

CRSP Database [6] (Daily Basis)

- Holding Period Return
- Price
- Number of Shares Outstanding

Compustat Database [26][67][66] (Quarterly Basis)

- GIC Groups [26]
- GIC Industries [26]
- GIC Sectors [26]
- GIC Sub-Industries [26]
- Index/S&P Membership Information [67]

Fama French Database[23]

- Fama/French 3 Factors [Daily]
- Momentum Factor (Mom) [Daily]

3.2 Financial News Headline

For financial news, we looked at historical webpages on the Internet Archive: Wayback Machine [34]. Wikipedia [76] defines it as "The Wayback Machine is a digital archive of the World Wide Web and other information on the Internet created by the Internet Archive". Using the algorithm provided at `github.com/hartator/wayback-machine-downloader`[30], we were able to extract the historical news articles from the newswire service *PRNewswire* [57]. Figure 1 below shows a sample article from which we extracted the headline.

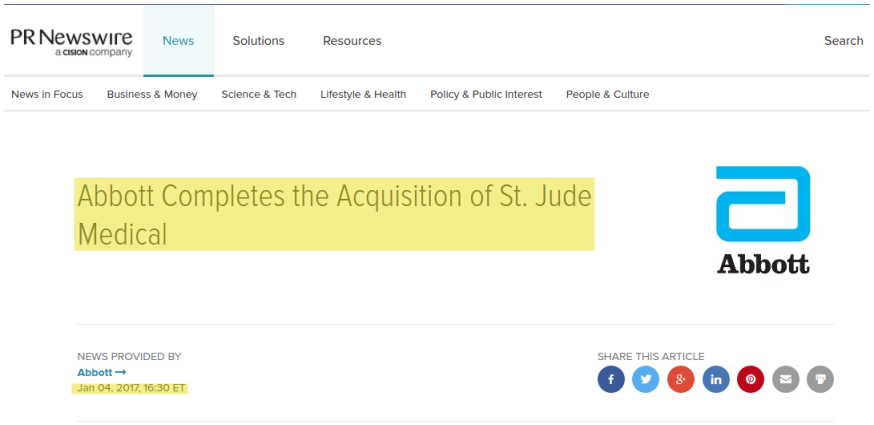


Figure 1: Sample article *http://www.prnewswire.com/news-releases/abbott-completes-the-acquisition-of-st-jude-medical-300385823.html*[30] extracted from the Wayback Machine archive. The headline and dates are highlighted.

One problem in parsing these articles lies in determining their date of publication. The format used to express the date varies widely across different websites. To counter this difficulty, we used the RegEx matching algorithm found in the *date-extractor* package. [17]

4 Methodology

The flow chart shown in Figure 2 captures our project's basic methodology:

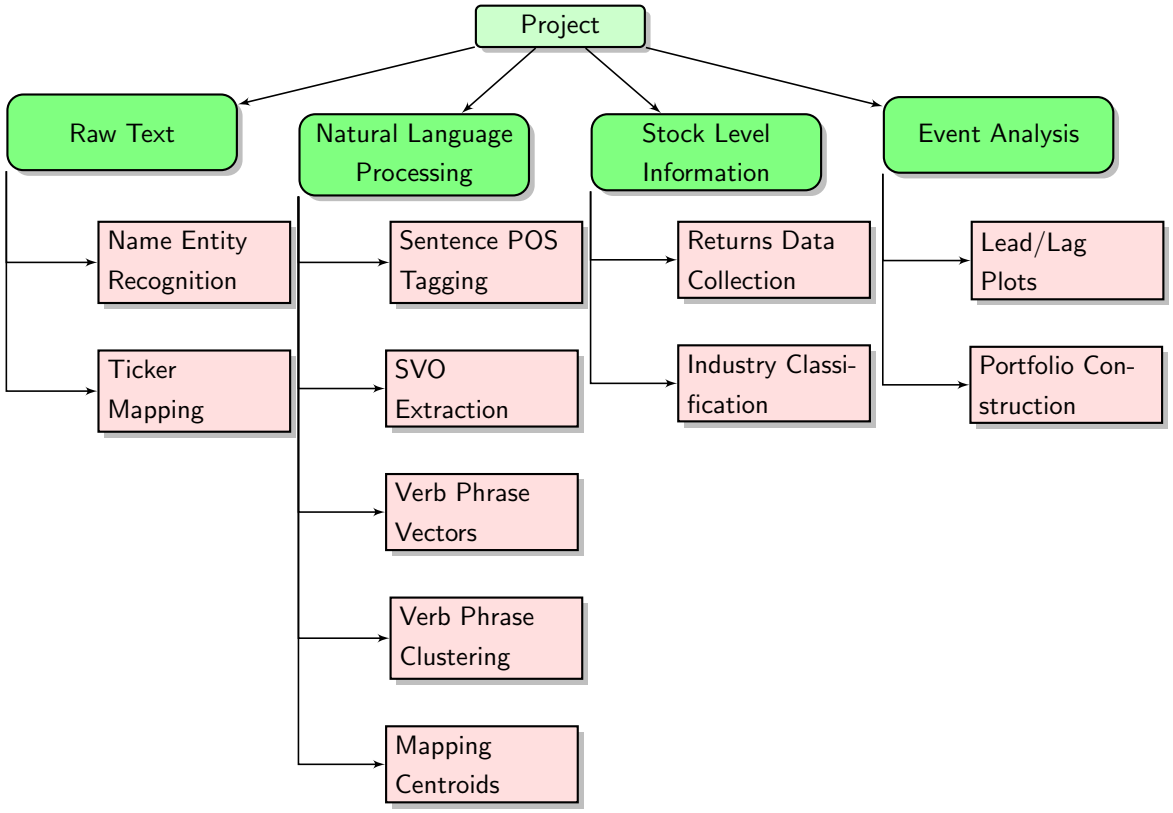


Figure 2: Project flow chart: This provides an overall breakdown of the project.

4.1 Raw Text

We obtained the raw text from the Wayback Machine [34] which contains all the headlines and date time stamps. Table 1 shows a few examples of the data collected:

Headline	Date
"Swing Trading Watch List: URI, WPX, ETN, GLW, ZTS" [71]	2016-09-26
"Corning (GLW) Crosses Pivot Point Support at $23.01" [11]	2016-09-23
"Interesting May 2017 Stock Options for GLW" [35]	2016-09-19
"When Will Corning Inc Split Its Stock Again?" [64]	2016-09-19
"3 Stocks to Hold for the Next 50 Years" [16]	2016-09-16

Table 1: A few sample headlines for Corning Incorporated (NYSE:GLW)

Note that:

- The headlines contain either stock specific information or industry/macro information.
- There are passive and active headlines. Passive headlines indicate information about the entity; active headline showcase an action performed by the entity.

For the purpose of this project, we limit ourselves to active headlines that relate to a specific company. These sentences follow the (Subject-Verb-Object) structure. This grammatical structure is advantageous because it is more amenable to POS tagging and NER using known NLP techniques.

4.1.1 Name Entity Recognition

The first stage in processing the raw headline data is Name Entity Recognition (NER). Sang et al., (2003)[72] defines named entities as "phrases that contain the names of persons, organizations and locations" . NER systems were initially developed using a rule-based methodology. However, machine learning techniques are now widely used [51].

In our problem setting, the NER process takes the headline as input and classifies each word in the headline into one of several categories. The purpose behind this is to identify the entities in a given headline as organizations before attempting to map these organizations to their corresponding tickers.

In our analysis, we used the Stanford NER implementation provided by Stanford's Natural Language Processing Group [36]. We have used the interface to Stanford's NER tool provided by Python's NLTK module, which was originally written in Java. Figure 3 illustrates how Stanford's NER operates via their web-interface [68].

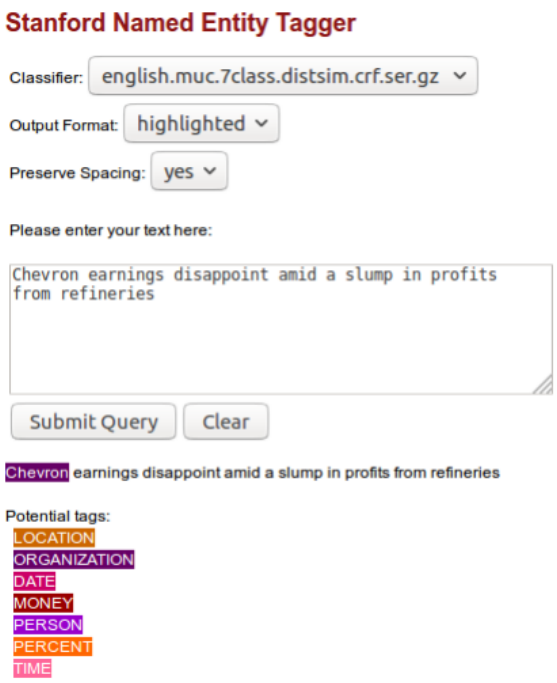


Figure 3: Stanford NER example

While there are several models available with varying numbers of classes to do the NER tagging, our experience with the data set has shown the 3-class model to be most successful in identifying organizations within a headline.

In brief, Stanford's NER was built by layering a Markov Chain Monte Carlo (MCMC) method on top of their usual Random Forest structure for identifying named entities [36]. The purpose of this was to incorporate non-local information within the text to reduce error in information extraction. In particular, Finkel et al. (2005) note that such a methodology can be used to improve NER systems[36].

Let us briefly illustrate how the NER process works on a sample headline. The following headline was published on Forbes.com [9] on March 18, 2015:

"Fastenal Moves Up In Market Cap Rank, Passing NVIDIA"

Before passing this to the NER operation, we apply a simple preprocessing step to remove punctuation and standardize the capitalization. During our initial tests with Stanford's NER, it was found that removal of stop words actually hindered performance, so we have elected to leave them in. Following this preprocessing step, the text then reads:

"FASTENAL MOVES UP IN MARKET CAP RANK PASSING NVIDIA"

Passing this headline through to the NER, we obtained the following list of tuples.

```
[('FASTENAL', 'ORGANIZATION'),
 ('MOVES', 'O'),
 ('UP', 'O'),
 ('IN', 'O'),
 ('MARKET', 'O'),
 ('CAP', 'O'),
 ('RANK', 'O'),
 ('PASSING', 'O'),
 ('NVIDIA', 'ORGANIZATION')]
```

Listing 1: NER classification results on sample headline

The NER has tagged two words as organizations in this headline. With the NER step now completed, we can attempt to map these tagged words to tickers.

4.1.2 Ticker Mapping

The next phase is to map our list of organizations to their corresponding tickers. We first need to build a database of company name and ticker pairs to which we can attempt to match our list of organizations compiled in the NER step.

Quarterly company financials were obtained from the WRDS database [62]. These financials are the data of company names and tickers to which we compare our organizations. In order to properly compare the strings from our list of organizations to those gathered from WRDS, we apply similar preprocessing transformations of capitalization and removal of punctuation. In addition, because the full official company names as listed on WRDS are not likely to appear in headlines, we also remove common words and phrases (ex. Inc, LLC, LTD, etc.) from the company name. The 25 most common words were then removed from company titles to increase the probability of successfully mapping organization names in headlines. Figure 4 below displays the frequency distribution of the 100 most common words found in the WRDS data:

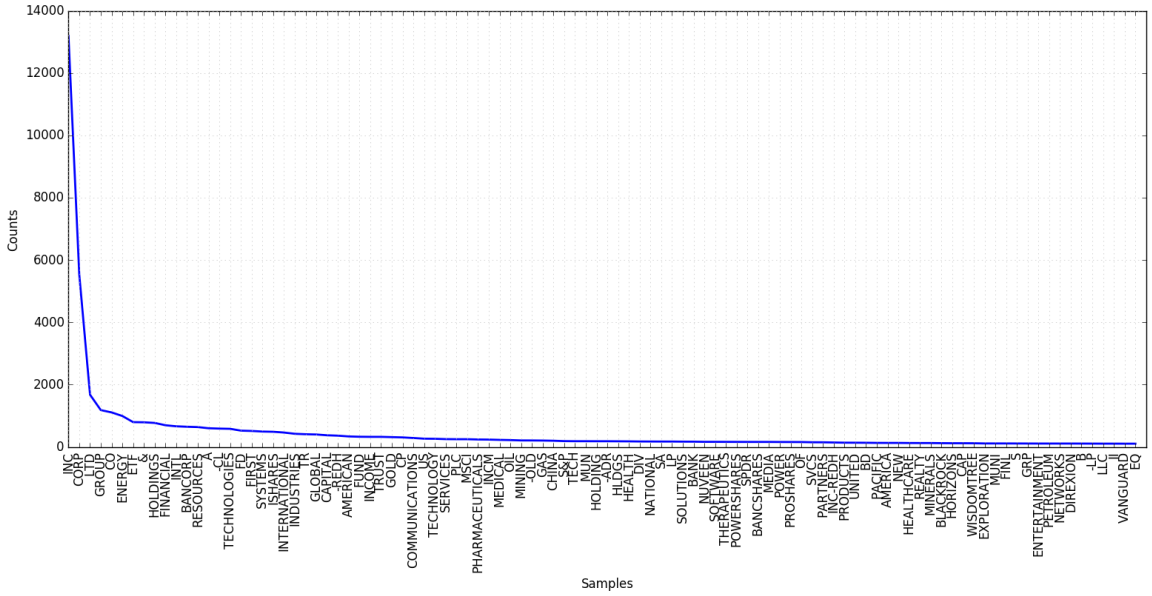


Figure 4: Distribution of most common words in WRDS dataset

With the database of name-ticker associations, we take into account one other consideration before searching our tagged organizations against this data. Because our data is tokenized, each tagged organization from the NER may correspond to only part of the company's name. To address this, we combine sets of adjacently tagged organizations so that we can pick up company names with multiple words. Applying this rule to our previous example, we would search the WRDS dataset against three possible names: "FASTENAL", "FASTENAL NVIDIA", and "NVIDIA". For each case, we then find potential string matches using the edit distance between the tagged organization versus the list of company names. If a successful match is found, we then edit our original headline to replace the identified company with its ticker. Additionally, we add a prefix of two underscores to each ticker name to make it easier to extract them at later stages of our analysis. Applying this to our example headline, we obtain:

"__FAST MOVES UP IN MARKET CAP RANK, PASSING __NVDA"

In addition to the identification with NER and subsequent ticker mapping, a second methodology used to identify firms within a headline takes advantage of the grammatical structure of the headline itself. We shall discuss this methodology in more detail in the following sections, but the main idea is to identify those headlines which follow the grammatical structure of Subject - Verb - Object. Similar to the case for NER, we can then attempt to map the subject (or object) against the list of company names and map it to a ticker. Take for example the headline:

"Corning Announces Investment in Versalume LLC" [15]

After identifying "Corning" as the subject of the headline, we can then simply map the subject of the sentence to its corresponding ticker and get:

"__GLW Announces Investment in Versalume LLC"

More details regarding the SVO methodology are discussed in the following sections.

4.2 Natural Language Processing

Natural Language Processing has multiple definitions. Chowdhury, G. (2003) [12] defines Natural Language Processing as "Natural Language Processing (NLP) is an area of research and application that

explores how computers can be used to understand and manipulate natural language text or speech to do useful things". The *useful things* in our project are the actions performed by firms.

4.2.1 Sentence POS Tagging

Part-of-Speech(POS) tagger is defined by Stanford [65] as "a piece of software that reads text in some language and assigns parts of speech to each word". POS Tagging involves training a model on some corpus, making it a supervised task.

The earliest corpus was the Brown Corpus [39]. However, the models that are trained these days are trained on the Penn tag set found in the Penn Treebank [46].

Initial models were linguistic models, where rules such as "no verb after an article" were hard coded. [27] and were followed by the Hidden Markov Models[40] and the Conditional Markov Models [58]. The latest models use a Linearly Dependency network [73]. These models are also called "bi-directional" since they also consider the tags on both sides of each word to predict the tag of the current word. Another class of POS taggers are the Perceptron or Machine learning class. One of the first of such models was Collins (2002) [13].

For the purpose of our research, we use the Python package textacy [10] (which is a layer built spaCy [32][19]) and StanfordPOSTagger [73]. Recall our previous example:

"Corning Announces Investment in Versalume LLC"

If we parse this headline through the StanfordCoreNLP POS tagger, we get the results shown in Table 2:

Word	POS Tag	Tag Meaning
Corning	NNP	"Proper noun, singular"
Announces	NNPS	"Proper noun, plural"
Investment	NNP	"Proper noun, singular"
in	IN	"Preposition or subordinating conjunction"
Versalume	NNP	"Proper noun, singular"
LLC	NNP	"Proper noun, singular"

Table 2: POS tagging results & interpretation for raw sample headline

Clearly, the tagger has mis-specified the verb "Announces". POS taggers are very sensitive to case in words. Financial headlines, however, use the title format which capitalizes each word. If we change "Announces" to "announces" and re-run this headline though the tagger, we get the results shown in Table 3.

Word	POS Tag	Tag Meaning
Corning	NNP	"Proper noun, singular"
announces	VBZ	"Verb, 3rd person singular present"
Investment	NNP	"Proper noun, singular"
in	IN	"Preposition or subordinating conjunction"
Versalume	NNP	"Proper noun, singular"
LLC	NNP	"Proper noun, singular"

Table 3: POS tagging results & interpretation for sample headline with lowercase verbs

4.2.2 Subject Verb Object Extraction

Our next objective is to extract the SVO structure from the sentences. In order to do so, we need to figure out the dependency structure of the sentence.

If we run the headline with lowercase verbs through the online weblink provided by Stanford Parser [69] we get the following results:

Figure 5: The textbox at the top allows us to enter our query. Walking through the results, we are provided with the query in raw text form, the POS tags for each word in the sentence, the dependency structure of the sentence in a tree format and, finally, the dependency structure of each word.

The results show that the sentence above consists of a main verb (linked to the ROOT) which is connected to the subject of the sentence, i.e. Corning in this case. Finally, towards the right of the Verb we have the Object, Investment.

There exist rule-based algorithms that are used to extract the SVO structure from a sentence. An example is Delia et al. (2007) [61], which shows an explicit algorithm in order to get the SVO triplets.

In our project, we compared a simple grammar rule-based approach with Stanford POS tagger[73] and textacy[10] which uses the in-built *subject_verb_object_triples* function. Both results extract the following information:

- Subject: Corning
- Verb: announces
- Object: Investment

Due to time limitations and textacy's speed, we chose textacy [10] for our purpose of converting unstructured sentence to a semi-structured format.

4.2.3 Verb Phrase Vectors

Once we apply name entity recognition and information extraction from POS tagging to our raw text we then define our verb phrase as the identified verb + object pair. From our continuing example we see that after having tagged Corning to Corning Inc's ticker(GLW) we get the verb phrase *announces Investment* by combining our Verb and Object.

In order to map this phrase to a vector space, we first need to be able to map a word to a vector space. Bengio et al. (2003) [4] came up with one of the earliest *classical neural network model* to accomplish this. The model is trained in order to predict a word given the previous n words. While the approach is simple, it is computationally slow [3]. Collobert & Weston (2008) [14] replaced the softmax function with a ranking method. Mikolov et al. (2013) [47] suggested the word2vec approach which uses a single-layer architecture and trains words based on information both sides of the word. The word2vec models have become increasingly popular [60]. The two distinct approaches suggested by Mikolov et al. (2013) [47] are the continuous bag-of-words (CBOW) model and the skip gram model. The CBOW model tries to predict a word based on the surrounding words on both sides whereas the skip gram model flips the problem and predicts the surrounding words of a given word. Finally, the Global Vector (GloVe) model introduced by Pennington et al. (2014) [56] explicitly trains the word vectors to match the co-occurrence statistics implicitly.

Pennington et al.(2014) [56] cite various benchmarks used to compare popular models. The word similarities benchmarks are as follow:

- "WordSim-353 (Finkelstein et al., 2001)"[22]
- "MC (Miller and Charles, 1991)"[49]
- "RG (Rubenstein and Goodenough, 1965)"[59]
- "SCWS (Huang et al., 2012)"[33]
- "RW (Luong et al., 2013)"[43]

Pennington et al.(2014) [56] show that GloVe models perform the best. A few GloVe models are shown below[55]:

Model Name	Number of Dimensions	Tokens	Vocabulary	Trained on
glove.6B	50/100/200/300	6B	400K	Wikipedia 2014 + Gigaword 5
glove.42B	300	42B	1.9M	Common Crawl
glove.840B	300	840B	2.2M	Common Crawl

Table 4: Descriptive overview of various GloVe models

For our project, we tried both the *glove.42B* and *glove.840B* models. However, due to computational limitations, we decided to choose the *glove.42B* model.

The dimensions in the GloVe models represent different characteristics or features of a word. A feature of this approach is the different interactions that are possible with the words. For example, if we

take the word vector for `king`, subtract the word vector for `man` and add the word vector for `woman`, the resulting word vector is very close to that of `queen`. [56][48]

$$v_{king} - v_{man} + v_{woman} \approx v_{queen}$$

Thus, the first step is to reduce the Verb and the Object parts of the sentence to a single vector. We achieve this by applying an average of each of the verb words and object words. For example, if our Subject, Verb, Object tuple is ((s_1, s_2), (v_1, v_2, v_3), (o_1, o_2)), we define our verb vector and object vector as follow:

$$v_{verb} = sum(v_{v1}, v_{v2}, v_{v3})$$
$$v_{obj} = sum(v_{o1}, v_{o2})$$

To convert our verb phrase into vectors, we have two different techniques:

1. Averaging: Take the average of the verb and phrase vectors such that $v_{phrase} = average(v_{verb}, v_{obj})$
2. Concatenation: Concatenate the verb and object vectors under each other such that $v_{phrase} = [v_{verb} v_{obj}]^T$

An issue with first technique is that the cluster centroids might lose the interpretability of the individual verbs and objects. On the other hand, the concatenation technique ensures that verbs and objects are separate in the cluster centroids. Since we plan to show sensibilities of cluster centroids, we selected the concatenation method of creating the phrase to vector representations.

4.2.4 Verb Phrase Clustering

After obtaining the verb phrase vectors, we cluster these verb phrases into bins. These bins represent the potential event clusters that will be used in the next step, i.e. event analysis. For the clustering step, we looked at the following approaches:

- **K-means clustering:** K-means is a very old clustering algorithm. While the idea can be traced back to Steinhaus(1956) [70] a more accessible description can be found in Macqueen(1967) [44] which is as follow:
  - The number of clusters k, to be identified within the input data, is decided initially.
  - k centers are then chosen from the input data in a random fashion.
  - Next, the algorithm iterates through each new data point and assigns them to the nearest cluster mean, on the basis of the maximum similarity (minimum distance) between the data points from all of the k cluster means. If the new data point is assigned to the i-th cluster, then the i-th cluster mean is updated accordingly (after incorporating the new data point coordinates).
  - Once each data point has been assigned to a cluster, the algorithm again iterates through each data point to ensure that the current cluster mean is still the nearest for all the data points. If it is not the case, then those data points are re-assigned to a new cluster and the process is continued until there are no new re-assignments in an entire iteration.
- **Affinity Propagation:** Frey & Dueck (2007) [24] introduced the Affinity Propagation clustering algorithm. The advantage of this algorithm is that the total number of clusters is not required as an input. The algorithm comes up with the optimal number of clusters for the given input data points using the similarity metrics. The similarity metrics between the word phrases is obtained using the following methodology:

– For each pair, we calculate the similarity matrix by taking the dot product of the vectors corresponding to each phrase and normalizing the same. The similarity between verb phrase i and j is obtained as follow:

$$sim_{i,j} = \frac{vec_i \cdot vec_j}{||vec_i||_2 ||vecj||_2}$$

where vec_i represents the vector corresponding to verb phrase i and vec_j represents the vector corresponding to verb phrase j.

Due to the large number of financial news to be clustered and memory storage limitations, we were not able to pursue the affinity propagation technique and have left it for future research.

4.2.5 Mapping Cluster Centroids

Upon obtaining our clusters, the next step is to map the corresponding centroid for each of them back to our text corpus so that we can gain interpretation regarding the actual meaning of each cluster. In order to perform this mapping, we construct a ball tree via Python's scikit-learn module.[54][5]

Liu et al,. (2006)[42] offer a detailed implementation of the algorithm. A ball-tree is often used to partition large sets of high dimensional data into a binary tree structure. Once constructed, it can be used to identify nearest-neighbors for a given vector. This methodology provides a framework for mapping cluster centroids to word vectors in our corpus.

In our study, we have used 100 as our maximum number of clusters. The GloVe model[56] we use has a vocabulary size of approximately two million words. In order to speed up the construction, we filter the corpus and remove non-relevant words such as stop words, single-token strings as well as the words not observed in our headline dataset.
An example of the effectiveness of our mapping can be seen using a couple of test cases which are known to have financial relevance. Consider the words "earnings" and "layoff" which both appear in our corpus. Constructing our tree and then querying it against these two words yields the nearest-neighbors shown in Table 5. Note that the words themselves appear because by construction they are zero-distance from themselves.

earnings	**layoff**
profit	lay-off
profits	layoffs
income	lay-offs
revenue	demotion

Table 5: Sample results for querying ball tree against known words

From the results on the simple test cases, we can see that this methodology yields sensible results. Naturally, as one extends the number of nearest-neighbors, fringe results will begin to appear at some point. During our tests with this framework we have found that selecting five closest neighbors provides a balance between sufficient depth while still retaining sensible words.

4.3 Stock Level Information

We use stock-related information from WRDS [62] in order to gauge the financial impact of company actions. For our project, we only focus on the S&P 500 stocks for the following reasons:

- There is not enough data available in form of news wires for smaller firms. This can cause issues while running factor return regressions.
- There is a liquidity risk premium[53] associated with smaller firms[1]. In order to avoid biases without creating a liquidity risk framework ourselves, we can constrain investments to liquid S&P 500 stocks.

The information is broken down into the following components below.

4.3.1 Returns Data Collection

For all the stocks that are traded, we use the holding period return in the CRSP dataset[6] from WRDS[62]. This adjusts for all stock splits, stock or cash dividends and assumes that any proceeds that are paid out are immediately re-invested in the stock.

4.3.2 Industry Classification

For each stock, we have GICS industry classification.[26] This classification allows us to perform industry neutral analysis as certain actions might be dominated by a particular industry. In order to ensure that our results do not erroneously attribute industry performance to an action, we can remove the equal-weighted industry mean and run the analysis.

4.4 Events Analysis

In order to understand the impact on returns caused by our identified event clusters, we have conducted a detailed event study investigation. In this section, we discuss the framework used to produce our lead/lag plots as well as briefly review some basic and well understood models of expected returns. Following this, we describe the optimization framework used to translate news headline clusters into portfolio weights.

4.4.1 Lead/Lag Plots

Let us begin with a brief review of some of the most famous asset pricing frameworks and their extensions. Firstly, let us discuss the Capital Asset Pricing Model (CAPM)[63][41]. In short, the CAPM provides a very simple framework which details the expected return of an asset based on the amount of market risk of the asset. This relationship is captured in the equation:

$$E[r] = r_f + \beta\left(E[r_m] - r_f\right)$$

In this equation, r_f is the risk-free rate, the expected return from the market is given by $E[r_m]$ and $E[r]$ denotes the expected return on our asset.

The quantity β captures the amount of market risk of our asset and is given by:

$$\beta = \frac{\mathrm{Cov}(r, r_m)}{\mathrm{Var}(r_m)} = \rho\frac{\sigma}{\sigma_m}$$

Here, ρ is the correlation coefficient between our asset and the market, and σ^2, σ_m^2 represent the variance of the asset and of the market respectively.

While the CAPM provides a nice starting point, there have since been many observed pricing anomalies which the CAPM fails to catch [21] [20]. Some of the most notable instances include the larger returns observed for small-cap stocks compared to what CAPM would predict [20]. In a similar vein, value stocks also exhibited significantly higher returns than what would have been expected by the CAPM alone. [20]

To build upon the framework developed in the CAPM, Fama and French[20] developed a factor model which incorporates both the size of the firm, as well as its book-to-market equity. The inclusion of such factors gives the model more explanatory power.

The equation for the expected asset return then takes a similar form to that of the CAPM, but now with these two new factors introduced we get:

$$E[r] = r_f + \beta_1 \left(E[r_m] - r_f\right) + \beta_2 \text{SMB} + \beta_3 \text{HML}$$

SMB and HML stand for "Small Minus Big" and "High Minus Low" and are defined as risk factor returns to size and book-to-market equity respectively [20]. These factors function as a type of risk premia.

Following the Fama-French Model, one of the next extensions was the inclusion of a momentum factor[8]. Momentum factors would suggest that increasing prices should be followed by further increases and falling prices should fall further still. Taking this factor into consideration, we obtain:

$$E[r] = r_f + \beta_1 \left(E[r_m] - r_f\right) + \beta_2 \,\text{SMB} + \beta_3 \,\text{HML} + \beta_4 \,\text{UMD}$$

where UMD stands for "Up Minus Down".

With an understanding of some of the various models used for expected returns, we can then incorporate them into our event-study framework to examine the impact that various clusters of news headlines have on asset returns. We hope to capture the abnormal returns, which are the total returns minus the expected stock returns from a given model.

Lead/Lag plots were produced to measure the impact of events on returns. Three methodologies were used in this endeavor, namely:

- Market Adjusted Returns
- Industry Adjusted Returns
- Fama-French + Momentum Adjusted Returns

We applied a 252 day rolling window and ran a regression to obtain estimations for our β coefficients as well as estimations for the abnormal returns. Further, the abnormal returns are then adjusted by the standard deviation calculated on another 252 day rolling window. Utilizing multiple expected return frameworks in this way allows us to more systematically investigate the effects our identified events had on returns while accounting for well known factor exposures. This allows us to more effectively isolate the impact that an event might have on returns. Our analysis showed that the event study results are similar across all methodologies. Hence, we only show market adjusted and scaled returns in our results in section 5.

4.4.2 Portfolio Weight Optimization

Having identified the event clusters, we transfer our results to a portfolio allocation framework[28]. Our strategy entails building a long/short portfolio following a total risk optimization scheme. Our aim is to produce portfolio weights such that we have no exposure to the market, size, and value while going long on stocks for which events have occurred, and shorting the same industry stocks for which there are no events. A more detailed view of this process is given below.

On any given day, we will have an input vector, $\mathbf{L}$, which indicates what stocks in the event cluster were active on that particular day. The dimensionality of $\mathbf{L}$ is $n \times 1$ where n is our *active stocks*. In order to calculate these active stocks for each given day, we take the S&P 500 stocks as of that day and remove the GICS industries for which there were no events. In our active stocks, we would like to go long on those stocks for which event occurred, and short the remaining. The next focus is then to construct our portfolio weight vector $\mathbf{w}$.

The initial choice would be to construct the minimum variance portfolio subject to a set of factor-neutral constraints:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \mathbf{w}^T \, \Sigma \, \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T \beta_i = 0, \; i = 1, 2, 3 \end{aligned}$$

where Σ is our covariance matrix for the returns of all stocks within our universe. The quantities β_1, β_2, and β_3 are our factor coefficients for the market, value, and size respectively. The setup of this problem looks sensible. However, there is the distinct problem that our optimization routine could simply return zero portfolio weights to satisfy the minimum variance condition as well as the constraints. To subvert this issue, we incorporate the vector $\mathbf{L}$ as well as $\mathbf{S} \equiv \mathbf{1} - \mathbf{L}$ into our constraints. We introduce them such that the optimization problem becomes:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \mathbf{w}^T \, \Sigma \, \mathbf{w} \\ \text{subject to} \quad & \mathbf{w}^T \beta_i = 0, \; i = 1, 2, 3 \\ & \mathbf{w}^T \mathbf{L} = c \\ & \mathbf{w}^T \mathbf{S} = -c \end{aligned}$$

where c corresponds to the size of our position on the long/short side. Reformulating this problem in terms of Lagrange multipliers, we seek to solve:

$$\underset{w}{\text{minimize}} \quad \mathbf{w}^T \Sigma \mathbf{w} - \lambda_1 \left(\mathbf{w}^T \beta_1 \right) - \lambda_2 \left(\mathbf{w}^T \beta_2 \right) - \lambda_3 \left(\mathbf{w}^T \beta_3 \right) - \lambda_4 \left(\mathbf{w}^T \mathbf{L} - c \right) - \lambda_5 \left(\mathbf{w}^T \mathbf{S} + c \right)$$

Solving this for the weight vector, we obtain the following closed form solution:

$$\mathbf{w} = \Sigma^{-1} \left(\lambda_1 \beta_1 + \lambda_2 \beta_2 + \lambda_3 \beta_3 + \lambda_4 \mathbf{L} + \lambda_5 \mathbf{S} \right)$$

Finally, we can solve for our set of Lagrange multipliers via the following set of equations:

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \end{bmatrix} = \begin{bmatrix} \beta_1^T \Sigma^{-1} \beta_1 & \dots & \beta_1^T \Sigma^{-1} \mathbf{S} \\ \vdots & \ddots & \\ \mathbf{S}^T \Sigma^{-1} \beta_1 & & \mathbf{S}^T \Sigma^{-1} \mathbf{S} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \\ 0 \\ c \\ -c \end{bmatrix}$$

Using the solutions to these sets of equations, we can then directly incorporate news headlines that belong to a given cluster into portfolio weights. Finally, as a last step to prevent any particular weight from becoming too dominant, we rescale the final portfolio weights using the following prescription:

$$\mathbf{w}^* = \frac{\mathbf{w}}{\max(|\mathbf{w}|)}$$

In essence, this rescaling of the portfolio weights actually has the same effect as adjusting the parameter c. However, doing so as a final step ensures that our portfolio allocations remain sensible and that we do not take on very large idiosyncratic bets. One important clarification to make is in the case when no news events occur for any of our clusters. Rather than shorting the entire set of stocks as the above

section would suggest, no positions are held if no events are detected.1

5 Results

5.1 Name Entity Recognition

With 1.8 million headlines obtained from the Internet Archive: Wayback Machine [34], upon running NER and SVO, approximately 400,000 headlines were successfully tagged as containing one or more company names. Figure 6 is a snapshot of the top 75 most common tickers that our methodology was able to tag. In general, we observe that the most common tickers are generally technology related firms. These findings are likely a combination of these two reasons:

- These firms appear in headlines with much greater frequency, thus giving our methodology a higher likelihood to detect them.
- The names of tech-related firms tend to be more distinct, making them easier for NER/SVO to recognize.

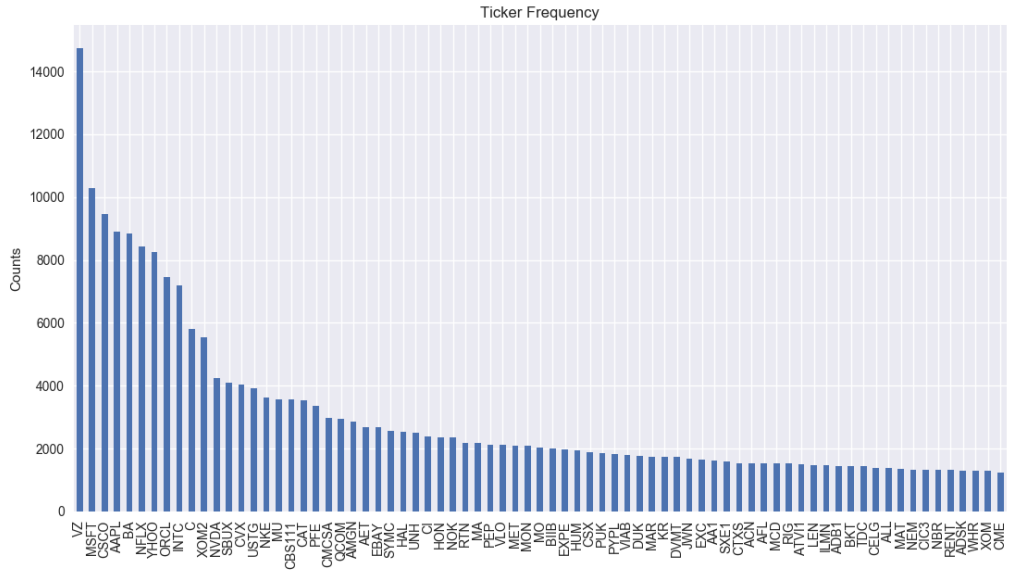


Figure 6: Distribution of most common tickers found by NER

5.2 Summary Statistics of Sample Data

We implemented three different filters on the 1.8 million headlines we obtained from the Internet Archive: Wayback Machine [34], after which we have 60,949 headlines relevant for the purpose of our study. The filters are as follows in the order in which we applied them:

- SVO script: To identify Subject, Verb and Objects from financial news headlines.
- Ticker mapping: To shortlist headlines which have a valid corresponding ticker.
- S&P 500 and Phrase Vector: To determine headlines related to S&P 500 stocks as well as those that consist of phrases for which we are able to find word vectors.

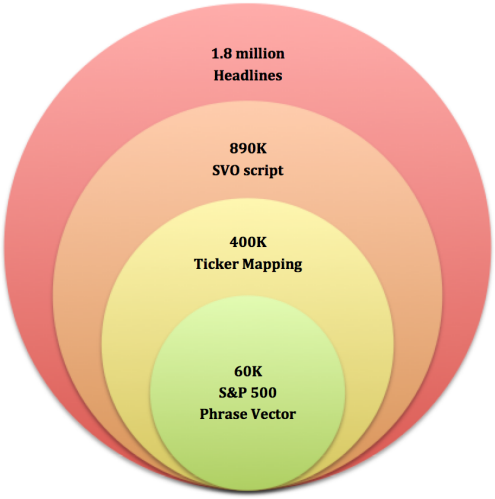


Figure 7: Layers of filtration on the initial dataset

Output examples of the parsed 60,949 financial headlines, including their mapped tickers, are displayed in Table 6 below:

Subject	Verb	Object	Date
CVX	Didn't Buy	Tesoros Hawaii Refinery	6/18/13
FB	Plans	IPO	11/28/11
LEN	Upgraded	To Buy	11/25/08
MRK	Crushes	Q3 Expectations	10/25/16
PLCN	Launches	Ad Campaign	9/23/13
SLB	To Reduce	Venezuela Operations	4/12/16

Table 6: Sample SVO results with ticker mapping for various headlines

The headlines span a timeline of 10 years from 2006 to 2017, with the majority of the headlines being more recent as shown in Figure 8 below:

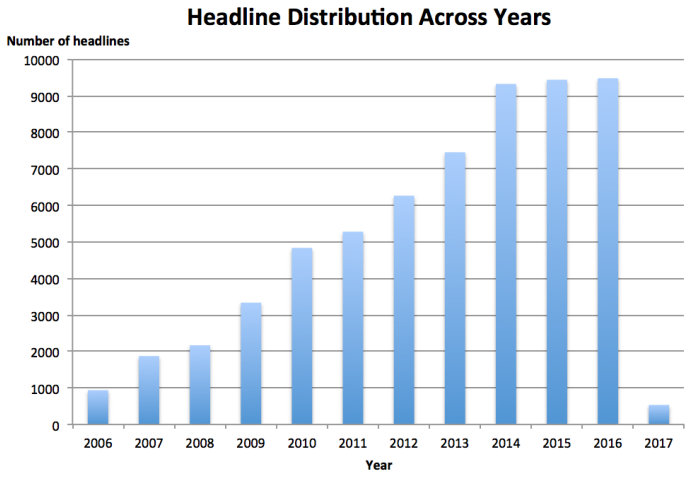


Figure 8: Distribution of number of headlines across time

The headline distribution across months shows an interesting seasonal pattern which makes sense con-

sidering that earnings reports and corporate actions are typically scheduled for the start of each quarter. Figure 9 illustrates this behavior.

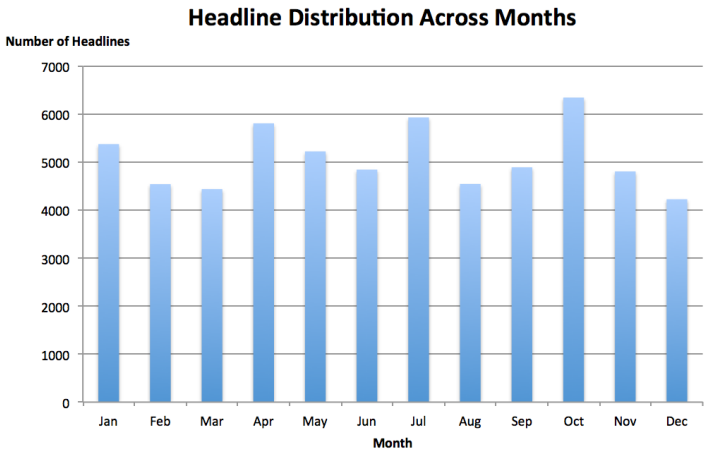


Figure 9: Monthly distribution of headlines

The 60,949 financial news headlines mapped to a total of 546 companies. The 10 most common tickers and their proportions of the entire population of 60,949 headlines are shown in Figure 10. The most commonly featured stock in the dataset of 60,949 headlines is AAPL at 6%, with Microsoft trailing behind at 4% and the rest at approximately 3% or less.

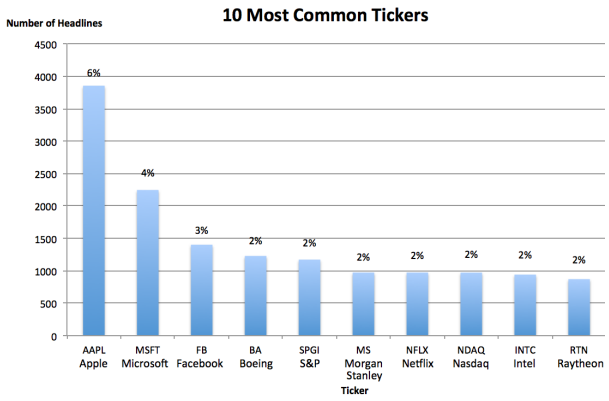


Figure 10: Most common tickers using SVO

We also zoomed in on each ticker to determine if there are any unexpectedly high spikes for any particular trading day. For instance, Figure 11 showing AAPL's cumulative distribution of headlines coverage corresponds well with the trend as seen in the headline distribution across the years for all 60,949 headlines. No obvious spikes are seen and the cumulative number of headlines increased on an exponential curve:

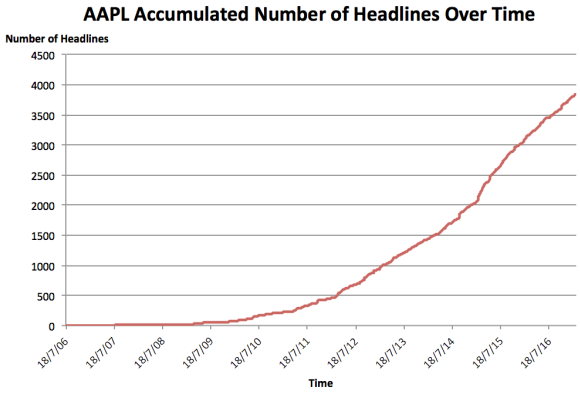


Figure 11: Cumulative number of headlines mentioning Apple over time

We have also attempted to determine if there is a relationship between market capitalization of the companies and their media coverage. In Figure 12, the chart on the left portrays the impression of an exponential relationship that is more clustered around the lower tail, such that the higher the log(market capitalization) of a firm, the higher the headline counts for that company. However, if we remove the three most common tickers - AAPL, MSFT and FB - then we see that there is no distinct relationship between log (market capitalization) and headline counts, as evident in the right chart of Figure 12.

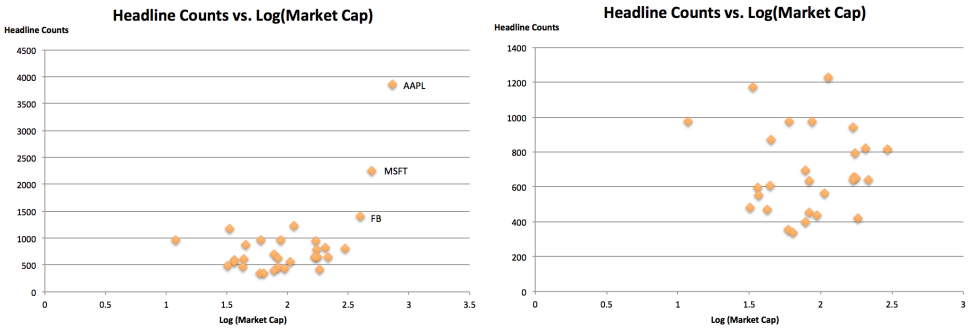


Figure 12: Number of headline counts versus log of market cap. Left: Includes top 30 firms with largest number of mentions. Right: Top three firms removed.

5.3 Observations

Our dataset contains headlines obtained from the period of 2006 to 2016. To test our framework we partition this data set into two sets. One from 2006 to 2014(in-sample) on which we apply our NLP methodology to identify event clusters. Using the second partition(out-sample), which spans 2015-2016, we confirm out of sample event performance and construct portfolios, on the previously identified clusters, using the weighting scheme mentioned in section 4.4.2.
In this section, we focus on observations from the outcomes of the clustering, covering years 2006 to 2014 (in-sample). Specifically, we highlight five events before briefly explaining about some fringe events that we observed as well as the effect of increasing the number of clusters.

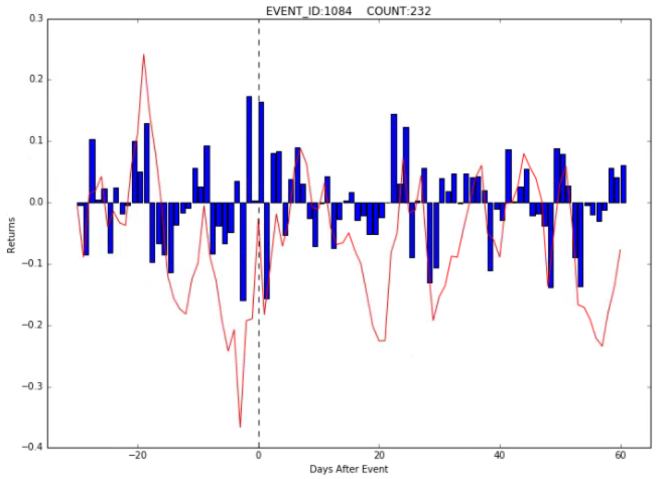
For each event, we compare and analyze the effects of increasing the number of clusters, say k, from 10 to 20, 50 and 100. In this subsection we elaborate on the following events:

- Dividend Declaration
- Oversold Conditions
- Approvals Received by Firms
- Signs Agreement
- Hires Advisor

In general, while increasing the number of clusters from 10 to 100, we notice two main trends:

- The number of returns scenarios increases. This is because as k increases, the amount of occurrences in each cluster shrinks. As a result, each cluster groups together headlines with greater similarities. Hence, higher numbers of clusters are better able to tease out the returns of varying subgroups within each cluster.
- As the number of returns scenarios increases, the Information Ratios accompanying them show greater insights into the events associated with each subgroup.

For each event mapped, our output consists of charts illustrating the cumulative and daily average returns across events, the accompanying sensibility keywords for clusters as well as the Information Ratios for time 1-20 days, 1-40 days and 1-60 days after the event. A typical chart is shown below for an illustration of how we analyze the results. This chart is randomly selected and the accompanying tables show the information ratios as well as keywords that identify the associated event and provide the graph with sensibility:



Verbs	Objects
reports	earnings
report	profit
reported	q1
following	q4
confirmed	q2

Days after Event	Information Ratio
1-20	-0.71
1-40	-0.16
1-60	-0.10

Figure 13: Left: Cumulative & daily returns. Center: Closest verb and object matches to cluster centroid. Right: Information ratio

The interpretation of the left chart in Figure 13 is as follows:

- X-axis: Time
- Y-axis: Return
- Vertical dashed line: Time of event occurrence
- Red line: Cumulative average return over time
- Blue bars: Average daily return at each point of time

From the chart, we observe that the cumulative returns (red line) decrease before reaching a low point around Day 20 after the event. After that, cumulative returns pick up due to positive daily returns (blue bars) approximately on Days 21 and 23. Overall the post-event return for this particular event remains volatile.

We also compute the annualized IR and t-stat for each of the horizons. The IR for an given horizon n is computed using the following formula[29]

$$IR_{1,n} = \sqrt{\frac{252}{n}} \times \frac{mean(\sum_{i=1}^{n} r_{i,j})}{std(\sum_{i=1}^{n} r_{i,j})}$$

Where $r_{i,j}$ is the scaled abnormal return for a stock on event j, i days after the event. We take the average across all the events.

Additionally we compute the $t - stat$ as follows[29]

$$t_{1,n} = IR_{1,n} \times \sqrt{\frac{n}{252}}$$

Our analyses in the following sub-sections are conducted in a similar fashion.
Also, all the event study results shown in the following sub-sections are for market-adjusted returns scaled by their standard deviations.

5.3.1 Event 1: Dividend Declaration

This event is identified by the following sensibility keywords:

Verbs	Objects
declares	dividend
declared	dividends
declare	earnings
considers	income
intends	quarterly

Table 7: Closest verb and object matches to centroid cluster.

The following charts show the daily and cumulative returns over time before and after the declaration of dividends for different number of clusters:

10 (left) & 20 (right) Clusters

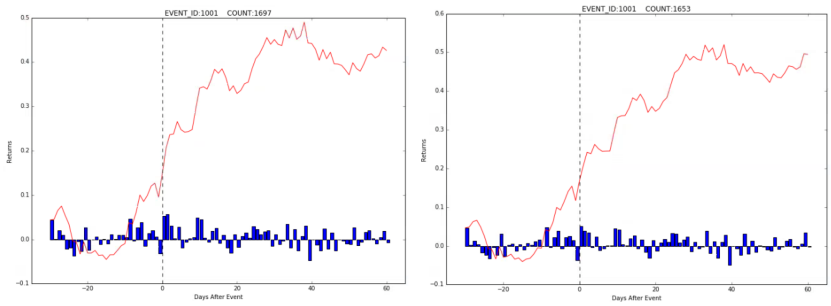


Figure 14: In-sample event study performance. Left: 10 clusters. Right: 20 clusters.

50 Clusters

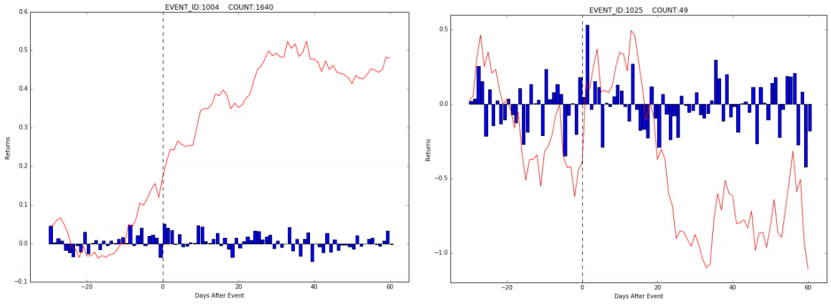


Figure 15: In-sample event study performance using 50 clusters. Left: Event ID 4/100. Right: Event ID 25/100.

100 Clusters

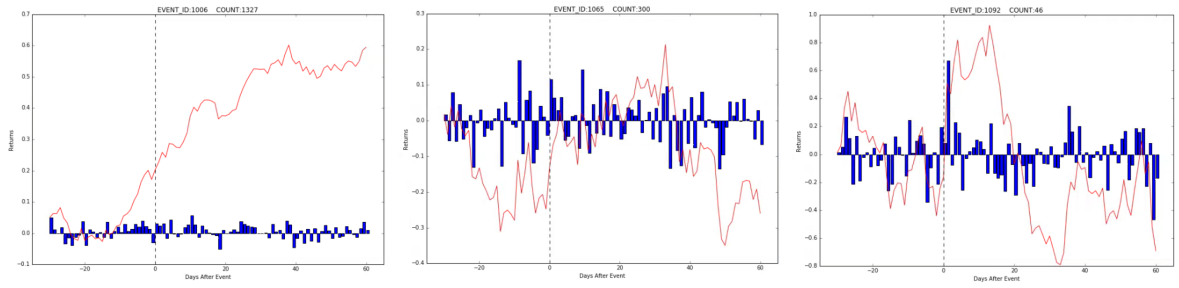


Figure 16: In-sample event study performance using 100 clusters. Left: Event ID 6/100. Center: 65/100. Right: Event ID 92/100.

Information Ratio

Number of Clusters	10	20	50		100		
Counts	1697	1653	1640	49	1327	300	46
Days after Event	Information Ratio (IR)						
1 to 20	0.63	0.62	0.63	0.08	0.60	0.53	0.32
1 to 40	0.75	0.77	0.78	-0.64	0.85	0.13	0.02
1 to 60	0.56	0.66	0.63	-1.32	0.80	-0.13	-0.96

Table 8: Data on IR based on number of clusters and holding period.

Interpretation & Analysis

This event amply describes the effectiveness of our model:

Even though the charts with 10 and 20 clusters (Figure 14) demonstrate positive cumulative returns on average, the IR increases slightly for $k = 20$ as compared to $k = 10$. This shows that as we increase the number of clusters we remove some noisy events from the clusters.

Once we have 50 clusters (Figure 15), we get two clusters corresponding to dividend declaration. One of the clusters with 1640 headlines shows positive cumulative returns (Left chart of Figure 15) whereas another with 49 headlines shows volatile and eventually decreasing cumulative returns (Right chart of Figure 15).

Increasing the number of clusters further to 100 clusters, as shown in Figure 16, allows us to further break down the headlines into three subgroups pertaining to dividend declaration, with positive, neutral and negative post-event performances. Specifically, the three broad sub-groups within the 100 clusters are as follow:

- Sub-group 1 with 1327 events: Clearly positive cumulative returns (Left chart of Figure 16).
- Sub-group 2 with 300 events: Slightly positive yet volatile trend of cumulative returns that turns negative as we go further away from the time of the event (Center chart of Figure 16).
- Sub-group 3 with 46 events: Highly negative reaction to the event but not immediately after the event (Right chart of Figure 16).

5.3.2 Event 2: Oversold Conditions

An event that merits discussion is illustrated in Figure 17.

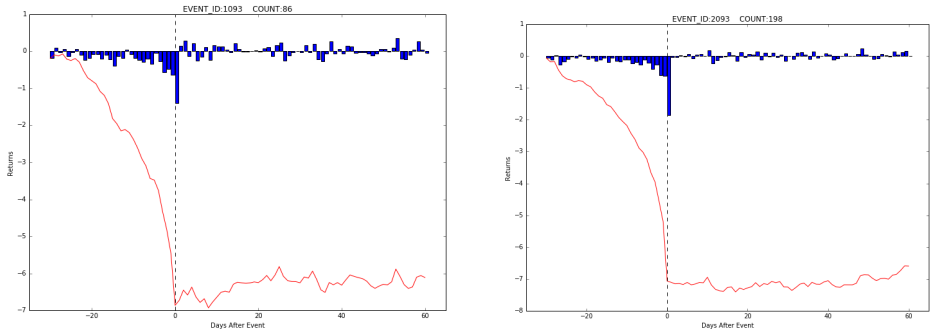


Figure 17: Event study for cluster 93/100. Left: In-sample. Right: Out-of-sample

This figure was produced after observing our in-sample results using 100 clusters. At first glance, the most striking part of the charts is the fact that the event has clearly had some impact on the cumulative returns: After the sharp decline prior to the event, the returns level out and become relatively stable.

The corresponding verbs and objects found for this event are detailed in Table 9. The verbs listed here tend to revolve around the central themes of transformations and expectations. On the other hand, the identified objects tend to be more related to conditions of over-selling and under-valuing. A possible interpretation of this cluster may revolve around news headlines expressing expectations of a company's stock price being lower than its fair price.

Verbs	Objects
is	oversold
becomes	bullish
seems	undervalued
comes	downside
makes	retest

Table 9: Closest verb and object matches to centroid cluster.

We observe that this behavior persists in our out-of-sample tests as illustrated on the right chart of Figure 17. The strong persistence of this behavior in our out-of-sample results provides us with confidence in the ability of the cluster to identify some meaningful news events.

Finally, we can observe how well the identification of events in this cluster translates into portfolio returns. Figure 18 displays the out-of-sample results for the backtest which obtained an IR of 0.44. This plot also illustrates that while the cumulative returns of this backtest are positive, they tend to flatten out over time.

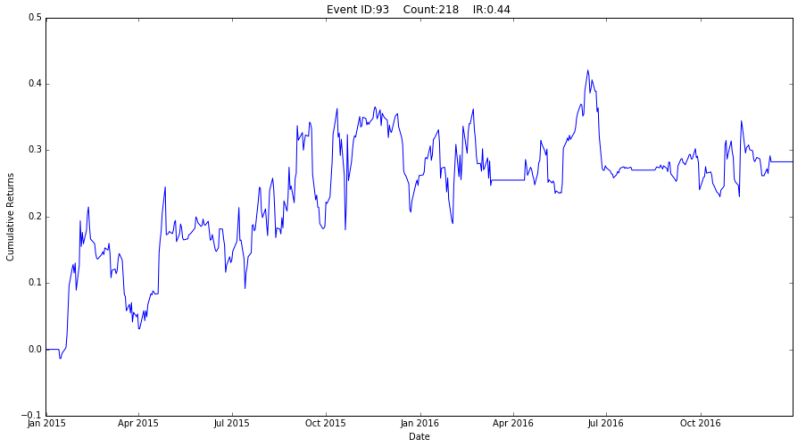


Figure 18: Backtest results for cluster 93/100

As a final point, the event study performance statistics for varying holding periods are presented in Table 10 below:

Horizon	IR	t-stat
1-20	1.72	0.48
1-40	1.09	0.43
1-60	1.20	0.58

(a) In-sample stats

Horizon	IR	t-stat
1-20	-0.67	-0.19
1-40	0.02	0.01
1-60	0.96	0.47

(b) Out-of-sample stats

Table 10: Left: In-sample IR and t-stat (2006-2014). Right: Out-of-sample IR and t-stat (2015-2016)

5.3.3 Event 3: Approval Story

Approvals received by firms appear to be captured by event cluster 36 out of 50, with the corresponding centroid mapping to the following keywords:

Verbs	Objects
receives	approval
sends	approved
obtains	consideration
being	bring
has	reason

Table 11: Closest verb and object matches to centroid cluster

The cluster centroid Verb component seems to be related to the act of obtaining or receiving, while the closest Objects are related to the words *approval* and *approved*. Overall, this event seems to pick up scenarios where positive outcomes are obtained for some approval the firm is seeking. The in-sample and out-of-sample event study results are shown in Figures 19 and 20 respectively:

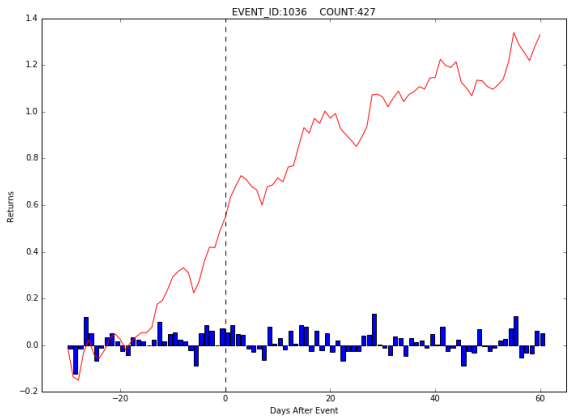


Figure 19: In-sample event study on cluster 36/50

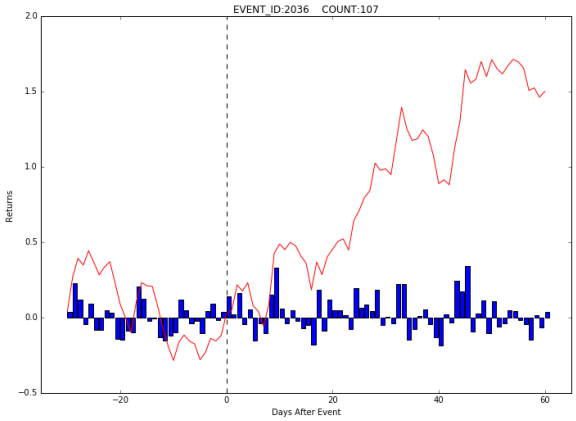


Figure 20: Out-of-sample event study on cluster 36/50

It appears that firms experience higher market-adjusted returns after obtaining approvals. While there is a positive return on the date of the event, it is not as prominent as the Oversold Event (discussed in 5.3.2). The out-of-sample IR results in Table 12 also show consistent positive performances for varying time horizons.

Horizon	IR	t-stat
1-20	1.58	0.45
1-40	1.52	0.60
1-60	1.64	0.80

(a) In-sample stats

Horizon	IR	t-stat
1-20	1.49	0.42
1-40	2.40	0.96
1-60	3.06	1.49

(b) Out-of-sample stats

Table 12: Left: In-sample IR and t-stat (2006-2014). Right: Out-of-sample IR and t-stat (2015-2016)

Additionally, the out-of-sample portfolio construction results in Figure 21 is positive.

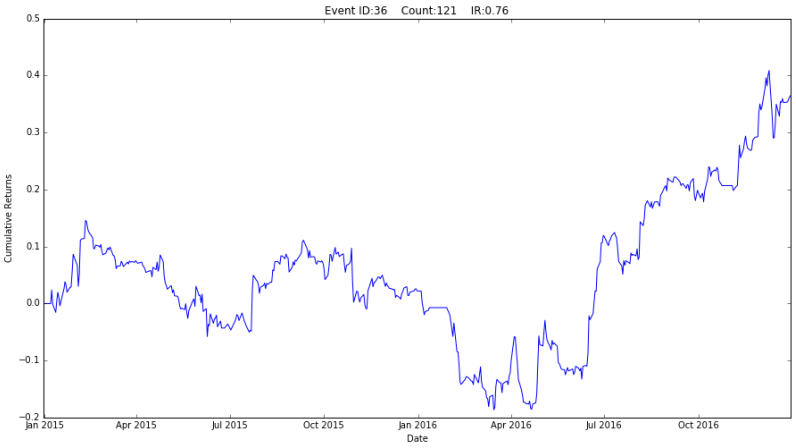


Figure 21: Out-of-sample backtest for cluster 36/50

Overall, the results suggest positive market-adjusted returns after gaining approvals of some sort by the firms. There is, however, still a large amount of noise in the results, which makes it challenging to determine if this event of gaining approval is truly indicative of future returns performance.

5.3.4 Event 4: Signs Agreement

Another interesting event which we identified is related to the signing of agreements, deals or contracts. The centroid for this cluster maps to the following words:

Verbs	Objects
signs	agreement
sign	contract
appears	agreements
suggests	contracts
sees	deal

Table 13: Closest verb and object matches to centroid cluster

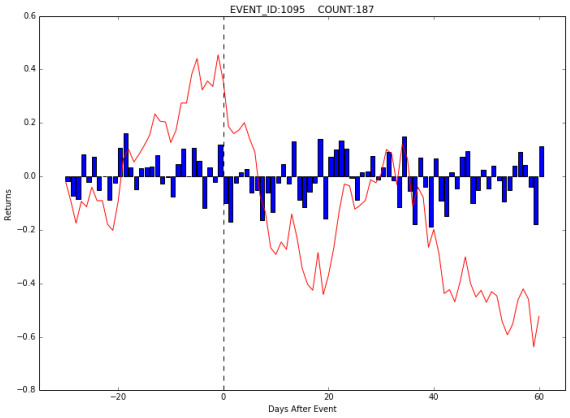


Figure 22: In-sample event study on cluster 95/100

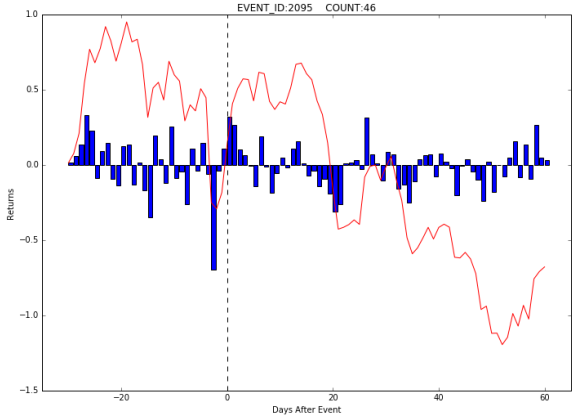


Figure 23: Out-of-sample event study on cluster 95/100

Both the in-sample and out-of-sample event studies results in Figures 22 and 24 show that there are subsequent negative returns after the event. While there is no direct economic rational for why a firm would have negative future returns after signing an agreement or deal, these results could possibly be due to the following reasons:

- The model not being able to capture all the features of the event.
- This result happens to be reflective of our sample data.

Manns et al,. (2012) [45] do mention that " there is no economically consequential market reaction to the disclosure of the details of the acquisition agreement" but do not comment on the longer term impact in aggregate. The statistics in Table 14 below show us the scale of negativity for this event:

Horizon	IR	t-stat
1-20	-2.39	-0.67
1-40	-1.31	-0.52
1-60	-1.77	-0.86

(a) In-sample stats

Horizon	IR	t-stat
1-20	-1.04	-0.29
1-40	-1.38	-0.55
1-60	-1.89	-0.92

(b) Out-of-sample stats

Table 14: Left: In-sample IR and t-stat (2006-2014). Right: Out-of-sample IR and t-stat (2015-2016)

Table 14 shows a slight pattern of reversion from 20 to 40 days after the event followed by further negative returns. The portfolio construction results in Figure 24 show that we have a limited number of events in our out-of-sample period. This is evident in the flat lines in the chart below:

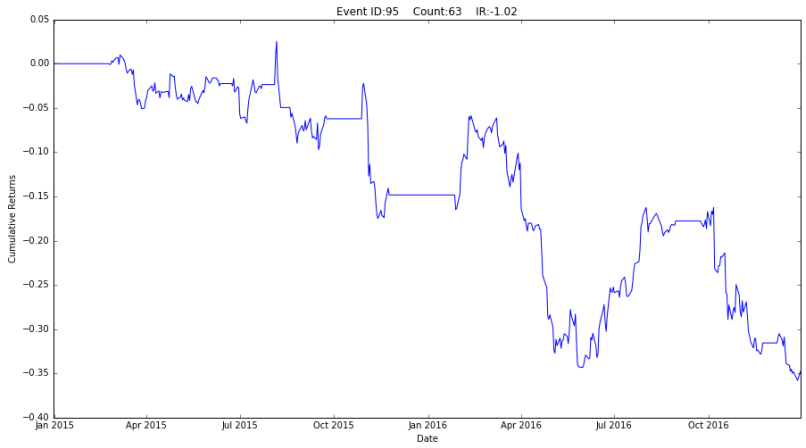


Figure 24: Out-of-sample backtest for cluster 95/100

This demonstrates an advantage of examining the impact of events over a period of time because it offers insights which may not be observable in the Lead/Lag plot framework.

5.3.5 Event 5: Hires Advisor

Examining another cluster that is related to the keywords in Table 15, we identify this event as being related to the act of hiring an advisor to consult with the firm.

Verbs	Objects
hires	advisor
hired	advisors
recruits	advisers
hiring	adviser
joins	investment

Table 15: Closest verb and object matches to centroid cluster.

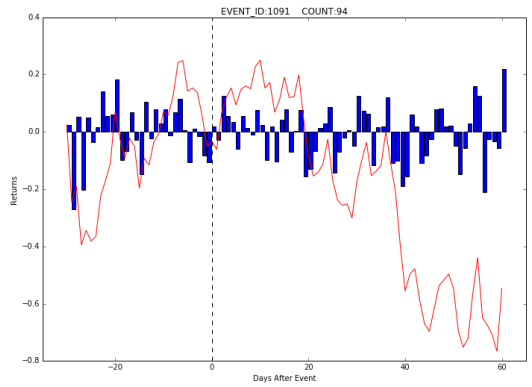


Figure 25: In-sample event study on cluster 91/100

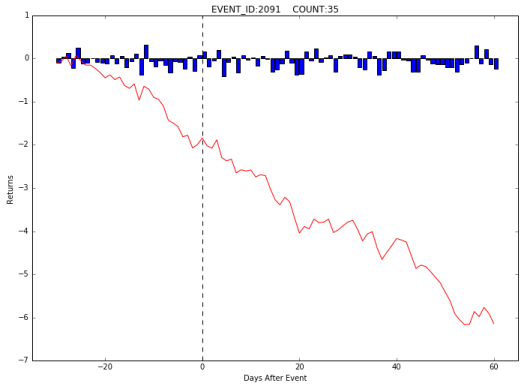


Figure 26: Out-of-sample event study on cluster 91/100

Based on the results of the in-sample and out-of-sample event study analysis displayed in Figures 25 and 26 respectively, we observe that the in-sample results are somewhat negative while the out-of-sample results are sharply negative. Intuitively, there seems to be no obvious reason why hiring an advisor might lead to such a negative performance. Although Mooney and Sibilkov (2012) [50] address the impact of hiring advisors on value, there is no one obvious effect across the board. No further investigation is conducted on the economic link since there might be features of the event that the framework does not capture. 16.

Horizon	IR	t-stat
1-20	-0.22	-0.06
1-40	-1.48	-0.59
1-60	-1.14	-0.56

(a) In-sample stats

Horizon	IR	t-stat
1-20	-7.31	-2.06
1-40	-6.39	-2.55
1-60	-10.85	-5.29

(b) Out-of-sample stats

Table 16: Left: In-sample IR and t-stat (2006-2014). Right: Out-of-sample IR and t-stat (2015-2016)

Based on the statistics for the event study in Table 16, there is a huge negative performance for the out-of-sample results which seems to be statistically significant. However, the in-sample results are much weaker. The difference between the performance might be attributed to either,

- Choosing an out-of-sample period which happens to capture the bulk of negative events
- The low count of the in-sample period (94) and out-of-sample period (35)

Finally, the portfolio construction results are as illustrated in Figure 27. We only show the out-of-sample results since the in-sample frequency is very low and doesn't allow us to create a well diversified portfolio.

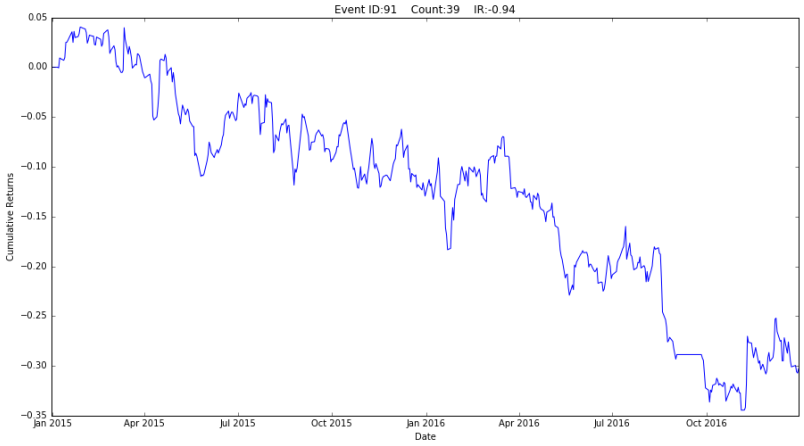


Figure 27: Backtest results for cluster 91/100

The strong negative effect of out of sample event study is seen in the portfolio construction results, albeit to a mitigated effect. From the absence of any long flat lines in the out-of-sample backtest results we can infer that these events were spread out over the entire period.

5.3.6 Inexplicable events

When going through the results, we came across a few cases where the events demonstrate good returns. However, we are unable to map it to any sensible financial phenomenon.

An example is shown here: The centroid for cluster 26 for $k = 50$ maps to the words shown in table 17

Verbs	Objects
is	put
be	same
not	possibly
being	bring
has	reason

Table 17: Closest verb and object matches to centroid cluster.

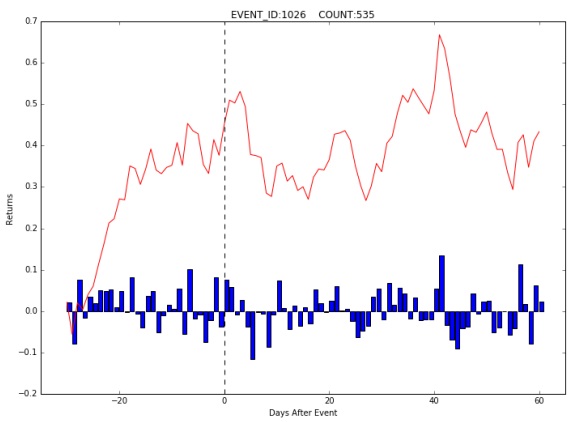


Figure 28: In-sample event study on cluster 26/50

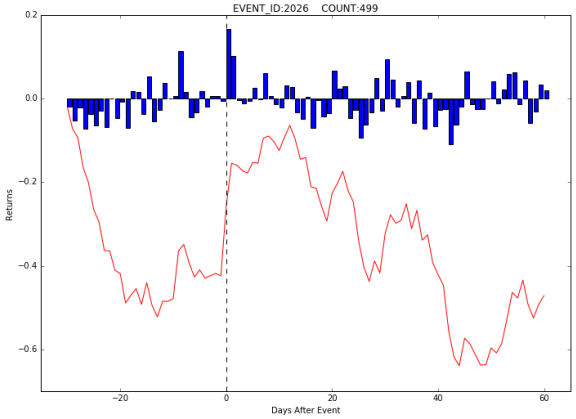


Figure 29: Out-of-sample event study on cluster 26/50

Horizon	IR	t-stat
1-20	-0.30	0.09
1-40	0.20	0.08
1-60	-0.03	-0.02

(a) In-sample stats

Horizon	IR	t-stat
1-20	0.11	0.03
1-40	-0.42	-0.17
1-60	-0.44	-0.22

(b) Out-of-sample stats

Table 18: Left: In-sample IR and t-stat (2006-2014). Right: Out-of-sample IR and t-stat (2015-2016)

Figures 28, 29 and Table 18 show that the returns from 1 to 20 days after the event are negative for the in-sample period. Figure 30 further shows the portfolio construction (backtesting) results with a holding period of 20 days:

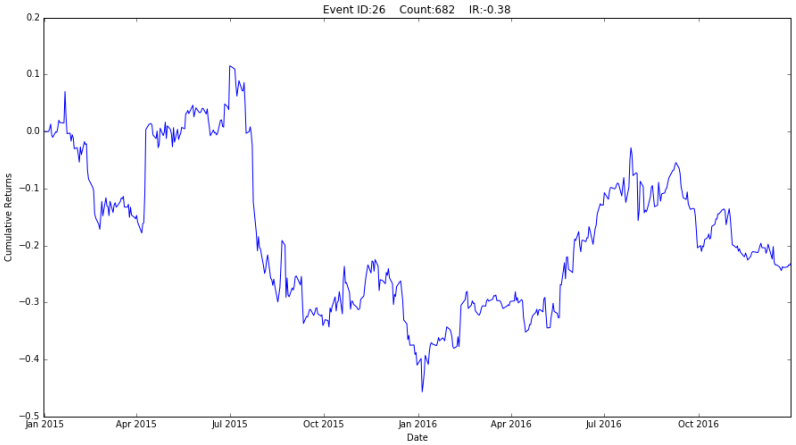


Figure 30: Out-of-sample backtest results for cluster 26/50

Even though the performance is strongly negative which indicates a shorting signal, we are unable to link the keywords found with any financial intuition. Thus, irrespective of performance, we ignore such events.

5.3.7 Relation to number of clusters

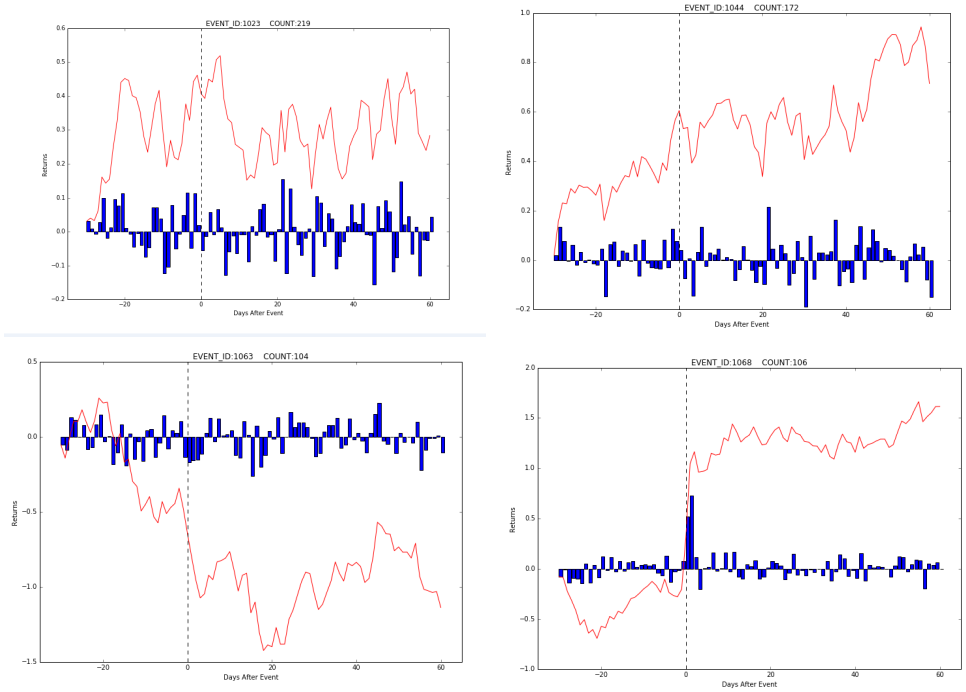


Figure 31: In-sample event study performance for events. Top left: *opening facilities*. Top right: *launching websites*. Bottom left: *employee layoffs*. Bottom right: *exceeding expectations*.

Event Id	Information Ratio			Hit Rate		
	1 to 20	1 to 40	1 to 60	1 to 20	1 to 40	1 to 60
23	-0.78	-0.34	-0.25	0.35	0.45	0.48
44	-0.98	-0.17	0.19	0.55	0.57	0.58
63	-2.62	-0.46	-1.06	0.5	0.57	0.5
68	3.64	2.38	3.01	0.65	0.47	0.58

Table 19: Information ratio and hit rate for events

Event Id	Verb				Object			
23	opens	closes	opening	begins	center	addition	facility	place
44	launches	launched	launch	launching	website	site	web	websites
63	cuts	cut	cutting	puts	jobs	job	employment	expect
68	beats	beat	beating	kicks	expect	expected	likely	coming

Table 20: Event sensibilities

Interpretation & Analysis

To gain a sense of the impact that the number of event clusters had on our results, we examined the behavior of our methodology after varying the number of clusters from 10 to a maximum of 100. Upon doing so, we observed three major results worth noting.

Firstly, by utilizing a larger number of clusters, we are able to more effectively recognize interesting events that are not seen for smaller values of k. Examples include events such as *employee layoffs* (event 63), *exceeding expectations* (event 68), *launching websites* (event 44) and *opening facilities* (event 23). The in-sample event study results for these events are displayed in Figure 31. The ability of the framework to more effectively pick out these unique events can be traced to the frequency with which such events appear in the news. Because these niche events occur at a lower frequency relative to the other major events such as *dividend declaration* and *earnings announcements*, they tend to get merged into clusters with the more common major events for lower values of k. This observation leads us to believe that by increasing the number of clusters, we can more effectively identify unique and infrequent events that may otherwise go overlooked. Table 20 shows the event sensibilities for these clusters. Based on Table 19 we see that the in-sample event studies for *employee layoffs* and *beats expectations* seems intuitive as the performance goes down after *employee layoffs* and increases after *beats expectations* events.

Secondly, we notice that by increasing the number of clusters, we tend to get numerous events corresponding to the same category. For instance, using $k = 100$ we have several events with tags related to earnings. Table 21 shows the breakdown of number of earnings related clusters for k equal to 10, 20, 50 and 100.

k	**Earning events**
10	1
20	2
50	6
100	8

Table 21: Count of earnings related events

Finally, as k increases, the total number of events per cluster decreases. For instance, Figure 32 shows the out-of-sample performance for $k = 100$ and event ID 36. As the event frequency in the out-of-sample data set is low, the out-of-sample studies for such events are not significant.

Therefore, even though increasing the number of clusters is helpful in identifying infrequent events, we cannot make the number of clusters very high as then we lose the significance of the results due to lower frequency of event occurrence.

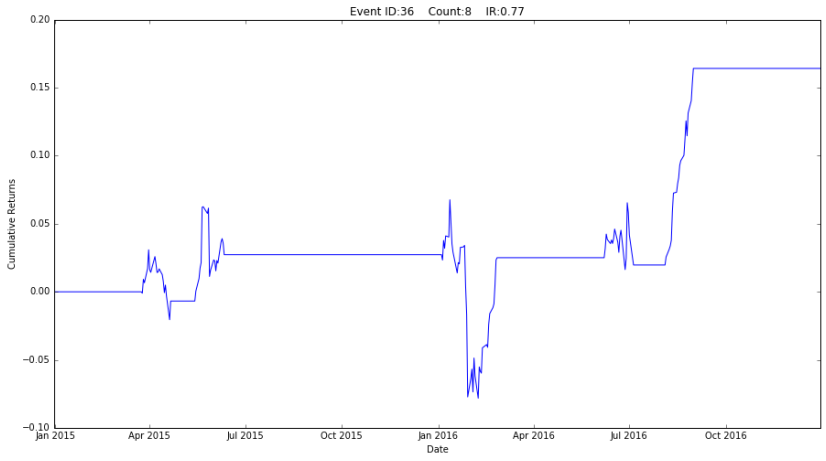


Figure 32: Out-of-sample performance for event 36/100

6 Conclusion

The previous results demonstrate the ability of our framework to operate as a successful prototype for parsing unstructured text data, clustering similar actions within headlines, and converting this information into a coherent portfolio strategy for each cluster. Of the results shown, the most successful event was the *receives approval* with an out-of-sample IR of 0.76. The worst performance was exhibited by the *signs agreement* event with an out-of-sample IR of -1.02.

In the selected event clusters shown, we have demonstrated that our framework is indeed successful in identifying some event clusters with significant implications and that they typically fall into one of three categories. First are events which display significance with regards to the event study and have phrase vectors with clear financial interpretations. The occurrence of these events demonstrates the success our methodology at least as a prototype. Second are events which have sensible phrase vectors but lack event study significance. While their impact for portfolio construction is quite limited, this class of results illustrates the framework's ability to parse headlines and create sensible clusters. Finally is the class of results which do display significance in the event study but have no clear financial interpretation with respect to their phrase vector. While events in this camp do seem to have some importance, the NLP portion of our framework cannot entirely ascertain their significance. Though they cannot be used on their own, one possible solution for future work might be to combine events in this category with other sources of information in an attempt to create a more clear picture.

Another salient feature of our project is the fact that everything used in this project was open source. Data was obtained via web scraping using the Wayback Machine, and all analysis was done using open source packages in Python. Without the significant overhead required by obtaining expensive data sources or proprietary machine learning and natural language processing tools, this project provides a very economical solution by combining many independent sources of publicly available data and open source software.

While these results demonstrate some initial successes, this project certainly contains some inherit limitations and harbors several opportunities for future improvement. In the following sub-sections we shall discuss some of these limitations and discuss some potential points of future research.

6.1 Limitations

Right now our set of text data is limited only to headlines and our focus has been only on companies in the S&P 500. Additionally, our current study is limited to headlines which follow the SVO grammar. Future extensions of the work should relax this restriction to accommodate a greater variety of grammatical structures in headlines. Finally, the gloVe model approach that we use is pre-trained on general English corpus. This causes words which have multiple meaning to have extra features. This word ambiguity can be improved upon by training the word vectors on Financial text only.

6.2 Future Research

This project harbors significant potential for future research. In brief, some of the major areas we would like to touch on in the future include the application of neural-networks to improve our model by better capturing the structural relationship between words in the sentence headline. Additionally,

we would like to be able to further refine the verb-clusters we have obtained to extend the efficacy of our methodology. As a final point, analyzing larger bodies of text would provide us with more information, however a method of differentiating new and novel information against past events is needed for this.

6.2.1 Neural Networks

As mentioned previously, Ding et. al. [18] has shown that using deep learning techniques can be successful in predicting stock movements using news data. A future goal for us would then be to work in a similar vein to train a neural network on unstructured news data to extract the most relevant features of sentences and increase our ability to capture structured relationships within a text. In addition to having increased predictive power, this route would ideally allow us to examine larger bodies of text as well. While much of the essence of the news article should be contained in the headline, one might expect to see enhanced results if full bodies of text are analyzed.

In addition to using neural networks to capture deeper structural relationships within text data, we could also apply a CNN as a layer on top of our collection of verb-clusters to find a more optimal way of combining them to produce a more effective trading signal.

6.2.2 Affinity Propagation

One of the choices that we faced in our project was selecting the appropriate number of clusters. Using an affinity propagation has an advantage of not choosing the parameter. It would be interesting to see what number of clusters the algorithm settles on.

6.2.3 Novelty

One major problem in identifying the actions taken in a given piece of news lies in the novelty of the information being described. A major complication arises in larger bodies of text in situations where news articles reference events that have happened in the past. At present, we subvert this issue by considering only headlines which specify information and actions happening in the present, but identifying past events which are referenced in news articles remains a large obstacle for correctly parsing full bodies of text.

References

[1] Amihud, Yakov. "Illiquidity and stock returns: cross-section and time-series effects." Journal of financial markets 5.1 (2002): 31-56.

[2] Aroomoogan, Kumesh. "How Quant Traders Use Sentiment To Get An Edge On The Market." Forbes. Forbes Magazine, 06 Aug. 2015. Web. 05 Mar. 2017. https://www.forbes.com/sites/kumesharoomoogan/2015/08/06/how-quant-traders-use-sentiment-to-get-an-edge-on-the-market/#75619e144b5d.

[3] Bengio, Yoshua, and Greg Corrado. "Bilbowa: Fast bilingual distributed representations without word alignments." (2015).

[4] Bengio, Yoshua, et al. "A neural probabilistic language model." Journal of machine learning research 3.Feb (2003): 1137-1155.

[5] Buitinck, Lars, et al. "API design for machine learning software: experiences from the scikit-learn project." arXiv preprint arXiv:1309.0238 (2013).

[6] "Calculated (or Derived) based on data from Securities Daily ©2017 Center for Research in Security Prices (CRSP), The University of Chicago Booth School of Business."

[7] Campbell, John Y., Andrew Wen-Chuan Lo, and Archie Craig MacKinlay. The econometrics of financial markets. Princeton University press, 1997.

[8] Carhart, Mark M. "On persistence in mutual fund performance." The Journal of finance 52.1 (1997): 57-82.

[9] Channel, Dividend. "Fastenal Moves Up In Market Cap Rank, Passing NVIDIA." Forbes. Forbes Magazine, 18 Mar. 2015. Web. 06 Mar. 2017. http://www.forbes.com/sites/dividendchannel/2015/03/18/fastenal-moves-up-in-market-cap-rank-passing-nvidia.

[10] Chartbeat, Inc, "textacy", GitHub repository, (2016) https://github.com/chartbeat-labs/textacy

[11] Chip, Brian. "Corning (GLW) Crosses Pivot Point Support at $23.74." N.p., n.d. Web. barchartmarketdata.aws.barchart.com/headlines/story/1474755/corning-glw-crosses-pivot-point-support-at-23-74.

[12] Chowdhury, Gobinda G. "Natural language processing." Annual review of information science and technology 37.1 (2003): 51-89.

[13] Collins, Michael. "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[14] Collobert, Ronan, and Jason Weston. "A unified architecture for natural language processing: Deep neural networks with multitask learning." Proceedings of the 25th international conference on Machine learning. ACM, 2008.

[15] "Corning Announces Investment in Versalume LLC." Versalume, Corning Incorporated, 28 June 2016, versalume.com/blogs/news/corning-announces-investment-in-versalume. Accessed 27 Jan. 2017.

[16] Crowe, Tyler. "3 Stocks to Hold for the Next 50 Years." The Motley Fool. The Motley Fool, 01 Jan. 1970. Web. 06 Mar. 2017. https://www.fool.com/investing/2016/09/16/3-stocks-to-hold-for-the-next-50-years.aspx.

[17] Daniel. J Dufour, "date-extractor", GitHub repository, (2015) github.com/DanielJDufour/date-extractor

[18] Ding, Xiao, Yue Zhang, Ting Liu, and Junwen Duan. "Deep Learning for Event-Driven Stock Prediction." Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015) (n.d.): n. pag. Web. https://www.ijcai.org/Proceedings/15/Papers/329.pdf

[19] Explosion AI, "spaCy", GitHub repository, (2015) https://github.com/explosion/spaCy

[20] Fama, Eugene F., and Kenneth R. French. "Common risk factors in the returns on stocks and bonds." Journal of financial economics 33.1 (1993): 3-56.

[21] Fama, Eugene F., and Kenneth R. French. "The capital asset pricing model: Theory and evidence." The Journal of Economic Perspectives 18.3 (2004): 25-46.

[22] Finkelstein, Lev, et al. "Placing search in context: The concept revisited." Proceedings of the 10th international conference on World Wide Web. ACM, 2001.

[23] French KR. 2007. Data Library, http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

[24] Frey, Brendan J., and Delbert Dueck. "Clustering by passing messages between data points." science 315.5814 (2007): 972-976.

[25] Gantz, John, and David Reinsel. "Extracting Value from Chaos." I D C I V I E W (n.d.): n. pag. Web. https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[26] "GICS." MSCI, MSCI INC, www.msci.com/gics. Accessed 27 Jan. 2017.

[27] Greene, Barbara B., and Gerald M. Rubin. "Automated grammatical tagging of English." (1971).

[28] Grinold, Richard C., and Ronald N. Kahn. "Information Analysis." Active Portfolio Management: A Quantitative Approach for Providing Superior Returns and Controlling Risk. New York: McGraw-Hill, 2000. N. pag. Print.

[29] Grinold, Richard C., and Ronald N. Kahn. "Information analysis." The Journal of Portfolio Management 18.3 (1992): 14-21.

[30] Hartator, "wayback-machine-downloader", GitHub repository, (2016) https://github.com/hartator/wayback-machine-downloader

[31] Herz, Frederick, Lyle Ungar, Jason Eisner, and Walter Labys. "Patent US20030135445 - Stock Market Prediction Using Natural Language Processing." Google Books. N.p., 22 Jan. 2002. Web. 14 Feb. 2017. http://www.google.com/patents/US20030135445

[32] Honnibal, Matthew, and Mark Johnson. "An Improved Non-monotonic Transition System for Dependency Parsing." EMNLP. 2015.

[33] Huang, Eric H., et al. "Improving word representations via global context and multiple word prototypes." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.

[34] "Internet Archive: Wayback Machine." Internet Archive: Wayback Machine, archive.org/web/. Accessed 17 Jan. 2017.

[35] Invest, BNK. "Interesting May 2017 Stock Options for GLW." NASDAQ.com. Nasdaq, 21 Nov. 2016. Web. 15 Feb. 2017. http://www.nasdaq.com/article/interesting-may-2017-stock-options-for-glw-cm712171.

[36] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370. http://nlp.stanford.edu/m̃anning/papers/gibbscrf3.pdf

[37] Knight, Will. "Will AI-Powered Hedge Funds Outsmart the Market?" MIT Technology Review. MIT Technology Review, 04 Feb. 2016. Web. 14 Feb. 2017. https://www.technologyreview.com/s/600695/will-ai-powered-hedge-funds-outsmart-the-market/.

[38] Kothari, S. P., and Jerold B. Warner. "The Econometrics of Event Studies." The Econometrics of Event Studies by S.P. Kothari, Jerold B. Warner :: SSRN. N.p., 20 Oct. 2004. Web. 15 Feb. 2017. https://papers.ssrn.com/sol3/papers2.cfm?abstract_id=608601.

[39] Kucera, H., and W. Francis. "A standard corpus of present-day edited American English, for use with digital computers (revised and amplified from 1967 version)." (1979).

[40] Kupiec, Julian. "Robust part-of-speech tagging using a hidden Markov model." Computer Speech & Language 6.3 (1992): 225-242.

[41] Lintner, John. "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets." The review of economics and statistics (1965): 13-37.

[42] Liu, Ting, Andrew W. Moore, and Alexander Gray. "New algorithms for efficient high-dimensional nonparametric classification." Journal of Machine Learning Research 7.Jun (2006): 1135-1158.

[43] Luong, Thang, Richard Socher, and Christopher D. Manning. "Better word representations with recursive neural networks for morphology." CoNLL. 2013.

[44] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 14. 1967.

[45] Manns, Jeffrey, and Robert Anderson IV. "The merger agreement myth." Cornell L. Rev. 98 (2012): 1143.

[46] Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank." Computational linguistics 19.2 (1993): 313-330.

[47] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[48] Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations." Hlt-naacl. Vol. 13. 2013.

[49] Miller, George A., and Walter G. Charles. "Contextual correlates of semantic similarity." Language and cognitive processes 6.1 (1991): 1-28.

[50] Mooney, Tim, and Valeriy Sibilkov. “Sealing the Deal: Is an Advisor’s Completion Expertise in Mergers & Acquisitions Value-Destroying?.” (2012).

[51] Nadeau, David, and Satoshi Sekine. “A survey of named entity recognition and classification.” Lingvisticae Investigationes 30.1 (2007): 3-26.

[52] Omohundro, Stephen M. “Five Balltree Construction Algorithms.” Berkeley: International Computer Science Institute, 1989.

[53] Pástor, Ľuboš, and Robert F. Stambaugh. “Liquidity risk and expected stock returns.” Journal of Political economy 111.3 (2003): 642-685.

[54] Pedregosa, Fabian, et al. “Scikit-learn: Machine learning in Python.” Journal of Machine Learning Research 12.Oct (2011): 2825-2830.

[55] Pennington, Jeffrey. ”GloVe: Global Vectors for Word Representation.” GloVe: Global Vectors for Word Representation. N.p., n.d. Web. http://nlp.stanford.edu/projects/glove/.

[56] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. “Glove: Global Vectors for Word Representation.” EMNLP. Vol. 14. 2014.

[57] PR Newswire: News Distribution, Targeting and Monitoring. PR Newswire Association LLC, n.d. Web. 27 Jan. 2017.

[58] Ratnaparkhi, Adwait. “A maximum entropy model for part-of-speech tagging.” Proceedings of the conference on empirical methods in natural language processing. Vol. 1. 1996.

[59] Rubenstein, Herbert, and John B. Goodenough. “Contextual correlates of synonymy.” Communications of the ACM 8.10 (1965): 627-633.

[60] Ruder, Sebastian. “An Overview of Word Embeddings and Their Connection to Distributional Semantic Models.” AYLIEN, 17 Jan. 2017, blog.aylien.com/overview-word-embeddings-history-word2vec-cbow-glove/. Accessed 27 Jan. 2017.

[61] Rusu, Delia, et al. “Triplet extraction from sentences.” Proceedings of the 10th International Multiconference Information Society-IS. 2007.

[62] Services, Wharton Research Data. “Wharton Research Data Services.” Wharton Research Data Services. N.p., n.d. Web. 06 Mar. 2017. https://wrds-web.wharton.upenn.edu/wrds/.

[63] Sharpe, William F. “Capital asset prices: A theory of market equilibrium under conditions of risk.” The journal of finance 19.3 (1964): 425-442.

[64] Smith, Rich. “When Will Corning Inc. Split Its Stock Again?” The Motley Fool. The Motley Fool, 01 Jan. 1970. Web. 06 Mar. 2017. https://www.fool.com/investing/2016/09/19/when-will-corning-inc-split-its-stock-again.aspx.

[65] “Software Stanford Log-linear Part-Of-Speech Tagger.” The Stanford Natural Language Processing Group. N.p., n.d. Web. http://nlp.stanford.edu/software/tagger.shtml.

[66] Standard and Poor’s Dataset. (2017). GICS Industry Classification data. 27 January 2017. Retrieved from Wharton Research Data Service.

[67] Standard and Poor’s Dataset. (2017). S&P 500 Membership data. 27 January 2017. Retrieved from Wharton Research Data Service.

[68] "Stanford Named Entity Tagger." Stanford Named Entity Tagger. N.p., n.d. Web. 07 Mar. 2017. http://nlp.stanford.edu:8080/ner/.

[69] "Stanford Parser." Stanford Parser. N.p., n.d. Web. http://nlp.stanford.edu:8080/parser/index.jsp.

[70] Steinhaus, H. (1956) "Sur la division des corps matériels en parties" Bulletin de l'academie' polonaise des sciences Cl. III — Vol. IV, No. 12, 1956

[71] "Swing Trading Watch List: URI, WPX, ETN, GLW, ZTS." TalkMarkets. N.p., n.d. Web. 15 Feb. 2017. http://www.talkmarkets.com/content/stocks--equities/swing-trading-watch-list-uri-wpx-etn-glw-zts?post=107333.

[72] Tjong Kim Sang, Erik F., and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, 2003.

[73] Toutanova, Kristina, et al. "Feature-rich part-of-speech tagging with a cyclic dependency network." Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, 2003.

[74] Turner, Matt. "A Giant Hedge Fund Used Artificial Intelligence to Analyze Fed Minutes — Here's What It Found." Business Insider. Business Insider, 18 May 2016. Web. 11 Feb. 2017.

[75] Unstructured Data: Friend Or Foe? (n.d.): n. pag. DataGravity, Inc. Web. https://datagravity.com/sites/default/files/resource-files/DG_wp_unstructureddatafriendorfoe_20140721.pdf.

[76] "Wayback Machine." Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc. 22 July 2004. Web. 10 Aug. 2004. https://en.wikipedia.org/wiki/Wayback_Machine

[77] Xie, Boyi, Rebecca J. Passonneau, and Leon Wu. "Semantic Frames to Predict Stock Price Movement." Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, n.d. Web.

7 Appendix

Cluster No.	Verb 1	Verb 2	Verb 3	Verb 4	Verb 5	Object 1	Object 2	Object 3	Object 4	Object 5
1	should	still	could	would	will	to	can	you	make	that
2	declares	declared	declare	considers	intends	dividend	dividends	earnings	income	quarterly
3	receives	puts	earns	awarded	buys	million	billion	100,000	500,000	200,000
4	to	will	can	would	be	expect	addition	bringing	same	possibly
5	reports	report	reported	following	confirmed	results	result	last	following	next
6	puts	considers	brings	pushes	sees	bringing	addition	possibly	expect	puts
7	comes	goes	puts	takes	makes	possibly	same	expect	bringing	addition
8	reports	report	reported	following	confirmed	quarter	earnings	profit	revenue	revenues
9	puts	brings	turns	suggests	sees	that	well	same	it	but
10	launches	announces	unveils	introduces	announce	addition	same	possibly	bringing	bring

Table 22: Similar Verbs and Objects for 10 Clusters. The columns show in similarity the center of each cluster to the closest verb or object in our vocabulary in descending order

Cluster No.	Verb 1	Verb 2	Verb 3	Verb 4	Verb 5	Object 1	Object 2	Object 3	Object 4	Object 5
1	puts	brings	sees	turns	suggests	that	well	it	same	can
2	declares	declared	declare	considers	intends	dividend	dividends	earnings	income	quarterly
3	is	has	be	being	still	possibly	expect	likely	same	put
4	named	called	names	name	chosen	one	same	another	put	well
5	leads	lead	leading	puts	turns	stocks	tech	market	gains	investment
6	reports	report	reported	suggests	expect	earnings	quarter	profit	revenue	revenues
7	reports	report	reported	following	claims	quarter	earnings	profit	fourth	revenue
8	to	will	can	would	make	addition	bringing	expect	same	possibly
9	receives	puts	considers	buys	brings	million	billion	100,000	500,000	dollars
10	can	will	be	would	could	expect	same	possibly	likely	bringing
11	puts	considers	demonstrates	pushes	suggests	addition	bringing	consider	possibly	expect
12	reports	report	reported	following	confirmed	results	result	following	last	similar
13	wants	expect	hoping	continues	supposed	to	can	you	make	them
14	receives	wins	earns	received	awarded	contract	contracts	deal	consideration	addition
15	puts	takes	turns	goes	comes	same	possibly	put	bring	addition
16	launches	announces	unveils	announce	announced	addition	same	possibly	bringing	bring
17	becomes	become	becoming	turns	became	component	components	example	addition	means
18	considers	puts	buys	chooses	intends	nallatech	boardvantage	254.50	knapp	arinso
19	puts	turns	brings	goes	comes	expect	puts	possibly	bringing	nallatech
20	is	be	will	not	would	to	can	that	you	it

Table 23: Similar Verbs and Objects for 20 Clusters. The columns show in similarity the center of each cluster to the closest verb or object in our vocabulary in descending order

Cluster No.	Verb 1	Verb 2	Verb 3	Verb 4	Verb 5	Object 1	Object 2	Object 3	Object 4	Object 5
1	tops	top	topped	sets	dresses	102.81	121.55	254.50	108.88	93.39
2	puts	brings	turns	sees	takes	that	well	can	it	one
3	announces	announced	announce	launches	unveils	addition	possibly	expect	same	likely
4	puts	takes	turns	goes	brings	possibly	same	bringing	put	bring
5	declares	declared	declare	considers	intends	dividend	dividends	earnings	income	quarterly
6	to	will	can	would	make	that	business	well	can	it
7	wants	expect	hoping	supposed	continues	to	can	you	make	be
8	launches	launch	launched	unveils	launching	addition	bringing	example	same	bring
9	puts	sees	suggests	brings	turns	sales	revenue	increase	profit	earnings
10	to	buy	can	get	want	puts	bringing	addition	possibly	bring
11	donates	donating	donated	donate	buys	100,000	200,000	500,000	50,000	million
12	reports	report	reported	following	confirmed	year	month	last	december	january
13	gets	hits	takes	goes	puts	spot	possibly	put	same	likely
14	can	will	be	would	could	possibly	expect	same	bringing	likely
15	becomes	become	becoming	turns	became	component	components	example	addition	means
16	to	acquire	can	will	would	bringing	addition	puts	utilizing	baxter
17	is	be	scheduled	will	not	to	report	that	be	can
18	extends	extending	extend	expands	defines	bring	expect	possibly	addition	consider
19	named	called	names	name	chosen	one	same	another	put	well
20	receives	puts	considers	brings	gives	million	billion	100,000	dollars	500,000
21	reports	report	reported	following	confirmed	results	result	similar	following	last
22	has	been	have	had	still	possibly	same	likely	expect	put
23	reports	report	reported	following	claims	quarter	fourth	half	third	earnings
24	leads	lead	leading	puts	turns	tech	technology	market	industry	sector
25	pushes	readies	downplays	puts	shuns	possibly	bringing	expect	same	addition
26	incorporated	declares	considers	declared	intends	dividend	quarterly	earnings	cash	dividends
27	is	be	not	being	has	put	same	possibly	bring	reason
28	bring	expect	bringing	might	put	possibly	expect	likely	bringing	same
29	shares	share	buys	stocks	owns	dropping	expect	likely	coming	possibly
30	puts	brings	turns	sees	suggests	nallatech	boardvantage	254.50	feedhenry	arinso
31	wins	awarded	won	winning	earns	contract	contracts	deal	agreement	basis
32	reports	report	reported	following	suggests	eps	q4	q3	q2	q1
33	raises	raise	raising	lowers	puts	expect	expected	likely	same	possibly
34	to	will	can	would	be	earnings	quarter	results	profit	revenue
35	reports	report	reported	following	confirmed	profit	earnings	profits	revenue	gains
36	are	be	is	not	being	stocks	markets	stock	commodities	investing
37	receives	sends	obtains	received	receive	approval	approved	consideration	regarding	prior
38	puts	turns	sees	bested	hoping	estimates	estimate	earnings	forecasts	q1
39	selects	initiates	considers	chooses	completes	addition	consider	possibly	bringing	present
40	to	will	can	would	make	bringing	addition	expect	possibly	bring
41	leads	lead	puts	turns	brings	stocks	stock	markets	market	trading
42	puts	brings	turns	bringing	takes	iphone	ipad	android	smartphone	apps
43	brings	adds	demonstrates	puts	incorporates	addition	bringing	possibly	bring	consider
44	celebrates	celebrating	celebrate	embraces	commemorates	years	year	month	coming	past
45	is	be	will	not	would	to	can	you	them	up
46	reiterates	considers	suggests	puts	believes	rating	ratings	rated	reviews	stars
47	puts	sees	turns	considers	brings	that	deal	financial	company	interest
48	receives	sends	received	earns	obtains	award	awards	awarded	winners	winner
49	to	will	can	would	make	million	billion	100,000	500,000	200,000
50	reports	report	reported	following	confirmed	quarter	earnings	profit	revenue	revenues

Table 24: Similar Verbs and Objects for 50 Clusters. The columns show in similarity the center of each cluster to the closest verb or object in our vocabulary in descending order

Cluster No.	Verb 1	Verb 2	Verb 3	Verb 4	Verb 5	Object 1	Object 2	Object 3	Object 4	Object 5
1	puts	brings	launches	introduces	bringing	service	that	can	it	be
2	awarded	award	awards	received	honors	million	billion	100,000	500,000	200,000
3	puts	sees	considers	turns	suggests	deal	offer	purchase	sell	offering
4	be	not	would	will	can	same	put	possibly	addition	expect
5	reports	report	reported	following	claims	quarter	fourth	half	third	earnings
6	announces	announced	announce	launches	unveils	consideration	addition	consider	possibly	prior
7	declares	declared	declare	considers	intends	dividend	dividends	earnings	quarterly	income
8	seen	seeing	might	see	expect	posting	possibly	giving	consider	putting
9	reports	report	reported	following	claims	results	result	similar	following	search
10	becomes	appears	become	is	seems	stock	stocks	shares	trading	market
11	approaches	suggests	aims	examining	considers	target	targets	targeting	aiming	likely
12	receives	sends	received	obtains	earns	award	awards	awarded	winners	winner
13	will	can	would	could	should	possibly	expect	bringing	puts	addition
14	to	will	can	would	make	that	well	business	can	it
15	awarded	wins	receives	earns	received	contract	contracts	agreement	deal	lease
16	launches	launched	launch	launching	unveils	utilizing	addition	example	capabilities	enable
17	set	sets	setting	put	comes	to	can	you	make	be
18	celebrates	celebrating	celebrate	embraces	commemorates	years	year	month	coming	past
19	to	will	can	would	be	conference	meeting	meetings	conferences	call
20	takes	puts	turns	goes	gives	spot	possibly	put	bring	again
21	to	will	can	would	make	puts	expect	possibly	bringing	addition
22	puts	sees	turns	brings	considers	that	government	regarding	likely	must
23	becomes	become	becoming	turns	became	component	components	example	addition	means
24	opens	closes	opening	begins	puts	center	addition	facility	place	outside
25	has	been	have	had	still	possibly	same	likely	expect	put
26	puts	turns	sees	brings	suggests	market	sales	business	company	that
27	expected	expect	expects	anticipated	expecting	to	can	you	that	it
28	is	be	scheduled	will	not	to	report	that	be	can
29	tops	top	topped	sets	dresses	puts	possibly	expect	102.81	121.55
30	to	will	can	make	help	addition	bringing	possibly	bring	same
31	makes	put	comes	puts	bring	possibly	bringing	same	expect	bring
32	announces	announce	announced	launches	unveils	addition	same	bring	bringing	well
33	hits	hit	hitting	strikes	puts	high	record	low	past	same
34	elects	appoints	nominates	chooses	selects	bennett	sullivan	knapp	miller	webb
35	reports	report	reported	following	suggests	earnings	quarter	profit	revenue	revenues
36	shares	share	buys	stocks	owns	drop	expect	likely	dropping	35
37	leads	lead	leading	puts	turns	tech	stocks	technology	market	industry
38	to	can	would	will	cut	jobs	job	employment	salaries	salary
39	reports	report	reported	following	suggests	eps	q4	q3	q2	q1
40	is	be	not	being	still	possibly	same	bringing	put	bring
41	wins	win	won	winning	loses	same	possibly	likely	reason	deal
42	to	can	will	would	make	same	possibly	bring	bringing	expect
43	reports	report	reported	following	claims	profit	increase	revenue	profits	expected
44	lifts	lift	lifted	pushes	pulls	puts	expect	bringing	wraps	ahead
45	launches	launched	launch	launching	unveils	website	site	web	websites	same
46	puts	sees	brings	suggests	turns	nallatech	254.50	boardvantage	feedhenry	arinso
47	reiterates	considers	suggests	puts	believes	rating	ratings	rated	reviews	stars
48	reports	report	reported	following	confirmed	results	quarter	year	last	result
49	to	will	can	would	make	million	billion	mln	dollars	100,000
50	is	be	will	not	would	to	can	you	them	up

Table 25: Similar Verbs and Objects for 100 Clusters, Part 1. The columns show in similarity the center of each cluster to the closest verb or object in our vocabulary in descending order

Cluster No.	Verb 1	Verb 2	Verb 3	Verb 4	Verb 5	Object 1	Object 2	Object 3	Object 4	Object 5
51	are	be	have	're	not	stocks	markets	stock	commodities	investing
52	donates	donating	donated	buys	sells	100,000	50,000	200,000	500,000	300,000
53	announces	announced	announce	launches	unveils	results	increase	result	expected	increased
54	to	acquire	can	will	would	bringing	addition	puts	utilizing	baxter
55	selects	chooses	selected	determines	identifies	addition	bringing	utilizing	puts	possibly
56	launches	launched	launch	launching	unveils	bringing	addition	bring	possibly	puts
57	unveils	introduces	launches	unveiled	announces	addition	same	possibly	bringing	bring
58	reports	report	reported	following	claims	traffic	september	december	january	october
59	receives	sends	received	receive	obtains	approval	regarding	consideration	prior	approved
60	to	will	can	would	not	quarter	earnings	results	profit	revenue
61	extends	extending	extend	expands	defines	bring	expect	possibly	addition	consider
62	gets	getting	takes	get	goes	possibly	expect	bringing	bring	puts
63	leads	lead	leading	turns	suggests	expect	gains	likely	possibly	bringing
64	cuts	cut	cutting	puts	breaks	jobs	job	employment	expect	working
65	named	called	names	name	chosen	one	another	same	put	well
66	declares	declared	declare	considers	intends	dividend	cash	dividends	earnings	income
67	wants	wanted	supposed	puts	put	to	can	you	be	make
68	downgraded	downgrades	downgrade	downgrading	upgraded	ago	possibly	gone	similar	yet
69	beats	beat	beating	kicks	puts	expect	expected	likely	coming	sees
70	want	would	could	do	can	same	put	well	might	but
71	announces	launches	introduces	announce	announced	date	dates	release	last	time
72	launches	introduces	announces	unveils	launched	program	programs	programme	education	plan
73	may	can	be	will	not	same	going	likely	bring	well
74	puts	turns	sees	suggests	goes	gas	oil	fuel	supply	energy
75	considers	intends	agrees	puts	pushes	consider	addition	possibly	likely	regarding
76	initiates	initiate	initiated	completes	selects	coverage	covering	benefit	addition	expect
77	puts	sees	turns	suggests	pushes	sales	market	profit	increase	revenue
78	demonstrates	brings	recognizes	incorporates	emphasizes	bringing	addition	bring	possibly	consider
79	offers	offering	provides	offer	provide	addition	consider	possibly	possible	bring
80	begins	starts	begin	continues	beginning	addition	bringing	possibly	consider	same
81	enhances	boosts	improves	accelerates	reduces	bringing	addition	expect	consider	bring
82	puts	brings	suggests	turns	sees	software	same	that	mobile	computer
83	leads	lead	leading	turns	puts	sector	market	industry	markets	investment
84	adds	adding	brings	gives	puts	addition	bringing	possibly	expect	puts
85	reports	report	reported	following	confirmed	earnings	profit	q1	q4	q2
86	puts	turns	sees	pushes	goes	level	higher	levels	same	well
87	sells	buys	sell	selling	owns	mln	bln	stake	shares	euros
88	puts	turns	sees	pushes	suggests	put	possibly	same	bringing	bring
89	announces	puts	launches	introduces	brings	billion	million	trillion	dollars	billions
90	shares	share	buys	owns	stocks	creeping	inching	slipping	fear	pushing
91	achieves	achieve	attains	delivers	demonstrates	addition	expect	result	possibly	likely
92	hires	hired	recruits	hiring	joins	advisor	advisors	advisers	adviser	investment
93	incorporated	declares	considers	declared	introduces	dividend	quarterly	earnings	cash	dividends
94	is	becomes	seems	comes	makes	oversold	bullish	undervalued	downside	retest
95	to	buy	can	get	want	puts	bringing	addition	possibly	bring
96	signs	sign	appears	suggests	sees	agreement	contract	agreements	contracts	deal
97	continues	hopes	wants	intends	hoping	to	can	make	you	them
98	puts	turns	sees	pushes	suggests	earnings	profit	profits	gains	expected
99	receives	puts	gives	brings	considers	million	billion	100,000	millions	dollars
100	is	still	comes	be	seems	buy	purchase	sell	buying	cheapest

Table 26: Similar Verbs and Objects for 100 Clusters, Part 2. The columns show in similarity the center of each cluster to the closest verb or object in our vocabulary in descending order