Universidad de San Andrés

Departamento de Economía

Maestría en Economía

Unsupervised Extraction of Market Moving Events with Neural Attention

Luciano Del Corro

DNI 29.150.471

Mentor: Daniel Heymann

Buenos Aires

Abril de 2020

Tesis de Maestría en Economía de
Luciano Del Corro

"Extracción no supervisada de eventos que mueven mercados con atención neuronal"

Resumen:En este trabajo presentamos un método para identificar eventos relevantes asociados con movimientos en el precio de las acciones sin datos anotados manualmente. Entrenamos una red neuronal basada en el mecanismo de atención, que dada una serie de títulos periodísticos, en un intervalo específico, predice el movimiento de precios con tres resultados posibles (baja, sin cambios, sube). El mecanismo de atención actúa de selector de inputs calculando un score normalizado de importancia para el embedding correspondiente a cada título. El promedio ponderado de los embeddings es utilizado para predecir el movimiento de la acción. Presentamos un análisis para entender si luego de que la red fuese entrenada, el mecanismo de atención no normalizado es capaz de producir un ranking global de eventos financieros. El ranking debería otorgar mayor importancia a los eventos financieramente más relevantes. En este estudio inicial utilizamos las categorías de las noticias como proxy de relevancia. Noticias que pertenezcan a categorías más relevantes deberían obtener un score de importancia relativamente más alto. Nuestros experimentos en los cuatro índices más relevantes de Estados Unidos indican que efectivamente el score calculado por el mecanismo de atención es más alto para aquellas categorías que intuitivamente resultan mas relevantes para explicar cambios de precios.

"Unsupervised Extraction of Market MovingEvents with Neural Attention"

Abstract: We present a method to identify relevant events associated with stock price movementswithout manually labeled data. We train an attention-based neural network, which given aset of news headlines for a given time frame, predicts the price movement of a given stockindex (i.e., DOWN, STAY, UP). An attention layer acts as an input selector; it computes anormalized weight for each headline embedding. The weighted average of the embeddingsis used to predict the price movement. We present an analysis to understand if, after thenetwork has been trained, the attention layer is capable of generating a global ranking ofnews events through its unnormalized weights. The ranking should be able to rank relevantfinancial events higher. In this initial study we use news categories as a proxy for relevance:news belonging to more relevant categories should be ranked higher. Our experiments onfour indices suggest that there is an indication that the weights indeed skew the global set ofevents towards those categories that are more relevant to explain the price change; this effectreflects the performance of the network on stock prediction.

Keywords: event extraction, neural attention, stock prediction, event ranking, unsupervised
Códigos JEL: C63, C81, C99, G14, G17:

Acknowledgements

I thank Johannes Hoffart for all his help and support. I would also like to thank the Universidad de San Andrés and especially the Department of Economics for the scholarship granted to pursue this master, which opened me so many doors and allowed me to get in touch with wonderful people. I would also like to thank Walter Sosa Escudero and Daniel Heymann all the motivation to pursue my career. Last but not least, I thank my wife Leticia for all her infinite and loving support, and the most amazing persons in the world: Dante and Eloy.

Universidad de
SanAndrés

Abstract

We present a method to identify relevant events associated with stock price movements without manually labeled data. We train an attention-based neural network, which given a set of news headlines for a given time frame, predicts the price movement of a given stock index (i.e., DOWN, STAY, UP). An attention layer acts as an input selector; it computes a normalized weight for each headline embedding. The weighted average of the embeddings is used to predict the price movement. We present an analysis to understand if, after the network has been trained, the attention layer is capable of generating a global ranking of news events through its unnormalized weights. The ranking should be able to rank relevant financial events higher. In this initial study we use news categories as a proxy for relevance: news belonging to more relevant categories should be ranked higher. Our experiments on four indices suggest that there is an indication that the weights indeed skew the global set of events towards those categories that are more relevant to explain the price change; this effect reflects the performance of the network on stock prediction.

Universidad de
SanAndrés

Chapter 1

Introduction

Anticipating stock price movements is one of the critical challenges in financial analysis. A specific event (or combination of events) might have a significant and even disruptive impact in the markets. Understanding which events may generate stock price movements is crucial for the financial industry.

In this paper, we present a method to identify relevant financial events without manually labeled data. We understand relevant events as those that contribute to stock price movements. We do not directly predict the importance of the events but use a method that automatically selects the more relevant ones to predict the price movement of a stock index.

The input of the method is a set of events in the form of news headlines, and the output is the price movement of a specific stock (i.e., DOWN, STAY, UP). The price movement is defined as the daily relative movement percentage that generates the most balanced distribution between the majority and minority output classes (e.g., for S&P500 this is +/- 0.03%). The headlines are embedded using BERT [4] contextualized embeddings. An attention layer [2] acts as input selector by computing for each headline in a time frame, a normalized weight. The weighted sum of the embeddings is used to predict the price movement.

We expect that this weight measures the contribution of a specific headline to the prediction. Here, we analyze to which extent those weights can be interpreted as the relative importance of each article to the price movement; we investigate if the attention layer allows us to generate a global ranking of events using its unnormalized weights so that it ranks higher the most relevant events when the dataset is analyzed in its entirety. In this exploratory work we use news categories as a proxy for event relevance: news belonging to more relevant categories should be ranked higher.

There is extensive literature in stock price prediction from news [11, 34, 6]. It has mostly focused on methods to exploit profitable trading opportunities by trying to predict the stock

price based on incoming news (i.e., the prediction should be solely based in recent past information). This approach presents a problem from an event extraction perspective as the event may have been already incorporated into the price before the news became public. This fact makes it difficult to develop a method that generalizes well; the association between the past input news and price movement becomes weak [20]. Our focus here differs from these approaches in that our ultimate goal is to identify which events are more informative to explain stock-prices movements. This allows us a more flexible approach as we are not conceptually bound to short time windows (in this paper we use 24 hours) or the publication time; for us, the stock prediction task is a proxy to identify relevant events.

We experiment on the most important US stock indices (i.e., S&P500, Nasdaq, Dow Jones, and Russell 1000) with the AP English news-wire subset of the Gigaword [12] dataset, a large corpus with more than 1,5 million articles spanning over 15 years (1994-2010). We first show that, as intuitively expected, the network tends to make better predictions if fed with news categories such as 'business' rather than 'sports' or 'entertainment.' Then we show that when fed with all categories at the same time, the network can skew the distribution via the attention weights towards meaningful categories, allowing us to infer that the network is able to extract stronger signals from more relevant events. Interestingly, we found that these two tasks are coupled as the performance (or lack of) in the first one seems to be correlated with the strength of the effect on the second one.

Chapter 2

Background and Related Work

2.1 Stock price prediction from the news.

Predicting stock prices from the news has been a long-standing goal. Stock prediction is the task of predicting stock prices with past information with a focus on maximizing profits. Deep learning methods have become the state-of-the-art for stock price prediction [13, 3]. The input usually consists of past technical data to capture the inertial component and several sources that reflect recent information such as social media streams, news, or other types of textual sources [35, 16].

As the goal of stock price prediction is to maximize profits, most of the literature focuses on the prediction solely based on recent past information. There has been a debate around the feasibility of predicting stock prices by looking at past news [20]. In this work, we do not necessarily need to enter this debate. From the event extraction perspective, it is irrelevant whether the news headline mentioning the event was published before or after the price movement; it is enough that the headline that generated the price movement is part of the input, no matter the length of the window. Our goal is to maximize event recognition performance and not trading profits.

Early work by [11] already showed a strong correlation between news articles and stock price movement. Multiple approaches were explored to extract signals from news such as sentiment analysis, named entities, semantic parsing, or neural encodings [27, 34, 19, 23]. Strategies that maximize profit also leverage other sources of information, such as past prices. The predictions usually run in chronological order.

Using news articles as input to neural networks requires the transformation of text into numerical input. This is achieved through the use of word embeddings.

2.2 Word Embeddings

The term *word embeddings* is used to refer to the representation of words as vectors in a low-dimensional continuous space. Word vectors are trained on large corpora, and each vector should be able to capture the context in which the word occurs. For instance, the words *king* and *queen*, or *table* and *chair*, are expected to be relatively close in a dimensional space. Even more, it has been shown that one could express certain semantic relations as word vector operations. For example, *king* - *man* = *queen* - *woman* [22].

The representation capacity of word vectors is not yet fully understood, but apart from semantic information, word vectors are known to capture syntactical [14], and even ontological [26] knowledge. From a methodological perspective, word embeddings have enabled the use of deep learning into the natural language understanding field.

Initially, word embeddings were static [21, 24] in the sense that each work would be associated with a single vector. These vectors were learned from large corpora. A word would be associated with a unique vector learned according to an operation between the vectors of the words in context.

2.2.1 Contextualized Embeddings

Static word representations present some limitations, especially for polysemous words, which have to share the same representation. To overcome these restrictions, recent work[25, 5] has focused on methods to generate contextualized representations, word vectors that are sensitive to the specific context in which they appear. This is achieved by training a neural network, which given a sentence or text span, assigns a vector to each word.

Contrary to static representations, in contextualized embeddings, the word vectors per-se are not learned but an artificial neural network that assigns them simultaneously for an input text. This means that each word can be associated with various vectors in different contexts. For instance, the word *president* in "The US president met with the Canadian president might have different representations in each occurrence (although probably quite related).

Another relevant aspect of the contextualized models is that the networks in which they rely on can be directly used as the backbone of the underlying task, which means that their weights can be further fine-tuned and accommodated to the underlying task (e.g., classification, sequence labeling, etc).

The use of contextualized representations generated tremendous improvements in diverse natural language understanding tasks. In this work, we use the contextualized vectors provided by the BERT model to encode the news headlines.

2.3 Event Extraction

Event extraction is the task of identifying event instances in texts[15]. Extracted events can be canonicalized or unncanonicalized. Unncanonicalized events are represented by the phrase in the original text ('US president signed a trade deal with China yesterday'). They are meant to be consumed by humans or further downstream tasks or applications. Canonical or machine-readable events are those whose arguments and relations are linked to a knowledge graph. This requires the extraction of attributes like *who*, *what*, *to whom*, *where*, *when*, *why*, and *how*, each of them linked to a knowledge base (*trade_deal(US, China, 2018_01_24)*). In this work, we focus on the identification of non-canonicalized events.

2.3.1 Event extraction of financial events.

One of the keys to predicting stock price movement is to understand which event may move the price in one or another direction. More recently, stock prediction literature has indeed focused on the explicit representation of events [6, 7, 1, 8, 17, 9, 29].

One line of work centered on the extraction of structured, canonical or semi-canonical events (linked to a knowledge base) [6–8, 23, 17], while another line has focused on non-canonicalized events (e.g., headlines or sentences referring to a relevant event) [9, 29].

[17] classify sentences into ten canonical events using manually annotated data from news articles mentioning seven companies. [9] overcome the limitation of relying purely on manually annotated data by using weak supervision to collect training data. The idea is to leverage Wikipedia to extract sentences containing events. A sentence is considered to hold a relevant event if, in a selected company Wikipedia article, the sentence is part of a specific section (e.g., history) and starts with a date. The model predicts, given an input sentence, if it is a financial event (as defined in the weakly supervised step before) or not. The text is encoded using BERT. In our approach, we adopt a fully unsupervised approach (i.e., we do not require semi-supervised annotations or other data sources), using only the price signal to detect the most relevant headlines.

[6], generate structured events from news headlines by using an open information extraction (Open IE) system [10]. The Open IE extractions are interpreted as events and linked to VerbNet and WordNet to generate canonicalized representations. [7] proposes to overcome the sparsity by using embeddings to represent the events, and [8] uses a knowledge base to improve those embeddings. Extracting event representations using Open IE is challenging. Open IE tends to generate a significant number of irrelevant and noisy facts, which eventually imposes the need for intense and informed data preprocessing as a way to improve the extractions. Here we postpone the canonicalization and concentrate on the recognition. Our

end-to-end alternative approach allows us to automatically select in one-shot the relevant events avoiding any involved preprocessing, compromise on the representation of the fact, or the use of any underlying relation extraction system.

Conceptually, the closest work to ours is probably [29]. They use an unsupervised method with a setting tailored for a visualization tool meant to be used by traders. Their focus is not event extraction but interpretability. The headlines presented to the network are already pre-selected for specific companies (i.e., the company has to be mentioned in the news). This pre-selection imposes a limitation as relevant news might not necessarily mention the most affected companies (e.g., an article about an oil price spike will not include every oil or car company or an article about a rate cut by the Fed will most surely not mention every financial institution). Our model allows us to present the network with all the news at once, as the goal is to allow the network to pre-select the relevant headlines. Also, their model does not explicitly look at the full headline but only at keywords which might be shared among multiple headlines; documents are encoded into bigrams, and relevant keywords are selected using LRP [1].

2.4 Attention Mechanism

To the best of our knowledge, the attention mechanism was introduced in natural language understanding by [2]. The idea of the attention mechanism is to allow the network to learn to focus on specific inputs. It is, not all inputs will contribute equally to a specific prediction,.The introduction of different attention mechanisms has had a significant impact in the field. In fact, attention is the base of the Transformer, one of the most successful neural architectures to date [30].

In this work, we follow [36], where the attention layer computes a weight for each of the inputs (headlines in our case) without any other element than the single input itself.

$$
\begin{aligned}
u_i &= tanh(W_h hl_i + b_h) \\
\beta_i &= u_i^T u_h \\
\alpha_i &= \frac{\beta_i}{\sum_t \beta_t} \\
v &= \sum_i \alpha_i hl_i
\end{aligned}
$$

where hl_i is the vector encoding of headline i, β_i is the unnormalized weight of headline i, u_h is a vector query representing all headlines, α_i is a normalized weight of the headline with

respect to all other input headlines, and v is the vector that summarizes all headlines in the input.

This type of attention is usually referred to as self-attention, as each input will determine its importance by itself. This allows us to use the unnormalized weight to generate a global ranking of headlines.

2.4.1 Attention as input selection.

A recent debate has erupted around the idea of using the attention mechanism [2] as a way of explaining model output [28, 18, 32]. The relation between attention weights and output has been unclear. In this regard, [18] found that different weight distributions can yield equivalent predictions. [28] and [18] concluded that attention weights are noisy and inconsistent and should not be used to explain a decision. [32] developed a set of tests to determine if attention weights are consistent enough to be taken as an explanation. In this paper, we are more interested in understanding the global effect of the attention layer and not its explainability in terms of single data point decisions. In our experiments, even though there is some variability, the global rankings generated by the unnormalized attention weights are consistent with expected results (i.e., which categories are more important to explain stock price movements) and generate consistent results across different runs and target stocks. A question that remains for future work is if we can indeed have more fine-grained consistency apart from the news categories.

Universidad de
SanAndrés

Chapter 3

Unsupervised event detection

We use the stock prediction task to identify events relevant to predict the stock price movements. This approach has the advantage that it does not require manually labeled data.

We use daily news headlines representing non-canonicalized events and their categories as input to predict the daily stock movement price. Note that no other data is used. Three classes represent the output: DOWN, STAY, UP with respect to the previous trading session. We include all headlines corresponding to a given date, and as target, we use the open price of the next trading session, assuming that all the information in the previous day news should be already incorporated in the stock price, including after-hour price movements. The full network is described in Figure 3.1.

More formally, each headline $hl_1, hl_2, \ldots, hl_k$ consisting of a (padded) sequence of N tokens $\{w_i\}_{i=1,\ldots,N}$, is encoded to vectors $\{\mathbf{hhl}_i\}_{i=1,\ldots,N}$ of length 768, using the pooled output of the BERT-base-uncased model:

$$\mathbf{hhl}_i = \mathrm{BERT}(hl_i)$$

Each headline comes with a corresponding single category label $hc_1, hc_2, \ldots$ (automatically labelled, details see "News Classification" in Section 4.2) which is embedded to a vector of length 30

$$\mathbf{hhc}_i = \mathrm{embed}(hc_i),$$

Both vectors are concatenated

$$\mathbf{h}_i = \mathbf{hhl}_i \oplus \mathbf{hhc}_i$$

and projected to a vector $\mathbf{h}_{p_i}$ of length 100 by a fully connected feed forward layer with ELU activation

$$\mathbf{h}_{p_i} = \mathrm{FF}_{\mathrm{ELU}}(\mathbf{h}_i)$$

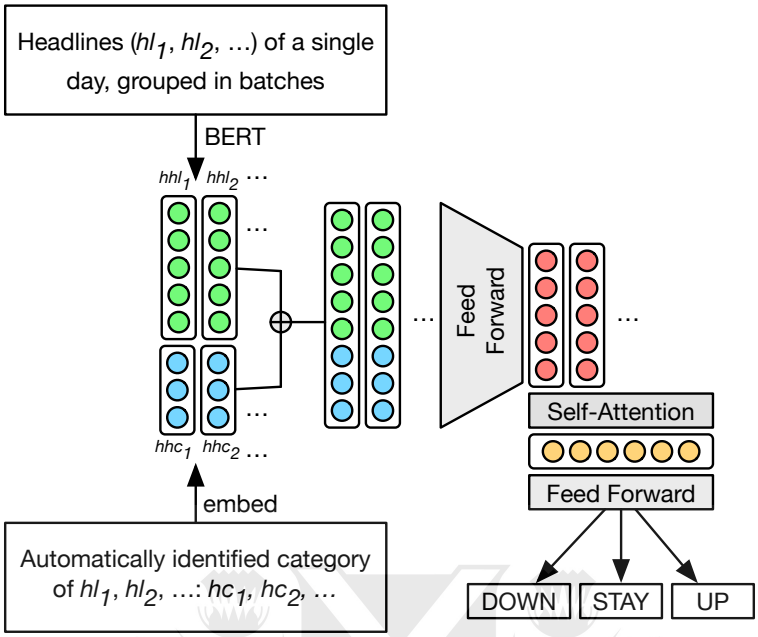


Fig. 3.1 Neural Network Layout for Stock Prediction

Following [36], an attention layer computes normalized weights for each headline of the input day, $\mathbf{H}_{p_d} := \{\mathbf{h}_{p_i}\}_{i=1,\ldots,k}$, and aggregates them according to those weights

$$\mathbf{h}_{a_s} = \text{SelfAttention}(\mathbf{H}_{p_d})$$

The final label l_i (DOWN, STAY, UP) and probabilities for each label are computed using a feed forward layer with softmax activation

$$l_i = \text{FF}_{\text{softmax}}(\mathbf{h}_{a_i})$$

Every input layer is normalized, and the weights are initialized using *He*. The dropout rate is 0.25. We use the out-of-the-box optimizer[1] that is provided by the official TensorFlow BERT repository. All BERT weights are fine-tuned during training.

[1] https://github.com/tensorflow/models/tree/master/official/nlp/bert

Chapter 4

Experimental Evaluation

In this chapter, we address two challenges: First, we need to understand if the network is indeed able to extract correct signals from the headlines and second if the attention layer is able to rank the events in terms of relevance.

The first set of experiments (Sec. 4.5) address the first challenge. It shows that, as intuitively expected, the price movement prediction for a single day is generally better if the model is trained on more relevant news categories. The idea is that the network should do better if it is fed with only business news than if it is fed, for instance, with entertainment headlines.

The second round of experiments (Sec. 4.6) analyzes if the attention layer via its unnormalized weights is capable of generating a global ranking of relevant headlines coherent with the previous results, when all headlines from all categories are provided as input. We expect that in the top-k positions, the distribution is skewed towards the relevant news category according to the results of the previous experiment.

Trading sessions	3777
Time-frame	11 Nov 1994 – 31 Dec 2010
Headlines	1,532,260
Mean	405.68
Std.	134.49
Min.	1
Max.	1213

Table 4.1 Dataset

4.1 Dataset

We used the AP headlines of the English Gigaword dataset [12], a collection of English newswire data with 1.5M articles published between 1994 and 2010. Regarding the stocks to be tested, we selected the most relevant US indices: S&P500, Dow Jones, Nasdaq, and Russell 1000, which we downloaded from Yahoo! Finance. Statistics of the dataset are displayed in Table 4.1.

4.2 News Classification.

We trained a news classifier on the TagMyNews [31] dataset. It consists of 32,567 headlines classified into 6 categories: 'business', 'entertainment', 'health', 'sci-tech', 'sport', 'us' and 'world'.

The input of the model is a single headline. The headline is embedded using the BERT-base-uncased pooled output, and the embedded headline serves as input to a fully connected layer that generates a binary classification score for each category. We use a dropout of 0.25 and the out-of-the-box optimizer provided by the official BERT TensorFlow distribution. The batch size was set to 120, and the max length of the headlines was limited to 15 WordPiece tokens [33]. We used early stopping with respect to accuracy to select the best model.

The size of the validation set was 0.2, and the model performance on this set was F1 0.85, which is in line with the state-of-the-art [37] for this dataset.

The news classifier was used to categorize the headlines from the Gigaword dataset. To assign a class to a headline, we considered the class with the highest score with a threshold of 0.5. Headlines for which every category was below the threshold were left unclassified and not used in the experiments. Table 4.2 shows the distribution of headlines per category. In total, 66,891 headlines were discarded, 4.37% of the total.

4.3 Preprocessing.

Given resource constraints, we are limited to 115 headlines per day with a maximum length of 15 WordPiece tokens each. To account for more than 115 news headlines, we created stratified subsets to generate several data points for one single day. The stratified samples were generated with respect to the headline categories. We discarded dates with less than 25 headlines for each of the four most prominent categories (i.e., 'world,' 'sports,' 'business', and 'us'), dropping 511 data points, a 13.53% of the total. We also remove headlines with less than 20 characters, which tend to be incomplete or noisy.

Category	Number of articles	%
world	596,899	38,96
sport	275,585	17.99
business	231,083	15.08
us	211,570	13.81
unclassified	66,891	4.37
entertainment	54,607	3.56
sci-tech	54,057	3.53
health	41,568	2.70

Table 4.2 Distribution of news per category

To generate the labels for each stock (i.e., DOWN, STAY, UP), we need to set a threshold to determine the classification of the price movement. This threshold will depend on the practitioners' strategy. Here we pick the threshold to generate a balanced distribution among the classes in order to avoid the network learning trivial decisions. We set for each stock a symmetric threshold between [1%, 0.1%] (in steps of 0.1) such that the distribution of classes between the majority and the minority class is the most balanced.

The final class distribution for each stock in the relevant dates and the corresponding thresholds are displayed on table 4.3.

Stock Index	Threshold	DOWN	STAY	UP
S&P500	+/- 0.3%	30.91%	33.61%	35.48%
Nasdaq	+/- 0.3%	30.48%	29.83%	39.69%
Dow Jones	+/- 0.3%	30.53%	33.23%	36.23%
Russell 1000	+/- 0.3%	29.47%	34.38%	36.15%

Table 4.3 Thresholds and class distributions

4.4 Training setup.

We ran our network on 3 Tesla V100 GPUs with a total batch size of 15. The test size was set to 0.2.

4.5 Stock price prediction

We ran the network on each news category separately to understand if the network was capable of extracting the right signals from the headlines. In this experiment we selected the model with the maximum accuracy, a maximum of 20 epochs, and a patience of 5 epochs.

As expected, 'business' headlines are more informative and consistent across the different indices for predicting the stock price movement. In fact, it is the only news category from which the network seems to extract a meaningful signal. For the rest of the news categories, the network does not perform much better than a random uniform choice. Except for business, the network is quite unstable, with most epochs not able to generate a precision or recall score above 0.

Regarding the individual indices, results are consistent across indices, except Nasdaq. This might be due to the generality of the dataset, most likely not suitable for Nasdaq but more appropriate for the more diversified S&P500, Dow Jones, and Russell 1000, which cover a more comprehensive range of sectors. Interestingly, the best result for Nasdaq was achieved with the sci-tech category.

Results when all news are included, with the exception of Nasdaq, are lower than the 'only business' setting. However it is still clear that informative signals are extracted.

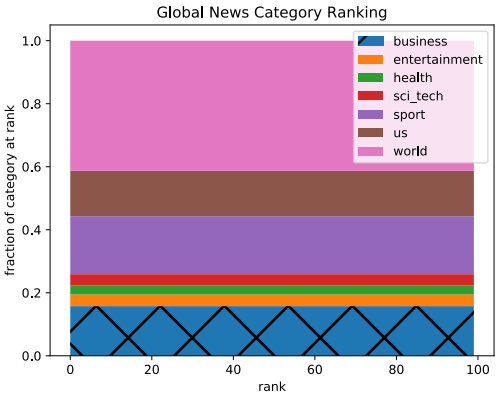
Table 4.4 shows the results for all categories.

News Category	S&P500	Nasdaq	Dow Jones	Russell 1000
business	**57.88**	43.64	**61.97**	**55.92**
us	40.13	38.45	42.02	39.59
world	41.83	44.89	39.66	38.73
sports	38.94	44.09	36.36	38.94
sci-tech	36.74	**44.96**	37.05	36.90
entertainment	34.57	42.33	37.98	38.14
health	34.57	40.78	34.26	35.81
all	52.92	45.22	54.99	54.49

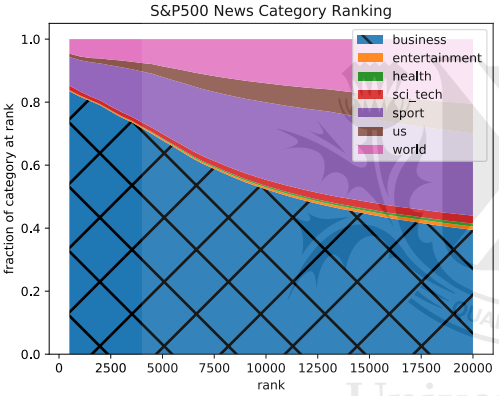
Table 4.4 Max accuracy of each news category on the stock prediction task

4.6 Event detection

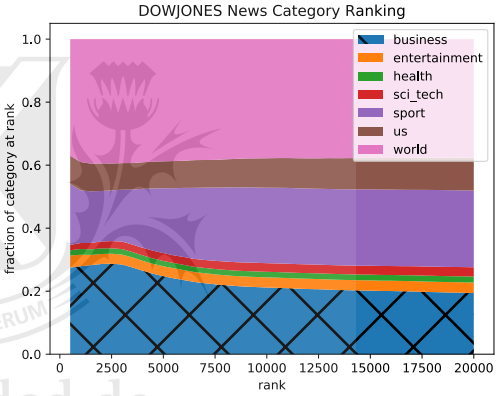
In this experiment we analyze if the attention layer weights can be used to generate a meaningful global ranking of the news headlines. We understand meaningful as ranking that



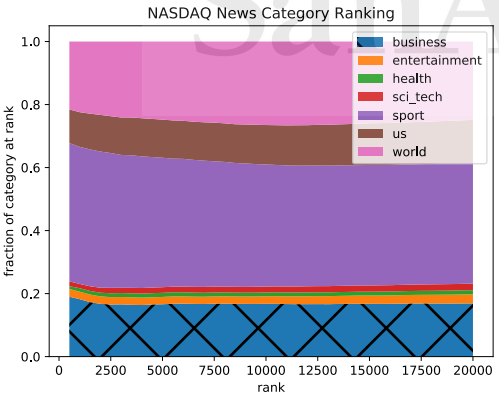
(a) Global



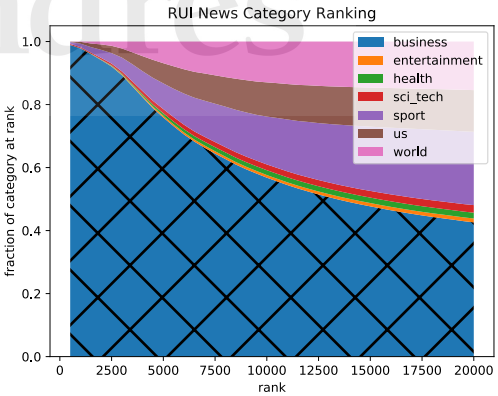
(b) S&P500



(c) Dow Jones



(d) NASDAQ



(e) Russell 1000

Fig. 4.1 Event Detection Results

favours news categories with stronger signals in the top positions; business news according to experiments in Section 4.5.

In this case we provided the entire set of news at once, regardless of the news category. After the network was trained, we used the unnormalized weight of the attention layer to rank all the headlines across all days. There are a total of 271,520 headlines in the test set. Results are the average over five runs.

Figure 4.1 shows the results of the experiment, showing the 20,000 headlines with the highest unnormalized weight during prediction. They show for a given rank (e.g., top 2500) the fraction of headlines with given categories up to that rank. Figures 4.1b, 4.1c, 4.1d and 4.1e display the distribution of categories for the top-k headlines for S&P500, Nasdaq, Dow Jones and Russell 1000 respectively. Additionally, Figure 4.1a shows the overall distribution of the test set. For S&P500, Nasdaq, Dow Jones, and Russell 1000, the top-k distribution is strongly skewed toward business news (compared to the global distribution), with only a minimal effect for Nasdaq.

This is consistent with the results of the experiments in Sec. 4.5. The three indices for which the network extracts signals from the news tend to generate changes in the distribution of categories, and the one that does not seem to extract any signal, Nasdaq, does not.

Regarding model training: For each stock index we selected the model with the minimum loss, a maximum of 20 epochs and a patience of two epochs. Note that the model selection strategy is different than in Section 4.5 where we maximized over accuracy. Minimizing loss reduces the variance of the results across different runs.

4.7 Anecdotal data

Table 4.5 shows the top 100 headlines over the whole timespan, ranked using the unnormalized attention layer weights of the model trained for S&P500. The examples clearly show that the model can discriminate market-relevant headlines from ones that are not: headlines highlighting general market trends such as stock movements, significant efforts by major companies, or commentary by public institutions make up the top ranks.

Results for a random single date (2003-3-20) show the same trend (Table 4.6). It is interesting to note that as for a single day specific news about stock movements are not many, the top ranking has also space for other relevant economic or political events.

Rank	Headline	Rank	Headline
1	FTSE 100 down 42.24 at 4,640.22	51	Dow Drops 53; Nasdaq Falls 49
2	FTSE 100 down 87.91 at 4,389.11	52	Dow Drops 53; Nasdaq Falls 49
3	Dow Drops 11; Nasdaq Falls 2	53	Agassi falls to Calleri at Nasdaq-100 Open
4	FTSE 100 down 36.13 at 3,993.54	54	Dow Drops 184; Nasdaq Falls 113
5	DuPont lowers 4Q profit outlook	55	Fed says US economy logged slower growth in the fall
6	Nasdaq Sheds 83; Dow Falls 8	56	Dow Down 24, Nasdaq Slips 4
7	Dow Falls 1.6; Nasdaq Drops 22	57	Nasdaq Falls 53; Dow Drops 55
8	Brooks Just Keeps Playing Through PGA Woes By MARK R. CHELLGREN	58	Dow Down 100, Nasdaq Falls 82
9	Dollar rebounds as sentiment improves	59	Dow Down 60.50; Nasdaq Off 8.78
10	Canada Falls to 0-2 at Little League World Series Eds: UPDATES with two	60	FTSE-100 down 8.21 at 4,300.87
11	games; Will be updated with late games	61	Dow Down 89, Nasdaq Falls 10
12	Delta plans more asset sales to build shareholder value	62	Dow crosses 10,000 as stocks continue their rally
13	Dow Drops 10, Nasdaq Climbs 4	63	FTSE-100 down 46.73 at 4,005.50
14	EBay reshapes itself as an easier, cooler place, but will buyers come?	64	Dow Falls 39; Nasdaq Gains 2
15	Nasdaq Falls From 5000; Dow Off 266	65	Maggert sent sliding with two-hole meltdown
16	Nasdaq Falls From 5000; Dow Off 235	66	U.S. stocks open quietly amid mixed earnings news
17	FTSE 100 down 38.33 at 3,850.73	67	Lighter, faster Serena Williams off and running at Rogers AT&T Cup
18	Dow Closes Down 124; Nasdaq Falls 64	68	Corning raises earnings forecast
19	NBA Contracts Soar Past Dlrs 100 Million By RONALD BLUM	69	Dow Down 249, Nasdaq Falls 100
20	FTSE 100 down 204.47 at 3,883.36	70	Down Falls 265, Nasdaq Down 179
21	SLOC says sales are booming	71	WPP returns to growth in 2003, is upbeat on 2004
22	Greenspan said economic prospects seem brighter, but growth will	72	Stocks rally in anticipation of possible rate cut this week Eds:
23	Financial crisis coverage dominates Loeb Awards	73	UPDATES with midday trading, ADDS details.
24	Dow Slips 15; Nasdaq Drops 48	74	Dow Ends Down 382; Nasdaq Falls 94
25	FTSE 100 down 184.28 at 4,209.93	75	Dow Ends Off 71.36; New Nasdaq High
26	Dow Down 156, Nasdaq Falls 30	76	Stocks open flat ahead of consumer confidence data
27	Dow Down 194, Nasdaq Slips 65	77	Dow Drops 16; Nasdaq Climbs 27
28	Dow Down 44; Nasdaq Dips 32	78	Hamilton battles at bat and in his life
29	Fed's interest rate stance jump-starts Wall Street; Dow gains 129.91	79	U.S. stocks finish flat as investors await catalysts to push stocks higher
30	Dow Falls 77; Nasdaq Dips 62	80	Dow Down 51, Nasdaq Slips 13
31	Henman keeps making it look difficult	81	Dow Down 109, Nasdaq Down 30
32	Dow Down 96, Nasdaq Falls 93	82	Tech stocks surge, but Dow falls short of 10,000 By LISA SINGHANIA
33	Dow Drops 92; Nasdaq Falls 55	83	In Pittsburgh, the big hits always keep on coming
34	Shea seats selling briskly at 869 a pair	84	Stocks Rebounds to Record High After Fed Move
35	Dow Drops 17; Nasdaq Down Fraction	85	Dow Falls 149; Nasdaq Drops 115
36	Nasdaq Off 189.22; Dow Ends Up 82.61	86	Dow Closes Down 292, Nasdaq 110
37	Dow Drops 145; Nasdaq Falls 57	87	Dow Ends Down 55; Nasdaq Falls 44
38	Dow Falls 26; Nasdaq Drops a Fraction	88	U.S. stocks turn mixed in early trading amid drop in retail sales, higher oil prices
39	Timken slashes earnings outlook, 700 jobs due to weak U.S. auto market	89	Avis Budget swings to 3Q loss due to hefty charges from Cendant breakup
40	Dow, Nasdaq Fall Slightly	90	Kiptanui Narrowly Misses Steeplechase Record, Bailey Fails In 100 Eds:
41	Hernandez falters in fifth inning of relief	91	will be updated with later events
42	U.S. stocks finish at four-year highs	92	Stocks make modest advance on better-than-expected earnings, economic
43	Dow Slips 33; Nasdaq Drops 6	93	news
44	FTSE 100 down 73.09 at 4,368.86	94	FTSE 100 down 33.70 at 4,246.28
45	Dow Down 64, Nasdaq Falls 45	95	U.S. stocks try for advance despite Best Buy warning, home construction report
46	Dow Drops 81; Nasdaq Falls 25	96	URGENT U.S. economy soars by 5.8 percent rate in first quarter, best
47	Beem's timing not the best at rainy Augusta	97	Nestle Profits Down 10 Percent In 1995
48	West Indies out for 152; follows on 225 in arrears	98	Dow Falls 106; Nasdaq Drops 38
49	Dow Off 134; Nasdaq Down 59	99	Florida teams cancel overseas trips because of war
50	Dow Slips 40; Nasdaq Falls 29	100	Dow Ends Off 66; Nasdaq Sheds 46

Table 4.5 Top 100 out of 271,520 headlines for S&P500

Rank	Headline
1	US stocks rebound from sharp sell-off; manufacturing reading boosts confidence
2	US stocks head for higher open after Wednesday ' s slide; Nike profit gives market a lift
3	Obama ties Iraq war to weak U.S. economy, McCain meets with Gordon Brown in London
4	US jobless claims rise by larger-than-expected amount
5	Credit Suisse slashes 2007 profit forecast after internal probe
6	US Vice President Cheney in Afghanistan to bolster government struggling with rising threats
7	Wall Street rebounds after big drop on profit-taking
8	Ireland ' s stock market jumps after financial regulator unveils probe into ' false rumors '
9	Umpire Hair plans to communicate better with players when he returns
10	Wall Street rises after Wednesday drop, but jittery after report shows spike in jobless claims
11	FedEx 3rd-qtr earnings fall 6 pct on high oil prices and slow US economy, beat Wall St. views
12	Drury and Dawes team to give Rangers 6th straight win over Devils
13	75-pound stingray leaps from water, kills Michigan woman sunbathing on boat in Florida Keys
14	Brazil investors jittery with declining commodities prices
15	Commodities prices plunge on stronger dollar, fund selling
16	Oil prices drop, near mid-US 102 after US report shows oil, gasoline demand softening
17	White House press secretary fumbles argument about Bush, says people required not to like him
18	JPMorgan Chase makes it difficult for third parties to make offer for Bear Stearns
19	EUROPE NEWS AT 1200GMT
20	Barnsley able to field Steele in FA Cup semifinal after signing him on loan for rest of season
21	Huge names in soccer converge as Beckham presents lifetime achievement award to Pele
22	Oil prices drop on concerns the slowing US economy is cutting demand
23	Towns across Midwest brace for more flooding as rivers continue to rise from heavy storms
24	Eddie O ' Sullivan resigns as Ireland rugby coach after back-to-back tournament flops
25	75-pound stingray leaps from water, kills Michigan woman on boat in Florida Keys
26	EUROPE NEWS AT 1900GMT
27	CONCACAF Champions Cup Glance
28	Iraq PM urges caution in choosing provincial officials
29	White House says Olympics are about athletes
30	Bulgarian unemployment at 7.26 percent in February
31	Manchester United wary of rivals despite 3-point lead
32	IAAF finds 10 positive results among more than 3,000 doping tests in 2007
33	Afghanistan ' s Karzai declines to say whether he ' ll seek another term as president
34	CONCACAF Champions: Mexico ' s Pachuca defeats Honduras ' Motagua 1-0 in
35	Bulgaria recognizes Kosovo as independent nation
36	Bill Clinton withdraws from Northern Ireland event celebrating 1998 peace accord
37	Vancouver organizers say no to boycott of Beijing Olympics
38	Court gives former Finnish NHL player Karalahti suspended sentence in drugs case
39	Federal Reserve says it will auction 75 billion in Treasury securities next week
40	Despite upset, Santos will play in Bolivia ' s altitude ' whenever necessary '
41	Heads of major UK banks meeting Bank of England governor Thursday
42	Pop songs, billboards help overcome fear of census after decades of war in Liberia
43	Israel on alert ahead of Purim holiday
44	US presidential hopeful McCain visits London, hails bravery of UK troops in Iraq
45	Denmark keeps security assessment unchanged after bin Laden threat
46	Iraqi government clears major obstacle to provincial elections
47	Serbia ' s Kostunica criticizes Bush ' s decision to allow U.S. military assistance for Kosovo
48	UN Security Council set to renew political mission in Afghanistan
49	Pakistan promises strict security for Asia Cup cricket teams
50	Back in England, Thaksin considers divesting business interests to retain Man City for life

Table 4.6 Top 50 results out of 105 total for S&P500 on random date (2008-03-20)

Chapter 5

Conclusion and future work

We presented an exploratory analysis, understanding the possibility to generate a ranking of relevant events in an unsupervised way. We showed that a simple neural network is able to extract informative signals from news, and that the attention layer was able to rank higher the most relevant news category.

Future work needs to focus on a more fine-grained analysis of the data. It should try to generate stable rankings beyond news categories and understand the limits of a purely unsupervised approach. It would be important also to understand when we can trust specific rankings, probably focusing the analysis on the attention layer [32].

Universidad de
SanAndrés

References

[1] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, (7):e0130140.

[2] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR*.

[3] Chong, E., Han, C., and Park, F. C. (2017). Deep learning networks for stock market analysis and prediction. *Expert Syst. Appl.*, 83(C):187–205.

[4] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

[5] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, pages 4171–4186.

[6] Ding, X., Zhang, Y., Liu, T., and Duan, J. (2014). Using structured events to predict stock price movement: An empirical investigation. In *EMNLP*, pages 1415–1425.

[7] Ding, X., Zhang, Y., Liu, T., and Duan, J. (2015). Deep learning for event-driven stock prediction. In *IJCAI*, IJCAI'15, page 2327–2333.

[8] Ding, X., Zhang, Y., Liu, T., and Duan, J. (2016). Knowledge-driven event embedding for stock prediction. In *COLING*, pages 2133–2142.

[9] Ein-Dor, L., Gera, A., Toledo-Ronen, O., Halfon, A., Sznajder, B., Dankin, L., Bilu, Y., Katz, Y., and Slonim, N. (2019). Financial event extraction using Wikipedia-based weak supervision. In *Second Workshop on Economics and Natural Language Processing*, pages 10–15.

[10] Fader, A., Soderland, S., and Etzioni, O. (2011). Identifying relations for open information extraction. In *EMNLP*, page 1535–1545.

[11] Gidófalvi, G. (2001). Using news articles to predict stock price movements.

[12] Graff, D., Kong, J., Chen, K., and Maeda, K. (2003). English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

[13] Hedayati, A., Moghaddam, M., and Esfandyari, M. (2016). Stock market index prediction using artificial neural network:. *Journal of Economics, Finance and Administrative Science*.

[14] Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *NAACL-HLT*, pages 4129–4138.

[15] Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F., and Caron, E. (2016). A survey of event extraction methods from text for decision support systems. *Decis. Support Syst.*, 85(C):12–22.

[16] Hu, Z., Liu, W., Bian, J., Liu, X., and Liu, T.-Y. (2018). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *WSDM*, page 261–269.

[17] Jacobs, G., Lefever, E., and Hoste, V. (2018). Economic event detection in company-specific news text. In *First Workshop on Economics and Natural Language Processing*, pages 1–10.

[18] Jain, S. and Wallace, B. C. (2019). Attention is not explanation. In *NAACL-HLT*, pages 3543–3556.

[19] Li, X., Xie, H., Chen, L., Wang, J., and Deng, X. (2014). News impact on stock price return via sentiment analysis. *Know.-Based Syst.*, 69(1):14–23.

[20] Merello, S., Picasso Ratto, A., Ma, Y., Luca, O., and Cambria, E. (2018). Investigating timing and impact of news on the stock market. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1348–1354.

[21] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013a). Distributed representations of words and phrases and their compositionality. In *NIPS*, page 3111–3119.

[22] Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *NAACL-HTL*, pages 746–751.

[23] Peng, Y. and Jiang, H. (2016). Leverage financial news to predict stock price movements using word embeddings and deep neural networks. In *NAACL-HLT*, pages 374–379.

[24] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

[25] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237.

[26] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *EMNLP-IJCNLP*, pages 2463–2473.

[27] Schumaker, R. P. and Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Trans. Inf. Syst.*, 27(2).

[28] Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *ACL*, pages 2931–2951.

[29] Shi, L., Teng, Z., Wang, L., Zhang, Y., and Binder, A. (2019). Deepclue: Visual interpretation of text-based deep stock prediction. *IEEE TKDE*, 31:1094–1108.

[30] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*, pages 5998–6008.

[31] Vitale, D., Ferragina, P., and Scaiella, U. (2012). Classification of short texts by deploying topical annotations. In *ECIR*, page 376–387.

[32] Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *EMNLP-IJCNLP*, pages 11–20.

[33] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*.

[34] Xie, B., Passonneau, R. J., Wu, L., and Creamer, G. G. (2013). Semantic frames to predict stock price movement. In *ACL*, pages 873–883.

[35] Xu, Y. and Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. In *ACL*, pages 1970–1:979.

[36] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. In *NAACL-HLT*, pages 1480–1489.

[37] Zeng, J., Li, J., Song, Y., Gao, C., Lyu, M. R., and King, I. (2018). Topic memory networks for short text classification. In *EMNLP*, pages 3120–3131.

Universidad de
SanAndrés