Financial Event Extraction Using Wikipedia-Based Weak Supervision

Liat Ein-Dor, Ariel Gera, Orith Toledo-Ronen, Alon Halfon, Benjamin Sznajder,
Lena Dankin, Yonatan Bilu, Yoav Katz and Noam Slonim
IBM Research, Haifa, Israel

Abstract

Extraction of financial and economic events from text has previously been done mostly using rule-based methods, with more recent works employing machine learning techniques. This work is in line with this latter approach, leveraging relevant Wikipedia sections to extract weak labels for sentences describing economic events. Whereas previous weakly supervised approaches required a knowledge-base of such events, or corresponding financial figures, our approach requires no such additional data, and can be employed to extract economic events related to companies which are not even mentioned in the training data.

1 Introduction

Event Extraction from text (Hogenboom et al., 2011; Ritter et al., 2012; Hogenboom et al., 2016) has been the subject of active research for over two decades (Allan et al., 2003). Detection and extraction of finance-related events have mostly focused on events described in news articles, which are likely to impact stock prices. In particular, previous work has sought to extract descriptions of events pertaining to a specific company, and analyzed how such events correlate with measures of that company's stock (price, volatility etc.). While much of the literature has focused on the prediction of stock prices (e.g., Ding et al., 2015; Xie et al., 2013), it is recognized that predicting future stock movements is a formidable challenge (see e.g. Merello et al., 2018); still, there are use-cases that might benefit from business-related event extraction from news.

One promising direction is enhancing the finance-related research performed by finance analysts. Such research typically requires reviewing a large body of news data under severe time constraints. We propose an automatic system for highlighting meaningful company-related news events that are likely to deserve the analyst's attention.

Work on economic event extraction often defines an ad-hoc taxonomy of events, and what constitutes an 'important event' for one might not be considered as such for another. For instance, the CoProE event ontology (Kakkonen and Mufti, 2011) includes events such as patent issuance and delayed filing of company reports, which are not considered by Du et al. (2016); similarly, while CoProE consider earnings estimates by analysts as events, Jacobs et al. (2018) examine instead analyst buy ratings and recommendations.

Outlining a comprehensive list of event types seems futile. For example, if a company's databases are hacked, this is certainly an influential event; but compiling an explicit and exhaustive event taxonomy that is sufficiently fine-grained to include all events such as this one is doomed to fail. At the same time, a formal event hierarchy is not necessarily required from an analyst's perspective. The strength of an automated system comes from the ability to process a large volume of news data and detect events of interest; automatically classifying these events into types is probably of secondary importance to an expert in the field.

Thus, our focus here is on a binary classification problem that is not type-based. This presents an interesting challenge, since the aim is not capturing the characteristics of predefined event types, but rather capturing general properties of relevant events.

The common NLP approach for economic event extraction has mostly made use of hand-crafted rules and patterns (Feldman et al., 2011; Arendarenko and Kakkonen, 2012; Xie et al., 2013; Hogenboom et al., 2013; Ding et al., 2014, 2015; Du et al., 2016). However, creating and maintaining such rules is time consuming, and further seems less suitable for our scenario, where no set

Proceedings of the Second Workshop on Economics and Natural Language Processing, pages 10–15
Hong Kong, November 4. ©2019 Association for Computational Linguistics

of underlying event types (which give rise to such rules) is assumed. Hence, here we follow a different, more flexible approach, that relies on a robust statistical learning framework for identifying relevant events. In particular, we adopt a supervised learning approach for identifying events related to a given company, and suggest to train a sentence-level classifier for this purpose. Given sentences from news articles discussing the company, the classifier aims to identify sentences containing events that would be of interest to the analyst. Since the sentences come from articles discussing the company, our main focus is on determining whether a sentence conveys an event worth considering, and not on ascertaining that it is related to the company.

Learning a supervised model requires annotated data. The standard approach for obtaining annotated data involves human annotation, which requires a substantial effort and limits the size of the data, which in turn may hinder the results. One way to overcome this problem is using weak supervision (Zhou, 2017), where labelled data is generated automatically using heuristics rather than manual annotation. Although such data may be noisier and less precise compared to standard labelled data, it enables to create much larger amounts of data at a significantly lower cost. Here we rely on content from Wikipedia to automatically generate a weakly-labelled sentence dataset for company events. We report experimental results that demonstrate the potential merit of our approach.

2 Related Work

Arendarenko and Kakkonen (2012) relied on a collection of hand-crafted detection rules in order to recognize 41 distinct company-related event types, and Du et al. (2016) used about 600 distinct patterns to cover 15 business event types.

More recently, machine-learning techniques were considered for this task. Jacobs et al. (2018) frame the problem as a multi-class classification task. They define a taxonomy of 10 event types, in addition to a ”no-event” class, and 7 companies of interest, and rely on manual annotation to train a sentence-level multi-class classifier. Testing several classifiers, they show that a linear SVM classifier attains the best results for most event types. While the current paper also adopts a supervised learning sentence-level approach, here the data is constructed based on weak labels, and the task is framed as a type-independent binary classification problem.

Rönnqvist and Sarlin (2017) used weak supervision in the context of financial events, focusing on bank distress events. They consider 101 banks for which 243 such events, and their date, are known. They then extract 386K sentences referring to these banks, and consider a sentence as describing a distress event if there is a matching event in the knowledge base mentioning the same bank and occurring near the publication date of the article from which the sentence was extracted. This approach requires a large knowledge-base of specific events, which is not readily available when moving from a confined event type (i.e. bank distress) to a diverse space of events. In this work we suggest a weak-label approach that aims to encompass a variety of relevant entities, event types and event occurrences.

3 Data

We used two types of datasets, one which is created automatically based on weak labels, and another which is based on manual annotation.

3.1 Weakly labelled datasets - Wikipedia

We leverage the content of Wikipedia articles describing companies as a source of influential events in the company's chronology.

In order to automatically identify ’positive’ sentences which likely describe noteworthy events, we rely on two observations: 1. Such events tend to appear within specific Wikipedia sections. 2. Sentences beginning with a date, specifically the *date-pattern* $['On/In/By/As\ of' + month + year]$, often describe an event. Thus, we manually created a lexicon of words which tend to appear in the titles of event-prone sections. A section whose title contains one of the following words is defined as an *event-section*: history, creation, leadership, corporate, acquisitions, growth, finance, financial, lawsuits, litigation, legal.

Given a company C, we select from its Wikipedia article all sentences appearing in an *event-section* and starting with a *date-pattern*. We remove the opening date and mark the sentences as positive examples with respect to C. All sentences which do not start with a *date-pattern* and are not in an *event-section* are considered as negative. To balance the dataset, we enforce an equal number of positive and negative examples by discarding sentences from the larger set. In addition, since many positive examples begin with either the company's name or the words "the company", we aim to balance the two classes in terms of sentences containing these patterns. The rest of the negative examples are chosen at random.

The procedure described above was used to create two datasets. The first, $S\&P\text{-}wiki$, is generated from Wikipedia articles of the companies on the S&P-500 index. A larger dataset, $Extended\text{-}wiki$, was later generated from Wikipedia articles of companies traded in one of five major stock exchanges[1], yielding 3.8K companies in total.

Each dataset was split into train and test sets based on dates - all positive examples up to 2018 are in the training set, and all those from 2019 are in the test set. Negative examples, which have no date attached, were split at random between the two sets, keeping the number of negative and positive examples equal within each set. Table 1 indicates the statistics of the resulting datasets, which will be released as part of this work.

Model	Train	Test
$SentiFM\text{-}binary$	8943 (0.2)	443 (0.2)
$S\&P\text{-}wiki$	6130 (0.5)	272 (0.5)
$Extended\text{-}wiki$	20074 (0.5)	908 (0.5)

Table 1: Data size (number of sentences) for the three models. The numbers in parenthesis indicate the percentage of positive samples.

Company	Articles	Sentences
Apple Inc.	438	10627
Facebook	302	6827
Qualcomm	120	3332
FedEx	67	1808
Anadarko Petroleum	91	1478
Xilinx	53	569
MGM Resorts International	32	463
Accenture	24	442
Allergan	35	421
Campbell Soup Company	27	307

Table 2: Number of articles and sentences in the $News$-2019 evaluation data.

3.2 Manually labelled dataset - SentiFM

To the best of our knowledge, the only manually annotated dataset for event detection in news articles is SentiFM (Jacobs et al., 2018). This dataset contains manual annotations of sentences into 10 predefined financial event types. However, this dataset is designed to solve a slightly different problem from the one explored in this paper. SentiFM was constructed in the context of a multi-class classification problem, whereas here we deal with a binary problem. Namely, we are not interested in event types, and do not assume there is a closed set of underlying types describing the events of interest. Indeed, it is possible that an event of interest might not be included in the SentiFM taxonomy, and hence a corresponding sentence would be labeled as negative. Despite these differences, we sought to examine how a classifier trained on the SentiFM data would perform on our task. To this end, we created a *binary* version of SentiFM, by considering all 'no-event' sentences as negative examples, and all event types as positives. We kept the original train/test split (see Table 1) and denote this data set as $SentiFM\text{-}binary$.

3.3 2019 News Sentences - $News$-2019

In order to evaluate methods for detecting company-related events within news data, we compile a set of sentences from news articles. Specifically, we selected the 10 S&P companies with the largest number of events from 2019 mentioned in their Wikipedia page (see Table 2). For each company, we retrieved all articles from 2019 on Seeking Alpha[2] that contained the company name in their title. We assume that this set of articles provides a good coverage of the company's events of interest during 2019. We applied sentence-splitting[3] on the retrieved articles, keeping only sentences 10-50 tokens long.

4 Experiments

The datasets described in Section 3 were used to train three event detection models. All classification models are based on BERT (Devlin et al., 2018), which has shown state-of-the-art results in many NLP tasks. We use a single-sentence input, and fine-tune the classifier with the $SentiFM\text{-}binary$, $S\&P\text{-}wiki$ and $Extended\text{-}wiki$ data sets. Henceforth, we will use these

[1] Hong Kong, London, NASDAQ, NYSE and Tokyo; Extracted via Wikipedia categories of these exchanges.

[2] seekingalpha.com; Transcriptions of company earning calls were filtered out due to their unique nature.

[3] using the NLTK library

names to refer to their corresponding BERT models. We use the BERT$_{\text{BASE}}$ model configuration, with maximum sequence length of 256, batch size of 16, dropout rate of 0.1 and learning rate of 5e-5. Each model was fine-tuned over 3 epochs, using a cross-entropy loss function.

4.1 Initial model evaluation

We first evaluate the performance of the three models on their corresponding test sets. As shown in Table 3, all models reach high performance when tested on the same type of data used in training. Next, we evaluate these models on the *Extended-wiki* test set (see Table 3). Notably, although less than 15% of the companies in *Extended-wiki* are in *S&P-wiki*, the latter model exceeds 90% precision and recall over the *Extended-wiki* test data. This suggests that the model is also able to detect events for companies that were not seen in training.

Model	Precision	Recall	F1
SentiFM-binary	0.97	0.96	0.96
S&P-wiki	0.97	0.92	0.94
Extended-wiki	0.93	0.95	0.94
SentiFM-binary	0.80	0.30	0.44
S&P-wiki	0.92	0.93	0.93
Extended-wiki	0.93	0.95	0.94

Table 3: Model performance on its test set (upper) and on the *Extended-wiki* test set (lower)

4.2 Identifying Wikipedia events in the news

Ultimately we are interested in the ability to detect events in the target domain of *news articles*. To validate performance over this domain, we used sentences from *News*-2019 and cross-referenced them with company events from Wikipedia. Specifically, we manually extracted events from 2019 from the Wikipedia pages of the companies in Table 2. For each event, we asked 3 annotators to mark all sentences from *News*-2019 which mention this event. In total, 26 of the Wikipedia events were mentioned in at least one sentence.

We then applied each of the three models to all the news sentences, and kept only the sentences that were classified as positive by the model. For each model, we measure the event recall rate as the fraction of Wikipedia events which are mentioned in at least one positively-classified sentence.

As expected, the recall rates of the Wikipedia-based models over the news data (Table 4) are lower than those achieved over Wikipedia data. This may be due to the difference in writing style between the two sources. Notably, even though *SentiFM-binary* was trained on news data, its recall is the lowest among the three models. This may be attributed to the mismatch between the event types in *SentiFM* and those in Wikipedia.

Sorting the positively-classified sentences by their model score, we also measure the average rank of the highest-scored mention of each event (Table 4). Clearly, the Wikipedia reference events do not fully cover all company-related events that occurred over this time period. Still, since we presume events mentioned in Wikipedia are relatively significant, we expect a good event-detection model to rank them among its top predictions.

Model	Recall	Avg. Rank
SentiFM-binary	0.38	153
S&P-wiki	0.73	21
Extended-wiki	0.77	19

Table 4: Model performance for identifying 26 Wikipedia events in the news data.

4.3 Identifying general events in the news

So far our experiments considered only Wikipedia events. However, there are likely numerous company-related news events that are not necessarily mentioned in the company's Wikipedia page. Thus, the question remains whether the Wikipedia-based models are able to detect such events as well. To this end, the top 20 model predictions of *SentiFM-binary* and *Extended-wiki* for the companies in Table 2 were annotated by three co-authors of this work. The guidelines were to determine whether a given sentence contains information which may have influence on the companys stock price, as such events presumably deserve the attention of a finance analyst. The annotation process was composed of two stages. First, each sentence was annotated by two labelers. Then, the sentences on which there was disagreement between the labelers (21% of the sentences) were annotated by a third annotator. Average agreement between the initial two annotators was 0.45 (Cohen's Kappa).

Table 5 shows the precision of the two models, compared to a baseline of randomly-selected sen-

Model	Precision
Random sentences	0.28
SentiFM-binary	0.70
Extended-wiki	0.74

Table 5: Average precision over the top-20 predicted events in the news evaluation data.

tences. The *Extended-wiki* model outperforms *SentiFM-binary*.

Finally, we wanted to analyze the diversity of events captured by the two models. For this purpose, we looked at the distribution of unique tokens in the top 200 predictions of each model, after filtering out stop words and the companies appearing in the list of Table 2. We sorted the remaining tokens by their frequency from highest to lowest, and computed the cumulative frequency as a function of the number of unique tokens. Figure 1 indicates that the top candidates of *Extended-wiki* capture a richer vocabulary than *SentiFM-binary*, which is dominated by a smaller group of tokens. For example, 20% of the tokens are covered by the 36 and 19 most frequent tokens in *Extended-wiki* and *SentiFM-binary*, respectively. Moreover, despite their similar precision values, the population of events captured by the two models is quite different - the overlap between their top candidates is less than 10% (18 out of 200 examples). This observation suggests that the models are complementary, and that there is potential benefit to combining them.

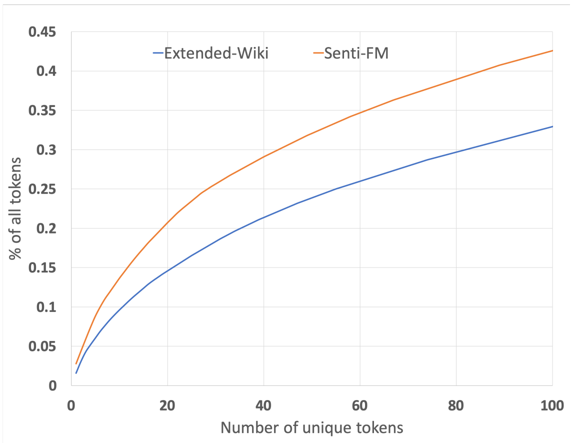

Figure 1: Cumulative token frequency over top model predictions.

5 Discussion

This paper focused on detecting ’important’ events in news articles, related to a specific company. We suggested to leverage information contained in Wikipedia to create weakly-labelled data, and proved the usefulness of the resultant classifier for the desired task. We believe that the results can be further improved by finding additional sources for weak-labels, e.g. by exploiting information from relevant knowledge bases.

The potential coverage of relevant events can be increased by retrieving articles which do not necessarily include the name of the considered company in their title. Extending our framework to pinpoint noteworthy events for a particular company, mentioned in articles that are not focused on that company, is a natural direction for future research. Such an extension will require adapting the weak labelled data and the corresponding classifiers to cope with an environment in which sentences are not necessarily relevant to the company.

References

James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 2003. Topic detection and tracking pilot study final report.

Ernest Arendarenko and Tuomo Kakkonen. 2012. Ontology-based information and event extraction for business intelligence. In *Proceedings of the 15th International Conference on Artificial Intelligence: Methodology, Systems, Applications*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1415–1425, Doha, Qatar. Association for Computational Linguistics.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI’15, pages 2327–2333. AAAI Press.

Mian Du, Lidia Pivovarova, and Roman Yangarber. 2016. Puls: natural language processing for business intelligence. In *Proceedings of the 2016 Workshop on Human Language Technology*, pages 1–8. Go to Print Publisher.

Ronen Feldman, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. 2011. The stock sonarsentiment analysis of stocks based on a hybrid approach. In *Twenty-Third IAAI Conference*.

Alexander Hogenboom, Frederik Hogenboom, Flavius Frasincar, Kim Schouten, and Otto Van Der Meer. 2013. Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications*, 64(1):27–52.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. In *DeRiVE@ ISWC*, pages 48–57. Citeseer.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska De Jong, and Emiel Caron. 2016. A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85:12–22.

Gilles Jacobs, Els Lefever, and Véronique Hoste. 2018. Economic event detection in company-specific news text. In *EcoNLP workshop at the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1–10. Association for Computational Linguistics.

Tuomo Kakkonen and Tabish Mufti. 2011. Developing and applying a company, product and business event ontology for text mining. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*.

Simone Merello, Andrea Picasso Ratto, Yukun Ma, Luca Oneto, and Erik Cambria. 2018. Investigating timing and impact of news on the stock market. In *2018 IEEE International Conference on Data Mining Workshops*.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1104–1112. ACM.

Samuel Rönnqvist and Peter Sarlin. 2017. Bank distress in the news: Describing events through deep learning. *Neurocomputing*, 264:57–70.

Boyi Xie, Rebecca Passonneau, Leon Wu, and Germán G Creamer. 2013. Semantic frames to predict stock price movement. In *Proceedings of the 51st annual meeting of the association for computational linguistics*, pages 873–883.

Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.