A Classification-based Approach to Economic Event Detection in Dutch News Text

Els Lefever and Véronique Hoste
LT3 Language and Translation Technology Team
Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
els.lefever, veronique.hoste@ugent.be

Abstract

Breaking news on economic events such as stock splits or mergers and acquisitions has been shown to have a substantial impact on the financial markets. As it is important to be able to automatically identify events in news items accurately and in a timely manner, we present in this paper proof-of-concept experiments for a supervised machine learning approach to economic event detection in newswire text. For this purpose, we created a corpus of Dutch financial news articles in which 10 types of company-specific economic events were annotated. We trained classifiers using various lexical, syntactic and semantic features. We obtain good results based on a basic set of shallow features, thus showing that this method is a viable approach for economic event detection in news text.

1. Introduction

In the financial domain, the way companies are perceived by investors is highly influenced by the news (Engle and Ng, 1993; Tetlock, 2007; Mian and Sankaraguruswamy, 2012). This news can be either linked to macroeconomic factors (like the overall economy and industry conditions), news from the geo-political front or company-specific factors (like the financial condition or announcements about dividend payments or stock-splits). Good news would tend to lift the market, while on the other hand, bad news would tend to dampen the markets growth, although it has also been observed (Engle and Ng, 1993) that there exists an asymmetric impact on the stock market from news, whereby negative news shows a greater impact on volatility than positive news. News might also have different effects on the market return depending on the overall state of the market. Good news in a bullish (positive) market showing confidence may be reacted on very differently from the same piece of news arrived during a bearish (pessimistic) market.

Event studies examine the behavior of companies' stock prices around certain economic events such as dividend announcements, stock splits, mergers and acquisitions, etc. (see MacKinlay (1997) for an overview of event study methods). The vast literature on event studies written over the past several decades has become an important part of financial economics (Kothari and Warner, 2007). In a corporate context, the usefulness of event studies arises from the fact that the magnitude of abnormal performance at the time of an event provides a measure of the (unanticipated) impact of this type of event on the wealth of the firms claimholders. Event studies also serve an important purpose in capital market research as a way of testing market efficiency. Studying the impact of specific events on the stock markets, however, is a labor-intensive process. This has prompted the use of text mining techniques for the automatic detection of economic events in news text. 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. Furthermore, automatic event detection can be of use for various financial applications such as algorithmic trading (Hogenboom, 2012).

Many of the existing approaches to the detection of economic events are pattern-based (i.e. rule-based). Appelt et al. (1993), for instance, apply a system which makes use of a domain pattern recognizer for the detection of joint venture events in English and Japanese text. Drury and Almeida (2011) also make use of phrase extraction patterns for the identification of business event phrases in news stories. Other pattern-based methodologies for the detection of certain types of economic events have been adopted by Arendarenko and Kakkonen (2012) and Hogenboom et al. (2013), who developed resp. the BEECON and SPEED systems. Both systems make use of domain ontologies and manually defined lexicon-semantic rules (e.g. "Company buys Company") for event pattern recognition. A drawback of rule-based information extraction methods is that creating rules is a difficult process, which requires a considerable amount of domain knowledge. Furthermore, defining a set of strict rules often results in low recall scores, since these lexico-semantic rules usually cover only a portion of the many various ways in which certain information can be lexicalized. Finally, rule-based systems are not easily portable to other languages and domains (or, in the case of event detection, to other types of events). 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. Whereas many researchers (Ahn, 2006; Hardy et al., 2006; Ji and Grishman, 2008) have successfully applied machine learning techniques for event extraction (and coreference) tasks, we are not aware of studies focusing on economic events that employ machine learning methods without making use of event extraction rules.

For this paper, we investigated the viability of a classification-based approach to economic event de-

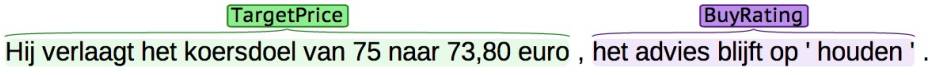

Figure 1: Annotation of economic events in brat.

tection based on an annotated corpus of Dutch financial news articles. We aimed at the detection of 10 types of company-specific events on the sentence level. The use of several lexical, syntactic and semantic features was investigated. We show that for the majority of event types, our classification-based approach obtains good results, even when using classifiers based on a limited amount of training data and incorporating only shallow lexical, syntactic and semantic features.

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. describes the set-up used to build the classifiers for economic event detection. In Section 4., the evaluation of these different classifiers is discussed. Finally, Section 5. formulates some conclusions and ideas for future work.

2. Corpus construction

In order to evaluate the viability of a classification-based approach to economic event detection, we created a corpus of Dutch news articles published in the Belgian financial paper De Tijd (between November 1st, 2004 and November 1st, 2013). Since our research is currently aimed at the detection of company-specific events, we collected articles reporting news on a predefined list of companies. We more specifically selected news texts of which the headlines mentioned at least one of the following 7 companies listed in the Bel 20 index: Delhaize, Belgacom, KBC, AB InBev, Solvay, Bekaert and Cofinimmo. The corpus used for the experiments described in this paper consists of 126 articles, containing 3,480 sentences (or 52,559 tokens) in total. Sentence splitting and tokenization of the corpus was performed using the LeTs Preprocess toolkit (Van de Kauter et al., 2013).
In the corpus, 10 types of company-specific economic events were manually identified, namely events regarding:

1. Profit
   e.g. *Bedrijfswinst groeit dit jaar 40 procent.* (English: Operating income will grow 40 percent this year.)

2. Turnover
   e.g. *Vicson moet voor 6 à 7 procent omzetgroei zorgen.* (English: Vicson has to ensure 6 to 7 percent revenue growth.)

3. Sales volume
   e.g. *Het wereldwijde volume verkocht bier zou nagenoeg stabiel blijven (+0,1%)* (English: Global beer sales volume should remain virtually stable (+0.1%))

4. Quarterly results
   e.g. *Maar volgens hem houden beleggers al rekening met zwakke tweedekwartaalresultaten.* (Engels: But according to him, investors are already accounting for weak second quarter results.)

5. Debt
   e.g. *AB InBev lost 2,5 miljard euro versneld af* (English: AB InBev to repay 2.5 billion at accelerated pace)

6. Target prices
   e.g. *Degroof verlaagde zijn koersdoel tot 80 euro.* (English: Degroof lowered its target price to 80 euros)

7. Buy ratings
   e.g. *Het koopadvies luidt 'verhogen'* (English: The buy recommendation says 'raise')

8. Dividend
   e.g. *De vraag is of de belegger daarvan een graantje meepikt* (English: The question is whether investors also benefit from this)

9. Share repurchase
   e.g. *Sommige analisten speculeren op een nieuw aandeleninkoopprogramma.* (English: Some analysts are counting on a new share buyback program.)

10. Merger/Acquisition (M&A) activity
    e.g. *Voorts kijken analisten uit naar de besparingen door de fusie met Anheuser-Busch.* (English: Furthermore, analysts are looking forward to the savings from the merger with Anheuser–Busch.)

As can be noticed from the examples, economic events can be expressed by means of simple lexicalizations (e.g. substantive "buy ratings"), whereas in other cases the event is rendered in a more implicit way. If we consider for instance the *dividend* example, the reader needs to have some financial background knowledge in order to correctly identify the reference to dividends. An additional challenge is caused by the very productive compounding system in Dutch, resulting in a high number of compounds written as one orthographic unit (e.g. *tweedekwartaalresultaten* (second quarter results)).

Human annotators marked all mentions of each of these event types using the brat rapid annotation tool (Stenetorp et al., 2012), a web-based tool for text annotation. Figure 1 shows an example of a sentence in which both a *TargetPrice*[1] and a *BuyRating*[2] event are annotated. In 1,032 of the 3,480 sentences, at least 1 event was marked; 109 of these sentences contain more than 1 event. To assess the reliability of the event annotations, we measured inter-annotator agreement on the events marked by 3 individual

[1]English translation: *He lowers the target price from 75 to 73.80 euros*

[2]English translation: *a 'hold' rating is maintained*

annotators in 10 articles from the corpus (consisting of 216 sentences and 3,202 tokens). For each annotator pair, we calculated F-score (van Rijsbergen, 1979) by considering the annotations made by the first annotator as the reference set, and the annotations of the second annotator as the test set. With an average F-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.
Based on the manual event annotations, we created one dataset per event type in which each sentence received a binary label indicating whether or not it contains a mention of the event type in question. Table 1 shows the number of positive instances (i.e. sentences) for each of the event types.

3. Experimental set-up

We conceived the economic event detection task as a binary classification task and built dedicated classifiers for the detection of mentions of each event type (on the sentence level). We experimented with different combinations of various lexical, syntactic and semantic features. The classifiers were trained using the LIBSVM package (Chang and Lin, 2011) with standard parameter settings (linear kernel function with $c = 1.0$).

Prior to feature extraction, the following **preprocessing** steps were taken: part-of-speech tagging, lemmatization and named entity recognition using the LeTs Preprocess toolkit (Van de Kauter et al., 2013) and dependency parsing by means of the Alpino parser (Bouma et al., 2001). Subsequently, a set of lexical, syntactic and semantic **features** was extracted from the linguistically preprocessed corpus.
A first set of **lexical** features contains various binary features:

- Token n-gram features: BOW features for all token unigrams, bigrams, trigrams and four-grams.
- Character n-gram features: BOW features for all character bigrams, trigrams and four-grams (within tokens).
- Features indicating the presence of numerals, symbols (e.g. %, $) and time indicators, which are detected using a list of words referring to a certain point in time (e.g. 'gisteren' - 'yesterday'). We expect these types of lexical items to often occur in conjunction with economic events, for instance in the sentence *KBC maakte vrijdag bekend dat het op basis van voorlopige cijfers verwacht dat de winst in 2004 met ruim 55 procent groeide*[3].
- Lemma features: BOW features for all lowercased lemmas.
- Disambiguated lemma features: BOW features for all lowercased lemma + PoS tag pairs. These are more abstract representations of the lemma features.

[3]English translation: *On Friday, KBC announced that based on provisional figures, it expects its profit to have grown by more than 55 percent in 2004.*

As we also wanted to explore whether the performance of the lemma and disambiguated lemma features could be improved by further abstraction based on semantic relation information, we also used the morphosyntactic hypernym detection module of Lefever et al. (2014) to identify hypernym relations between simple and compound nouns such as 'winst' - 'bedrijfswinst' ('profit' - 'operating profit'). Using this module, we detected possible hypernyms for each noun in the corpus and incorporated the hypernyms of each lemma into the feature vectors.

As **syntactic** information, we integrated 4 features for each main part-of-speech category: binary (category is present or not), ternary (category occurs 0, 1 or more times), absolute (number of occurrences for the category) and frequency (frequency of the category). We furthermore also included BOW features for all dependency relations. As proposed by Joshi and Penstein-Rosé (2009) (for the task of opinion mining), we do not only consider the lexicalized dependency relations, but also try to generalize the dependency features by backing off the head and/or modifier to its PoS tag or lemma. This resulted in 9 types of dependency relation features.

Finally, we also incorporated shallow **semantic** information into the feature vectors. Four different named entity features were stored for each of 6 NE types (person, organization, location, product, event and miscellaneous) and for all types considered together: binary (presence of NEs), absolute (number of NEs), absolute tokens (number of tokens contained in a NE) and frequency of tokens (frequency of tokens contained in a NE).

4. Evaluation

For each of the 10 event types, we trained SVM classifiers on the annotated news corpus using different combinations of the feature groups discussed in Section 3., ranging from basic token n-gram features to more complex lexical, syntactic and semantic feature groups. The leave-one-out cross-validation results of the incremental evaluation process can be found in Table 2, which lists all precision, recall and F-scores. We gradually added one feature group at a time to the previous experimental set-up: for the first experiment we only used token n-grams, for the second experiment token n-grams + character n-grams, for the third experiment token n-grams + character n-grams + numerals, symbols and time indicators, etc. For each event type, the best scores are indicated in bold.

Although our classifiers incorporate rather shallow features and were trained on a limited amount of data, we obtain good results for the extraction of most company-specific economic events. The experimental results for the different event types demonstrate high precision scores, especially for the classifiers trained using shallow token and character n-gram features. Recall scores are low for some event types, but based on the results for the other events, we believe a major improvement of these scores is possible by relying on LOD information sources and extending the annotated dataset. The performance of the lemma and

Event type	Profit	Turnover	Sales Volume	Quarterly Results	Debt	Target Price	Buy Rating	Dividend	Share Repurchase	M&A
# Sentences	294	208	97	126	51	80	79	79	15	123

Table 1: Number of sentences in the corpus in which mentions of the different event types occur (of 3,480 sentences in total).

Features		Profit	Turnover	Sales Volume	Quarterly Results	Debt	Target Price	Buy Rating	Dividend	Share Repurchase	M&A
token n-grams	*Prec*	**84.88**	**79.01**	**85.71**	57.97	**100.00**	**96.05**	94.34	**95.65**	0.00	67.74
	Rec	49.66	61.54	18.56	31.75	23.53	91.25	63.29	27.85	0.00	17.07
	F	62.66	69.19	30.51	41.03	38.10	93.59	75.76	43.14	0.00	27.27
+ char. n-grams	*Prec*	77.05	73.87	65.57	62.50	76.92	93.98	94.52	79.63	**100.00**	66.67
	Rec	63.95	70.67	41.24	51.59	39.22	**97.50**	87.34	54.43	**13.33**	**39.02**
	F	69.89	72.24	50.63	56.52	51.95	**95.71**	90.79	64.66	**23.53**	49.23
+ numerals / symbols / time indicators	*Prec*	76.95	73.10	65.08	64.08	76.92	93.98	94.52	78.95	100.00	66.67
	Rec	63.61	69.23	**42.27**	52.38	39.22	97.50	87.34	56.96	13.33	39.02
	F	69.65	71.11	**51.25**	57.64	51.95	95.71	90.79	66.18	23.53	49.23
+ lemmas	*Prec*	76.42	74.49	63.49	64.76	76.92	93.98	94.52	80.36	100.00	67.61
	Rec	63.95	70.19	41.24	53.97	39.22	97.50	87.34	56.96	13.33	39.02
	F	69.63	72.28	50.00	58.87	51.95	95.71	90.79	66.67	23.53	49.48
+ dis. lemmas	*Prec*	75.50	73.23	61.67	63.81	74.07	93.98	94.52	80.36	100.00	68.57
	Rec	63.95	69.71	38.14	53.17	39.22	97.50	87.34	56.96	6.67	39.02
	F	69.24	71.43	47.13	58.01	51.28	95.71	90.79	66.67	12.50	**49.74**
+ hypernyms	*Prec*	75.79	73.00	61.90	65.09	74.07	93.83	94.52	79.31	100.00	66.67
	Rec	**64.97**	70.19	40.21	54.76	39.22	95.00	87.34	58.32	13.33	39.02
	F	**69.96**	71.57	48.75	59.48	51.28	94.41	90.79	67.15	23.53	49.23
+ Part-of-Speech	*Prec*	76.00	75.00	65.00	65.09	80.00	93.83	94.44	79.66	100.00	63.01
	Rec	64.63	70.67	40.21	54.76	**47.06**	95.00	86.08	59.49	13.33	37.40
	F	69.85	72.77	49.68	59.48	**59.26**	94.41	90.07	68.22	23.53	46.94
+ dependency	*Prec*	77.64	76.24	72.55	65.38	84.00	93.75	94.52	79.69	100.00	69.23
	Rec	62.59	**74.04**	38.14	53.97	41.18	93.75	87.34	**64.56**	6.67	36.59
	F	69.30	**75.12**	50.00	59.13	55.26	93.75	90.97	**71.33**	12.50	47.87
+ named entities	*Prec*	76.76	76.24	70.59	**67.31**	84.00	93.75	**94.59**	79.69	100.00	**70.15**
	Rec	62.93	74.04	37.11	**55.56**	41.18	93.75	**88.61**	64.56	6.67	38.21
	F	69.16	75.12	48.65	**60.87**	55.26	93.75	**91.50**	71.33	12.50	49.47

Table 2: Leave-one-out cross-validation results for the different feature group combinations and event types - precision (*Prec*), recall (*Rec*) and F-scores (*F*) in %.

disambiguated lemma features is sometimes improved by also taking into account the hypernyms of the lemmas. The contribution of the syntactic and semantic feature groups differs depending on the event type at hand.

5. Conclusions and future work

We presented proof-of-concept experiments for a classification-based approach to detecting economic events in Dutch news text. For this purpose, we manually annotated a corpus of news articles and experimented with various lexical, syntactic and semantic features. We obtained good results using basic shallow features and a limited amount of annotations.

In future work, we will optimize the basic classifiers presented in this paper by expanding our annotated corpus and performing a qualitative analysis and feature selection. To improve recall, we will add deeper semantic features by exploiting databases such as DBpedia (Lehmann et al., 2014) and Cornetto (Vossen et al., 2008) and by integrating an ex-

isting Dutch coreference resolver (Hoste, 2005; Hendrickx et al., 2008) into the system. Additionally, our goal is to develop a similar classification framework for economic event detection in English text.

After optimizing the classifiers for the detection of these 10 company-specific economic events, we will extend our approach to a cascaded system for the more fine-grained detection of certain sub-events (which have already been annotated in the corpus of financial news articles). For instance, after having identified a mention of a profit event, we will assign it to one of several sub-event categories such as profit increase, profit decrease, etc. Finally, we do not only want to detect mentions of certain economic events, but also aim at extracting other useful pieces of information related to these events, such as the name and sector of the company the event is related to (information we already partly have at our disposal through NER and which will further be extended through the used of LOD resources) and the time at which the event occurred.

6. References

Ahn, D. (2006). The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia.

Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., Kameyama, M., and Tyson, M. (1993). SRI: Description of the JV-FASTUS system Used for MUC-5. In *Proceedings of the 5th Conference on Message Understanding (MUC 1993)*, pages 221–235, Baltimore, MD, USA.

Arendarenko, E. and Kakkonen, T. (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.

Bouma, G., van Noord, G., and Malouf, R. (2001). Alpino: Wide-coverage Computational Analysis of Dutch. In *Computational Linguistics in The Netherlands 2000: Selected Papers from the Eleventh CLIN Meeting*. Rodopi.

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

Drury, B. and Almeida, J. a. J. (2011). Identification of Fine Grained Feature Based Event and Sentiment Phrases from Business News Stories. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS '11)*, Sogndal, Norway.

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

Hardy, H., Kanchakouskaya, V., and Strzalkowski, T. (2006). Automatic Event Classification Using Surface Text Features. In *Proceedings of the 2006 AAAI Workshop on Event Extraction and Synthesis*, pages 36–41, Boston, MA, USA.

Hendrickx, I., Hoste, V., and Daelemans, W. (2008). Semantic and Syntactic Features for Anaphora Resolution for Dutch. In *Lecture Notes in Computer Science, Volume 4919, Proceedings of the CICLing-2008 conference*, pages 351–361, Haifa, Israel.

Hogenboom, A., Hogenboom, F., Frasincar, F., Schouten, K., and van der Meer, O. (2013). Semantics-Based Information Extraction for Detecting Economic Events. *Multimedia Tools and Applications*, 64(1):27–52.

Hogenboom, F. (2012). Financial Events Recognition in Web News for Algorithmic Trading. In *Ninth International Workshop on Web Information Systems Modeling (WISM 2012) at Thirty-First International Conference on Conceptual Modeling (ER 2012)*, volume 7518 of *Lecture Notes in Computer Science*, pages 368–377. Springer.

Hoste, V. (2005). *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, Antwerp University.

Ji, H. and Grishman, R. (2008). Refining Event Extraction through Cross-document Inference. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 254–262, Columbus, OH, USA.

Joshi, M. and Penstein-Rosé, C. (2009). Generalizing Dependency Features for Opinion Mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316, Suntec, Singapore.

Kothari, S. and Warner, I. (2007). Econometrics of event studies. pages 2–32. Elsevier.

Lefever, E., Van de Kauter, M., and Hoste, V. (2014). Evaluation of Automatic Hypernym Extraction from Technical Corpora in English and Dutch. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 490–497, Reykjavik, Iceland.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2014). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*.

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

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

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (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.

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

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L., and Hoste, V. (2013). LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London.

Vossen, P., Maks, I., Segers, R., and van der Vliet, H. (2008). Integrating lexical units, synsets and ontol-

ogy in the Cornetto Database. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 1006–1013, Marrakech, Morocco.