Economic Event Detection in Company-Specific News Text

Gilles Jacobs, Els Lefever and Véronique Hoste
Language and Translation Technology Team, Ghent University
Groot-Brittanniëlaan 45
9000 Ghent, Belgium
{gillesm.jacobs,els.lefever,veronique.hoste}@ugent.be

Abstract

This paper presents a dataset and supervised classification approach for economic event detection in English news articles. Currently, the economic domain is lacking resources and methods for data-driven supervised event detection. The detection task is conceived as a sentence-level classification task for 10 different economic event types. Two different machine learning approaches were tested: a rich feature set Support Vector Machine (SVM) set-up and a word-vector-based long short-term memory recurrent neural network (RNN-LSTM) set-up. We show satisfactory results for most event types, with the linear kernel SVM outperforming the other experimental set-ups.

1 Introduction

In the financial domain, the way companies are perceived by investors is influenced by the news published about those companies (Engle and Ng, 1993; Tetlock, 2007; Mian and Sankaraguruswamy, 2012). Tetlock (2007), for example, tried to characterize the relationship between the content of media reports and daily stock market activity, focusing on the immediate influence of the Wall Street Journal's 'Abreast of the Market' column on U.S. stock market returns. One of his major findings was that high levels of media pessimism robustly predict downward pressure on market prices.

To provide some insights into the way markets react to new information about companies, financial economists have conducted event studies. These event studies measure the impact of a specific event on the value of a firm (MacKinlay, 1997). They offer insight into the extent to which shareholders of acquired firms gain better returns during mergers, or examine the behavior of companies stock prices around events such as dividend announcements or stock splits. Studying the impact of specific events on the stock markets, however, is a labor-intensive process, starting with the identification of a given event, the estimation of abnormal returns to separate the general movement of stock returns from an individual stock return, followed by a number of statistical tests seeking evidence to support the event's economic significance. Since identifying news published about certain events in an automatic way enables researchers in the field of event studies to process more data in less time, and can consequently lead to new insights into the correlation between events and stock market movements, automatic techniques have been proposed to detect economic events in text.

Most of the existing approaches to the detection of economic events, however, are knowledge-based and pattern-based (Arendarenko and Kakkonen, 2012; Hogenboom et al., 2013; Du et al., 2016). These use rule-sets or ontology knowledge-bases which are largely or fully created by hand. The Stock Sonar project (Feldman et al., 2011) notably uses domain experts to formulate event rules for rule-based stock sentiment analysis. This technology has been successfully used in assessing the impact of events on the stock market (Boudoukh et al., 2016) and in formulating trading strategies (Ben Ami and Feldman, 2017). Other approaches conceptualize economic event detection as the extraction of event tuples (Ding et al., 2015) or as semantic frame parsing (Xie et al., 2013).

A drawback of knowledge-based information extraction methods is that creating rules and ontologies is a difficult, time-consuming process. Furthermore, defining a set of strict rules often re-

Proceedings of the First Workshop on Economics and Natural Language Processing, pages 1–10
Melbourne, Australia, July 20, 2018.

sults in low recall scores, since these rules usually cover only a portion of the many various ways in which certain information can be lexicalized. Thus, the need for flexible data-driven approaches, which do not require predefined ontological resources, arises. Rönnqvist and Sarlin (2017) provide an example of successful data-driven, weakly-supervised distress event detection based on bank entity mentions. Here, bank distress events are conceptualized as mentions of bank entities in a time-window and no typology classification is assigned. We are not aware of any published data-driven, supervised event detection approaches for the economic domain. However, in general domain event extraction, as embodied by projects such as ACE (Ahn, 2006) and ERE/TAC-KBP (Mitamura et al., 2016), supervised methods for extraction of event structures are predominant because of their promise of improved performance.

As discussed in Sprugnoli and Tonelli (2017), the definition of events in the field of information extraction differs widely. In this work, we employ a conceptualization of economic event detection as '*retrieving textually reported real-world occurrences, actions, relations, and situations involving companies and firms*'. Unlike other supervised data-driven 'event extraction' tasks such as in the ACE/ERE programs (Aguilar et al., 2014), we do not conceptualize events as structured schemata/frames, but more limited as textual mentions of real-world occurrences. The task presented here is often also referred to as event 'mention', 'nugget', or 'trigger' detection. The classification experiments described here are currently at the sentence-level, but our event annotation scheme is token-level.

In this paper, we tackle the task of economic event detection by means of a supervised machine learning approach, which we expect will be able to detect a wider variety of lexicalizations of economic events than pattern-based approaches. We consider economic event detection as a sentence-level multi-label classification task. The goal is to automatically assign the presence of a set of predetermined economic event categories in a sentence of a news article.

In previous work on the Dutch counterpart of this dataset, (Lefever and Hoste, 2016) has shown that SVM classification obtained decent results. Here, we compare two different machine learning approaches, viz. a rich feature set Support Vector Machine (SVM) approach, and a word-vector-based sequence long short-term memory recurrent neural network (RNN-LSTM) approach. We show that supervised classification is a viable approach to extract economic events, with the linear kernel SVM obtaining the best classification performance.

The remainder of this paper is structured as follows. In Section 2, we present the annotated corpus of financial news articles we constructed. Section 3 introduces our two classification approaches to economic event detection, followed by an overview of the results in Section 4. In Section 5, we conduct an error analysis to gain insights in the main shortcomings of the current approach. Section 6 formulates some conclusions and ideas for future work.

2 Data Description

In this section, we describe the SentiFM economic event dataset collection and annotation. The annotated dataset consists of an English and Dutch news corpus. While in this paper the focus is on English, we refer to Lefever and Hoste (2016) for a pilot study on Dutch event detection and a description of the Dutch event data. A reference to where to download the SentiFM dataset can be found in Section 7.

The goal of the SentiFM dataset is to enable supervised data-driven event detection in company-specific economic news. For English, we downloaded articles from the newspaper The Financial Times using the ProQuest Newsstand by means of keyword-search. The keywords were manually determined based on a subsample of random articles as being indicative to one of the event types. All articles were published between November 2004 and November 2013. The articles had at least one of the following seven companies in the title: Barclays, BHP, Unilever, British Land, Tesco, Vodafone, and BASF. These companies were selected because they are highly ranked in several market indexes while situated in different sectors/industries. This facilitates corpus collection as there is more news content due to the companies' status. Sectorial diversification is necessary to avoid specialization to one particular industry. For instance, six out of 10 highest market cap companies in the S&P500 index currently belong to the IT sector. In total, we collected 497 news articles

containing 2522 annotated company-specific economic events.

In the corpus, 10 types of company-specific economic events were manually identified:

Buy ratings A recommendation to purchase the security from an analyst. As event mentions, we include rating announcements, forecasts, performance, buy/sell/hold advice, and rating upgrades/downgrades/maintained.

Debt Event mentions pertaining to company debt and debt ratios. We include debt announcements, forecasts, increases, reductions, and restructuring.

Dividend A dividend is a distribution of a portion of a company's earnings paid to its shareholders. We include dividend announcements, forecasts, payments, none payments, stable yields, raises, and reductions.

Merger & acquisition Mergers and acquisitions refers to the consolidation of companies or assets involving at least two companies. We include announcements, forecasts, and cancellations of a merger/acquisition.

Profit Financial benefits that are realized when the amount of revenue exceeds expenses. We include declarations and forecasts of profit, positive and negative (losses) profit, lower than, higher than, as expected, increased, decreased, and stable profits.

Quarterly results Events pertaining to the quarterly report as a set of financial statements issued by a company. We include declaration of publication, forecasts, strong, weak, improved, declined, stable, better than, worse than, and as expected results.

Sales volume The quantity of goods and services sold over a certain period. We include declarations and predictions of sales volumes figures, increased, decreased, stable, better than, worse than, as expected sales volumes.

Share repurchase Share buyback events by a company including announcements and forecasts of share repurchases.

Target price Events on the projected price level of a security. We include announcements, forecasts, price raised, reduced, or maintained.

Turnover The number and frequency of securities traded over a certain period. We include declaration and prediction of turnover figures, increased, decreased, stable, worse than, better than, and as expected turnover.

These events and activities pertain to the specific instances of companies mentioned in the articles. The event typology was manually and iteratively constructed on a corpus subsample by an economic domain specialist. It is notable that this event typology overlaps largely with the independently created StockSonar typology (Feldman et al., 2011) and SPEED ontology (Hogenboom et al., 2013). These studies also used a manual and iterative approach to constructing a descriptive typology of company-specific economic events. It is unsurprising that the event types are highly similar.

Human annotators marked all mentions of each of these event types at the token level, using the Brat rapid annotation tool (Stenetorp et al., 2012), a web-based tool for text annotation. Events are linked to the earliest preceding company mentions with an 'about_company' relation (this relation is duplexed into 'acquiring_company' and 'target_company' for Merger & acquisition events). Discontinuous token spans and annotating multiple event types are allowed. Two annotators were involved in the first pass annotation phase. The gold standard was subsequently produced by an adjudication phase. The event annotation guidelines for English were ported from Dutch. To assess the reliability of the event annotations, we measured inter-annotator F_1-score on the events marked by 3 individual annotators in 10 articles from the Dutch corpus (consisting of 216 sentences and 3,202 tokens). With a cross-averaged F_1-score of 78.41% for the 3 annotator pairs, we can conclude that the annotated corpus is a reliable dataset for the task of economic event detection.

All texts were pre-processed (tokenized and sentence-splitted) using the LeTs Preprocess Toolkit (Van de Kauter et al., 2013).

The present task is sentence-level detection of event types, so one sentence instance can be assigned multiple event classes. Multiple labels are assigned to 3.81% ($n = 380$) of all sentence instances. An overview of the different event types and their total frequency is given in Table 1.

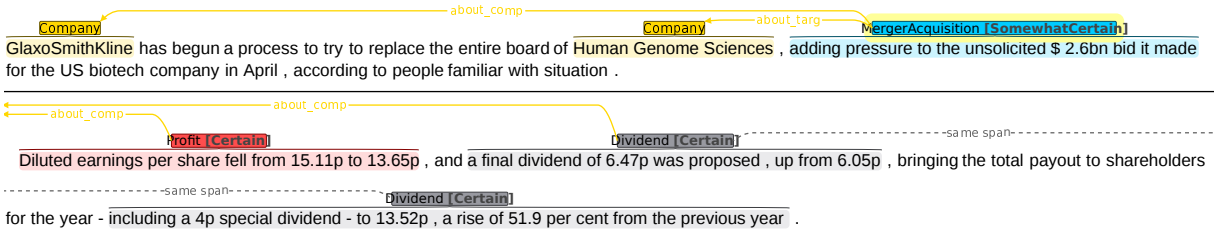

Figure 1: Annotated sentence examples from the Brat annotation tool.

Event type	Type ratio	# sentence instances
No Event	NA	7823 (75.62%)
BuyRating	9.00%	227 (2.19%)
Debt	2.38%	60 (0.58%)
Dividend	7.22%	182 (1.76%)
MergerAcquisition	10.03%	253 (2.45%)
Profit	25.81%	651 (6.29%)
QuarterlyResults	10.59%	267 (2.58%)
SalesVolume	19.31%	487 (4.71%)
ShareRepurchase	2.42%	61 (0.59%)
TargetPrice	3.73%	94 (0.91%)
Turnover	9.52%	240 (2.32%)
Total	2522 events/10345 sentences (24.38%)	

Table 1: Event type distribution in the SentiFM English economic dataset and sentence level counts (as used in experiments).

3 Experimental Set-up

For this study, the task of economic event detection is conceived as a sentence-level classification task. We decided on comparing two different machine learning approaches: an SVM approach requiring offline feature engineering, and a word-vector-based sequence RNN-LSTM approach.

The SVM approach incorporates a rich feature set with syntactic and lexical feature engineering. We built one SVM classifier per event, predicting whether the event was present in the sentence or not, in effect recasting the problem as a one-vs-rest binary classification task for each class. The RNN-LSTM is tested both as a multi-label single model classifier and a one-vs-rest set-up.

Performance estimation is done on a random hold-out test split (10%), whereas cross-validation experiments were carried out on the hold-in set (train set of 90%) for both hyper-parameter optimization and validation of generalization error.

Per event type, precision, recall, and F_1-score are reported for each approach on the hold-out test set. We do not report accuracy because it is not an apt performance indicator in the case of class imbalance. Cross-validation results on the training set are not reported due to space constraints, but followed the same trends as the reported test results with no indication of over-fitting.

3.1 Support Vector Machines

For the first set of experiments, a support vector machine model was built per economic event type in a one-vs-rest set-up applying two different kernels: (1) the *linear* kernel with default LIBSVM hyperparameters and (2) a hyper-parameter optimized version of the *RBF* kernel. The optimal weights for the c and g parameters for the RBF kernel were obtained by means of a 5-fold grid search on the training data for each event type. All experiments were carried out with the LIBSVM package (Chang and Lin, 2011).

In a first step, the data set was linguistically pre-processed by means of the LeTs Preprocessing Toolkit (Van de Kauter et al., 2013), which performs lemmatization, part-of-speech tagging, and named entity recognition. Consequently, a set of lexical and syntactic features were constructed on the basis of the pre-processed data.

Lexical features The following lexical features were constructed: token n-gram features (unigrams, bigrams and trigrams), character n-gram features (trigrams and fourgrams), lemma n-gram features (unigrams, bigrams and trigrams), disambiguated lemmas (lemma + associated PoS-tag), and a set of features indicating the presence of numerals, symbols, and time indicators (e.g. *yesterday*).

Syntactic features As syntactic features, we extracted three features for each PoS-category: binary (presence of category in the instance), ternary (category occurs 0, 1 or more times in the instance)

and total number of occurrences of the respective PoS-label. In addition, similar features (binary, ternary, and frequency) were extracted for 6 different Named Entity types: person, organization, location, product, event, and miscellaneous.

3.2 Recurrent Neural Net LSTM

The RNN-LSTM approach was implemented using the Keras neural networks API (Chollet et al., 2015) with TensorFlow as back-end (Abadi et al., 2015). We employ a straightforward neural architecture: the input-layer is a trainable embedding layer which feeds into an LSTM block. The LSTM block is connected to an output layer with a sigmoid activation function. Bi-directionality of the LSTM-layer is tested in hyper-parameter optimization. We use the Adam optimization algorithm with binary cross-entropy loss function. The embedding layer turns positive integers, in our case hold-in set token indexes, in dense vectors with fixed dimensionality. An existing word embedding matrix can be used in the input-layer which tunes pre-trained word vectors.

Three embedded inputs were tested with the multi-label set-up: 200 dimensional GloVe (Pennington et al., 2014) word vectors trained on the hold-in set, 300 dimensional GloVe vectors trained on a 6 billion token corpus of Wikipedia (2014) + Gigawords5B[1] (henceforth, 6B corpus), and no pre-trained embeddings. The latter means our classifier trains embedded word-representations (with a fixed dimensionality of 200) itself based on the token sequences of the hold-in set. We evaluated our own GloVe models on an analogy quality assessment task provided with the word2vec source code[2]. We picked the highest dimensional word vector model from the top ten ranking on the analogy task. We excluded lower dimensional vectors because preliminary tests have shown that higher dimensional pre-trained vectors obtained better scores.

We first tested a multi-label and subsequently a one-vs-rest approach in which a binary classifier is trained for each economic event class. The multi-label approach requires one full training iteration compared to one for each of the 10 classes in one-vs-rest and is much less computationally expensive. For this reason we limit the tested word-vector inputs to the 6B GloVe word vectors in the one-vs-rest approach. These input vectors outperformed others in the multi-label experiments considering F_1-score per label, as well as the hold-in set vectors in preliminary tests using limited iteration randomized search testing.

The following model hyper-parameters were set by 3-fold random search with 32 iterations. The winning hyper-parameters are chosen by prevalence-weighted macro-averaged F_1-score over the multi-label prediction.

RNN-LSTM hyper-parameter	Setting
Bidirectionality on LSTM layer	Enabled or disabled
LSTM unit size	$d \in \{134, 268, 536\}$
Dropout rate	$r \in \{0.0, 0.2\}$
Recurrent dropout rate	$rr \in \{0.0, 0.2\}$
Batch size	$b \in \{64, 128, 256, 512\}$
Training epochs	$e \in \{32, 64, 128\}$

Table 2: RNN-LSTM model hyper-parameters.

In the next section, the best model hyper-parametrization as determined by prevalence-weighted macro-averaged F_1-score will be discussed.

4 Experimental Results

We present per class results of the SVM one-vs-rest approach in Table 3 and for the RNN-LSTM in Table 4 for multi-label and Table 5 for one-vs-rest. Even though our classifiers were trained on a limited amount of data, we obtain satisfactory results for the detection of company-specific economic events for most event types. Overall precision scores are promising, especially for the SVM-based approach and the RNN-LSTM with hold-in trained word vectors.

The best overall results are obtained by the linear kernel SVM which obtained far better recall than any other model. The one-vs-rest RNN-LSTM systems comes in at a close second and outperforms its multi-label counterparts by a large margin. Including lexical and syntactic features seems to be worthwhile when compared to the straight-forward word vector/token sequence approach used with the RNN-LSTM.

The best RNN-LSTM multi-label model is outperformed by the linear kernel SVM approach and is on par with the optimized RBF kernel approach. The pre-trained GloVe vectors trained on our own dataset performed best out of the three input meth-

[1] https://nlp.stanford.edu/projects/glove/

[2] https://code.google.com/archive/p/word2vec/

ods with a prevalence-weighted macro-averaged F_1-score of 0.66 on hold-out. The GloVe vectors trained on the 6B corpus obtain worse precision but slightly better recall, resulting in a comparable F_1-score of 0.64. The 6B GloVe inputs obtain better scores on more classes, but their macro-averaged score is hurt by not detecting any of the Debt class instances. Not feeding pre-trained embeddings to our network shows the worst performance of all classifiers (F_1-score of 0.54).

Event type	Precision	Recall	F_1-score
Linear kernel one-vs-rest			
BuyRating	**0.95**	**0.91**	**0.93**
Debt	**0.50**	**1.00**	**0.67**
Dividend	**0.62**	**0.73**	**0.67**
MergerAcquisition	**0.56**	**0.40**	**0.47**
Profit	0.75	0.74	0.75
QuarterlyResults	0.82	0.53	0.64
SalesVolume	0.88	**0.75**	**0.81**
ShareRepurchase	**1.00**	**0.50**	**0.67**
TargetPrice	**1.00**	**0.75**	**0.86**
Turnover	**0.91**	**0.77**	**0.83**
avg	**0.80**	**0.71**	**0.73**
Optimized RBF one-vs-rest			
BuyRating	**0.95**	**0.91**	**0.93**
Debt	**0.50**	**1.00**	**0.67**
Dividend	0.54	0.64	0.58
MergerAcquisition	0.00	0.00	0.00
Profit	**0.80**	**0.76**	**0.78**
QuarterlyResults	**0.83**	**0.56**	**0.67**
SalesVolume	**0.94**	0.65	0.77
ShareRepurchase	**1.00**	**0.50**	**0.67**
TargetPrice	**1.00**	**0.75**	**0.86**
Turnover	0.87	**0.77**	0.82
avg	0.74	0.65	0.67

Table 3: Hold-out test precision, recall, and F_1-scores per type for the linear and optimized RBF kernels of the feature-engineered SVM one-vs-rest approach. **Boldface** indicates best performance within the SVM set-up. Underline indicates best of all tested systems.

In both one-vs-rest approaches, we trade off computation time for performance compared to multi-label systems. This approach also has the advantage that a separate classifier is produced for each class. At prediction time, we can thus trivially apply the best available classifier algorithm from both the SVM and RNN-LSTM systems for each class. When combining classifiers in this manner an average score of 0.81% precision, 0.74% recall, and 0.75% F_1-score is reached, improving over the best scoring single algorithm system.

Event type	Precision	Recall	F_1-score
Hold-in set GloVe multi-label			
BuyRating	**0.91**	**0.91**	**0.91**
Debt	**1.00**	**0.50**	**0.67**
Dividend	0.50	0.36	0.42
MergerAcquisition	0.32	0.24	0.27
Profit	0.75	**0.81**	0.78
QuarterlyResults	**0.87**	0.38	0.53
SalesVolume	**0.92**	0.67	0.77
ShareRepurchase	0.80	**0.67**	0.73
TargetPrice	**1.00**	**0.50**	**0.67**
Turnover	**0.95**	0.69	0.80
avg	**0.80**	0.57	**0.66**
6B corpus GloVe multi-label			
BuyRating	0.86	0.82	0.84
Debt	0.00	0.00	0.00
Dividend	0.50	**0.55**	0.52
MergerAcquisition	**0.40**	**0.32**	**0.36**
Profit	0.82	0.79	**0.81**
QuarterlyResults	0.77	**0.68**	**0.72**
SalesVolume	0.84	**0.73**	**0.78**
ShareRepurchase	**1.00**	**0.67**	**0.80**
TargetPrice	0.75	0.75	**0.75**
Turnover	0.90	**0.73**	**0.81**
avg	0.68	**0.60**	0.64
No pre-trained word vectors multi-label			
BuyRating	0.81	0.59	0.68
Debt	0.33	0.50	0.40
Dividend	**0.75**	**0.55**	**0.63**
MergerAcquisition	0.21	0.12	0.15
Profit	**0.83**	0.33	0.47
QuarterlyResults	0.67	0.35	0.46
SalesVolume	0.86	0.61	0.71
ShareRepurchase	0.60	0.50	0.55
TargetPrice	**1.00**	**0.50**	**0.67**
Turnover	0.88	0.58	0.70
avg	0.69	0.46	0.54

Table 4: Hold-out test precision, recall, and F_1-scores per type for RNN-LSTM for different word vector input. **Boldface** indicates best performance within RNN-LSTM multi-label approach. Underline indicates best of all systems.

5 Error Analysis

We performed a detailed error analysis on the best classifier in order to gain insights in the main shortcomings of the current approach.

Event type	Precision	Recall	F_1-score
6B corpus GloVe one-vs-rest			
BuyRating	0.88	0.95	0.91
Debt	0.50	0.50	0.50
Dividend	0.55	0.55	0.55
MergerAcquisition	0.58	0.44	0.50
Profit	0.81	0.74	0.77
QuarterlyResults	0.84	0.47	0.60
SalesVolume	0.81	0.76	0.79
ShareRepurchase	0.75	0.50	0.60
TargetPrice	1.00	1.00	1.00
Turnover	0.94	0.65	0.77
avg	0.77	0.66	0.70

Table 5: Hold-out test precision, recall, and F_1-scores per type for the one-vs-rest RNN-LSTM with 6B GloVe corpus word vectors. Underline indicates best of all systems.

In general, we noticed that a fair amount of event types are characterized by strong lexical clues. As an example, we can cite the following *BuyRating* example, where the unigrams *upgraded, hold* and *buy* can be considered lexical indicators of this category:

(1) *Repair and maintenance group Homeserve, which also reports on Friday, rose 2.8 per cent to pound(s)17.54 after RBS upgraded from "hold" to "buy".*

Most of the event categories, however, show a **large variety of possible lexicalizations**. This is illustrated by examples 2 and 3 for SalesVolume, examples 4 and 5 for ShareRepurchase, and examples 6, 7 and 8 for Turnover:

(2) *This could raise doubts about Vodafone's target of reaching 10m subscribers by the end of the current financial year.*

(3) *It will increase the number of Barclays' customers in France by 25 per cent.*

(4) *Last week, Engelhard scotched hopes of a negotiated deal with BASF, after three months of ding-dong talks, unveiling instead a defence strategy centred on a planned Dollars 1.2bn share buy-back at Dollars 45 a share.*

(5) *So far, free cash flow has been used to finance share buybacks and dividend increases.*

(6) *The mobile network reseller also forecast mid-teen percentage growth in service revenue, far better than most analysts had expected in a tough UK market.*

(7) *However, revenues from voice and text fell in the period.*

(8) *Arun Sarin yesterday sought to dispel fears about slowing revenue growth at Vodafone by saying the mobile phone company would make more acquisitions in Africa and Asia.*

In addition, some of the lexical clues are **ambiguous** in the sense that they occur with various event categories. This is for instance the case for *buy*, which can be informative to predict the *BuyRating* (Example 9) as well as the *MergerAcquisition* (Example 10) event categories:

(9) *EMI eased 1.19 per cent to 252p in spite of a buy recommendation from Deutsche Bank.*

(10) *G4S led the blue-chip risers amid continued speculation that shareholders may block its pound(s)5.2bn deal to buy ISS, the office cleaning group.*

In future work, we intend to improve the lexical coverage by increasing the data set size, but also by adding semantic knowledge from structured resources. The following *BuyRating* event has not been detected, but this could be the case if *downgrade* could be correctly identified as a lowering in rating (viz. moving the rating from a buy to a hold, or a hold to a sell).

(11) *The weak oil price and a downgrade from RBS did the damage.*

The same holds for the following *MergerAcquisition* example, where *takeover* should be semantically clustered together with *acquire, acquisition*, etc.

(12) *News that Hewlett-Packard was preparing a $10bn takeover offer for the software maker came too late for London traders to react.*

Furthermore, for some event categories, the evaluation set is too limited to draw reliable conclusions. As can be noticed in Table 6, which lists the number of instances per category in the test set, the *Debt* and *TargetPrice* evaluation sets contain less than five test items. Collecting and annotating

Event type	# test instances
BuyRating	22
Debt	2
Dividend	11
MergerAcquisition	25
Profit	58
QuarterlyResults	34
SalesVolume	51
ShareRepurchase	6
TargetPrice	4
Turnover	26
Total	994

Table 6: Economic event type distribution in the evaluation set.

additional data should lead to a better coverage for all event categories.

Another source of wrong classification was due to annotation errors in the data set. This is illustrated by Example 13, where the *buyRating* event was not labeled, and Example 14, where the *dividend* label was lacking:

(13) *Morgan Stanley repeated " underweight " advice in a note sent to clients overnight .*

(14) *ECS argues Verizon Wireless is a "passive investment" for Vodafone because it last received a dividend in 2004-05, worth Pounds 923m.*

Finally, the error analysis also revealed that some strong lexical clues are not always picked up by the classifier to correctly predict the event category. We assume this might be due to the very large feature space, as the SVM classifier is now trained on more than 300,000 bag-of-words features. In addition to the skewed data distribution, this large feature set makes the machine learning task very challenging. Therefore, we expect the classification performance to improve by performing feature selection to determine which sources of information are most relevant for solving this learning task. Having a good mechanism to select informative bag-of-words features should allow to correctly predict the economic event in case lexical clues are present in the sentence. In this case, the following sentence should definitely be classified as a *MergerAcquisition* event:

(15) *The acquisition would give CIBC control of FirstCaribbean with a stake of 87.4 per cent .*

6 Conclusions

This paper presents a dataset and classification experiments for company-specific economic event detection in English news articles. Currently, there is little to no data resources and experiments for supervised, data-driven economic event extraction. The task was approached as a supervised classification approach and two different machine learning algorithms, an SVM and RNN-LSTM learner, were tested for the task. For our SentiFM event dataset, we have shown that a feature-engineered SVM approach obtains better performance than an RNN-LSTM word-vector system. The results show good classification performance for most event types, with the linear kernel SVM outperforming the RBF kernel SVM and RNN-LSTM set-ups. We demonstrated that data-driven approaches obtain good recall and can capture variation in lexicalizations of events to a satisfactory extent.

There is still plenty of room for improvement: more annotated data and augmentative resources are needed to further offset ambiguous event expressions. In future work, we will design a more fine-grained event detection model that also extracts the token span of the event below the sentence level. Furthermore, we will work on detecting subevents currently contained in our annotations: e.g. BuyRating: outperform, hold, sell, upgrade, etc. As feature engineering seems to pay off for the extraction of economic events, we will integrate additional linguistic information by adding semantic knowledge from structured resources such as DBpedia and dedicated ontologies for economics (e.g. the NewsEvent ontology (Lösch and Nikitina, 2009) and derived CoProE ontology (Kakkonen and Mufti, 2011)) as well as syntactic information extracted from dependency parses.

7 Data availability

The SentiFM company-specific economic news event dataset and annotation guidelines as used in this paper are available for download from `https://osf.io/enu2k/` (Van de Kauter et al., 2018). This repository also contains replication data including the vectorized feature data and test split.

8 Acknowledgment

The work presented in this paper was carried out in the framework of the SENTiVENT project aspirant grant of the Research Foundation - Flanders.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems.

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53.

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8. Association for Computational Linguistics.

Ernest Arendarenko and Tuomo Kakkonen. 2012. Ontology-Based Information and Event Extraction for Business Intelligence. In *Artificial Intelligence: Methodology, Systems, and Applications*, volume 7557 of *Lecture Notes in Computer Science*, pages 89–102. Springer.

Zvi Ben Ami and Ronen Feldman. 2017. Event-based trading: Building superior trading strategies with state-of-the-art information extraction tools. SSRN Working Paper 2907600.

Jacob Boudoukh, Ronen Feldman, Shimon Kogan, and Matthew P Richardson. 2016. Information, trading, and volatility: Evidence from firm-specific news. SSRN Working Paper 2193667.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2:27:1–27:27.

François Chollet et al. 2015. Keras. `https://github.com/fchollet/keras`.

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.

Robert F. Engle and Victor K. Ng. 1993. Measuring and Testing the Impact of News on Volatility. *The Journal of Finance*, 48(5):1749–1778.

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.

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*, page 24. ACM.

Marjan Van de Kauter, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Véronique Hoste. 2013. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120.

Marjan Van de Kauter, Gilles Jacobs, Els Lefever, and Vronique Hoste. 2018. SentiFM company-specific economic news event dataset (English).

Els Lefever and Veronique Hoste. 2016. A classification-based approach to economic event detection in dutch news text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*, pages 330–335. European Language Resources Association (ELRA).

Uta Lösch and Nadejda Nikitina. 2009. The newsEvents ontology: an ontology for describing business events. In *Proceedings of the 2009 International Conference on Ontology Patterns-Volume 516*, pages 187–193. CEUR-WS. org.

A. Craig MacKinlay. 1997. Event Studies in Economics and Finance. *Journal of Economic Literature*, 35(1):13–39.

Ghulam Mujtaba Mian and Srinivasan Sankaraguruswamy. 2012. Investor Sentiment and Stock Market Response to Earnings News. *The Accounting Review*, 87(4):1357–1384.

Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. 2016. Overview of TAC-KBP 2016 event nugget track. In *Text Analysis Conference*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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

Rachele Sprugnoli and Sara Tonelli. 2017. One, no one and one hundred thousand events: Defining and processing events in an inter-disciplinary perspective. *Natural Language Engineering*, 23(4):485–506.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, pages 102–107, Avignon, France.

Paul C. Tetlock. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139–1168.

Boyi Xie, Rebecca J. 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 (Volume 1: Long Papers)*, pages 873–883. Association for Computational Linguistics.