SENTiVENT: enabling supervised information extraction of company-specific events in economic and financial news

Gilles Jacobs¹ · Véronique Hoste¹

Accepted: 14 September 2021 / Published online: 8 October 2021
© The Author(s) 2021

Abstract We present SENTiVENT, a corpus of fine-grained company-specific events in English economic news articles. The domain of event processing is highly productive and various general domain, fine-grained event extraction corpora are freely available but economically-focused resources are lacking. This work fills a large need for a manually annotated dataset for economic and financial text mining applications. A representative corpus of business news is crawled and an annotation scheme developed with an iteratively refined economic event typology. The annotations are compatible with benchmark datasets (ACE/ERE) so state-of-the-art event extraction systems can be readily applied. This results in a gold-standard dataset annotated with event triggers, participant arguments, event co-reference, and event attributes such as type, subtype, negation, and modality. An adjudicated reference test set is created for use in annotator and system evaluation. Agreement scores are substantial and annotator performance adequate, indicating that the annotation scheme produces consistent event annotations of high quality. In an event detection pilot study, satisfactory results were obtained with a macro-averaged F_1 -score of 59% validating the dataset for machine learning purposes. This dataset thus provides a rich resource on events as training data for supervised machine learning for economic and financial applications. The dataset and related source code is made available at <https://osf.io/8jec2/>.

Keywords Event extraction · Economic events · Financial information extraction · Annotation scheme · English corpus · Event detection

✉ Gilles Jacobs
gilles.jacobs@ugent.be
Véronique Hoste
veronique.hoste@ugent.be

¹ Language & Translation Technology Team, Ghent University, UGent VTC Mercator A, Abdisstraat 1, 9000 Ghent, Belgium

1 Introduction

In the economic domain, information extraction from text is highly popular for making available fundamental knowledge present in economic text, such as business news (Day & Lee, 2016; Khedr et al., 2017), regulatory disclosures (Cavar & Josefy, 2018; Feuerriegel & Gordon, 2018), and social media (Gémar & Jiménez-Quintero, 2015; Oliveira et al., 2017). Extracting factual data (using named entity recognition, relation extraction, and event extraction) or subjectivity data (using sentiment analysis) from economic text has the ability to enhance the available numerical data on markets with fundamental information for financial applications. In business and finance, events encapsulate new information on the market and automatically collecting novel event data thus has applications in stock prediction (Bholat et al., 2015; Chen et al., 2019; Nardo et al., 2016; Nassirtoussi et al., 2014; Zhang et al., 2018), risk analysis (Hogenboom et al., 2015; Wei et al., 2019), policy assessment (Karami et al., 2018; Tobback et al., 2018), brand management (De Clercq et al., 2017; Geetha et al., 2017) and marketing (Rambocas & Pacheco, 2018) (Fig. 1).

While learning-based methodologies are currently predominantly used in NLP research, the domain of economic event extraction is still dominated by pattern-based approaches. In order to allow for data-driven supervised event extraction for company specific news, we discuss in this work the construction of the SENTiVENT Economic Event corpus, a representative corpus enabling both NLP and market research. The SENTiVENT annotation scheme aims to be compatible with the benchmark ACE/ERE event datasets.

We believe this work delivers the following contributions to the field of economic event processing:

- The first manually annotated SENTiVENT news corpus of fine-grained economic events: we annotated full ACE/ERE-like event schemata with attributes (type, subtype, negation, modality, realis), event co-reference, and

Fig. 1 Examples of annotated economic event schemata with argument roles and attributes. Boldface indicates event trigger spans which can be discontinuous

arguments (participant/filler) enabling structured event extraction in the economic domain for English. There are no available datasets on fine-grained economic events in the field of economic information extraction. Compatibility with ACE/ERE-like event nuggets allows straight-forward porting of state-of-the-art models in Event Extraction for which ACE2005 remains a benchmark dataset.

- Representative collection and typology creation methodology: we applied a dataset collection approach for obtaining a corpus representative of the business news genre minimizing topical specialization across sectors. Coverage and representativeness are guiding principles in the creation of the event typology. The choice of which events to include in event typologies are usually arbitrary and not transparent. We applied a typology creation method of iterative refinement resulting in annotator agreement on type and subtype identification that is substantial-(almost) perfect.
- A sentence-level event detection pilot study producing satisfactory results proving this dataset is adequate for machine learning purposes.

The remainder of this paper is organized as follows. First, Sect. 2 discusses related work on event extraction and information extraction in the economic domain. Section 3 describes our article collection approach used to obtain a corpus representative of business news. Section 4 presents the annotated units and definitions in the corpus, the tools and procedure used for annotation. Section 5 covers the inter-annotator agreement study. Next, we present the properties and statistics of the annotated corpus in Sect. 6. Finally, we elaborate on the methodology and results of our event detection pilot study in Sect. 7. We conclude with some notes on the availability of the corpus and prospects for future research.

2 Related research

2.1 General domain fine-grained event extraction

Our work on economic event annotation and processing can be accommodated within the rich history of automatic event processing. Starting with a pilot study in 1999, the ACE (Automatic Content Extraction) annotation scheme and programme was highly influential in event processing. The objective of the ACE program was to develop technology to automatically infer entities being mentioned in text, the relations among these entities that are directly expressed, and the events in which these entities participate (Doddington et al., 2004). The ACE programme also periodically organized “Event Detection and Recognition” evaluation competitions in which event extraction corpora were released (Walker et al., 2006; Consortium, 2005), thus leading to different types of methodologies to tackle the problem of event extraction. Some years later, the ERE (Entities, Relations, Events) standard was conceived under the DARPA DEFT program as a continuation of ACE with the goal of making annotation easier and more consistent. It consolidated and simplified the ACE typology and removed many complex annotation features into the Light

ERE standard (Aguilar et al., 2014; Bies et al., 2016). Both in the ACE and Light ERE annotations, an event is defined as an explicit occurrence involving participants, belonging to a pre-defined set of types/subtypes. Subsequently, expansions to this standard were added as Rich ERE (Song et al., 2015). The typology and attributes were expanded slightly and more attention was devoted to event coreference (“Event hoppers”). The DEFT ERE standards formed the basis of the Event Nugget annotation scheme proposed by Mitamura et al. (2015b). An event nugget was defined as a semantically meaningful unit referring to an event. The event trigger could be a single word as in ACE/Light ERE or a discontinuous multi-token expression. The Knowledge Base Population track of the Text Analysis Conference (TAC-KBP) organized several shared tasks based on this event nugget standard (Mitamura et al., 2015a, 2016a, 2017). Our annotation scheme is inspired by the Rich ERE Event annotation schemes (Linguistic Data Consortium, 2016, 2015a, 2015b) as to provide compatibility with on-going research in event extraction, where ACE and ERE datasets remain dominant benchmarks.

2.2 Events in the economic domain

Business and finance are domains in which new information is produced and consumed at a high rate due to the economic proposition of information acquisition. In practice, investors and analysts widely use text data such as news announcements and commentary on recent events as a major source of market information (Oberlechner & Hocking, 2004). In the financial domain, the way companies are perceived by investors is influenced by the news published about those companies (Engle and Ng, 1993; Mian and Sankaraguruswamy, 2012; Tetlock, 2007).

Aside the economic finality, there is also academic interest in automated event extraction: financial economists conduct so-called “event studies” to provide some insights into the way markets react to new information about companies. These event studies measure the impact of a specific event on the value of a firm (MacKinlay, 1997). Identifying news published about certain events in an automated way enables researchers in the field of event studies to process more data in less time. This can consequently lead to new insights into the correlation between events and stock market movements.

Most of the existing approaches to the detection of economic events, however, are *knowledge-based and pattern-based* (Arendarenko & Kakkonen, 2012; Chen et al., 2019; Du et al., 2016; Hogenboom et al., 2013). These use rule-sets or ontology knowledge-bases which are largely or fully created by hand and do not rely on fully manually annotated supervised datasets for machine learning. The Stock Sonar project (Feldman et al., 2011a) 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). Along the same line, Hogenboom et al. (2013) relies on a hand-crafted financial event ontology for pattern-based event detection in the economic domain and incorporates lexicons, gazetteers, PoS-tagging and morphological analysis.

A rich variety of representations of events in economic news exists in experiments on stock market prediction. Here event information is used as an information source in machine learning-based price or movement prediction. On a continuum of coarse- to fine-grained event representations as features, we can distinguish approaches using raw text as embeddings (Del Corro & Hoffart, 2020; Hu et al., 2018; Xu & Cohen, 2018), to topic modeling (Nguyen and Shirai, 2015; Zhou et al., 2015)), to *subject, predicate, object* tuples (Ding et al., 2015, 2016; Zhang et al., 2018) or *entities, time, location, key words* tuples (Zhou et al., 2015) to fine-grained semantic frames (Xie et al., 2013) and ACE/ERE-like event structures (Chen et al., 2019; Yang et al., 2018). The latter rely on rule-based data generation methods (distant supervision) to produce ACE/ERE-like event data as no supervised economic datasets are available. Chen et al. (2019) highlights the improved performance in stock prediction of more fine-grained event structures viz. coarse-grained representations.

Several *semi- or distantly supervision* approaches exist in which seed sets are manually labeled or rule-sets are used to generate or enhance training data. However, these often lack the granularity and structure of ACE-like events and are annotated at the document level. Qian et al. (2019), for example, used a semi-supervised approach for event type detection using word vector clustering in Chinese economic news. Clusters were manually labeled for event type and unseen documents were classified by cluster similarity. The authors annotated 4 event types (“Financing”, “M&A”, “IPO”, “Delisting”) on a gold-standard evaluation corpus of 14,556 documents containing article headlines and leads from October 2011 to February 2017. Rönqvist and Sarlin (2017) used weak supervision in the financial context of bank distress events. They relied on a knowledge base of known bank distress events and thus collected 386K sentences. A sentence describes a weakly supervised event when the article’s publication date is near the date at which the bank distress event occurred. Here, bank distress events are conceptualized as mentions of bank entities in a time-window and no typology classification is assigned. Ein-Dor et al. (2019) adopted a weakly-supervised training dataset for identifying company-related events at the sentence-level in English news articles. Their main interest was in the binary task of detecting sentences containing events. Sentences from a company’s Wikipedia article were selected if they appeared in an event-section and started with a date-pattern (such as ‘On/In/By/As of’ + month + year). The training set was re-balanced to enforce an equal number of positive and negative instances. The target evaluation dataset consisted of articles scraped through Seeking Alpha containing the name of the 10 most eventful S&P companies in Wikipedia for 2019. As test annotations, events were manually collected from 2019 from the Wikipedia pages of those companies. All sentences containing reference to these were marked positive (26 different events). Yang et al. (2018) also used a weak-supervision, hybrid approach in creating an event dataset for Chinese economic news. They relied on a financial event knowledge database consisting of labelled event trigger and argument samples of nine pre-determined events collected by financial professionals. The text data came from official announcements released by companies on the web and obtained from Sohu securities net. A weakly supervised dataset was created by dictionary matching the

knowledge-base event triggers and arguments against the corpus. Han et al. (2018) describe a hybrid approach to ACE-like event extraction in Chinese business news. They labeled triggers, argument types, and event types for 1500 Chinese news articles from the Sina.com news aggregator with 8 economic event types. Unfortunately, detailed information on the annotation scheme was missing and it was not clear to what extent manual annotation has been applied. In their experimental work, the authors mainly relied on a knowledge-based methodology with ML-based expansion using trigger and argument dictionaries and a pattern-matching approach for argument extraction. They expanded a handmade event trigger dictionary by finding synonyms using a Word2Vec-based word vector model of a background corpus of Chinese business news.

Few *strictly supervised approaches* exist in the literature, precisely due to the lack of annotated training data.

Our previous work in the SentiFM project (Van de Kauter et al., 2015a, 2015b) mainly focused on annotating implicit and explicit sentiment in Dutch and English economic text. Besides sentiment annotations, event annotations were made, but were limited to annotating continuous trigger spans of 10 events types and 64 subtypes. Only one argument relation was tagged linking the company involved to the event span. No other event attributes were annotated. The English SentiFM corpus consists of 497 news articles containing 2522 annotated events. The English subcorpus was collected from The Financial Times using keyword search where the title had to contain one out of seven companies. The Dutch part has 126 articles on seven BEL20 companies containing 1152 events. Several supervised event detection experiments have been conducted on this dataset (Jacobs et al., 2018; Lefever & Hoste, 2016). Our current corpus has more fine-grained event annotations with additional attributes, the typology is more diverse, counts more annotated events and has a higher event density meaning more relevant events are tagged. It has been crawled randomly making it more representative of the business news genre.

We are not aware of any other published fine-grained ACE/ERE-like event extraction datasets for the economic domain. This work presents the first attempt at the creation of a gold-standard event dataset in the economic domain with incorporating an agreement/annotation quality study and detailed guidelines.

3 Dataset collection

First, we crawled English online business news articles on all companies in the S&P500 from various sources over a period of 14 months (June 2016–May 2017) using the business news aggregator Yahoo! Finance¹ API. Next, we filtered all articles pertaining to a hand-selected set of 30 companies based on publication frequency and sector diversification.² Finally, we selected a subset of articles for each company based on temporal spread, topical diversity, and quality criteria for linguistic annotation. The goal of this selection method was to assure temporal

¹ <https://finance.yahoo.com/>.

² Sector diversity is defined as areas of business providing differing types of goods and services.

Table 1 Overview of companies on which articles are annotated in the English SENTiVENT corpus

Company	Industry	Ticker	# Docs
AT&T Inc.	Telecommunication Services	T	10
AbbVie	Health Care	ABBV	14
Advanced Micro Devices	Information Technology	AMD	15
Alphabet Inc Class C	Information Technology	GOOG	14
Amazon.com Inc.	Consumer Discretionary	AMZN	15
American Airlines Group	Industrials	AAL	15
Apple Inc.	Information Technology	AAPL	17
Bank of America Corp	Financials	BAC	7
Boeing Company	Industrials	BA	17
Celgene Corp.	Health Care	CELG	5
Chesapeake Energy	Energy	CHK	4
Chevron Corp.	Energy	CVX	8
Chipotle Mexican Grill	Consumer Discretionary	CMG	5
Costco Wholesale Corp.	Consumer Staples	COST	4
Duke Energy	Utilities	DUK	8
Facebook, Inc.	Information Technology	FB	15
Ford Motor	Consumer Discretionary	F	16
General Electric	Industrials	GE	14
General Motors	Consumer Discretionary	GM	12
Goldman Sachs Group	Financials	GS	5
Johnson & Johnson	Health Care	JNJ	7
Kinder Morgan	Energy	KMI	5
Microsoft Corp.	Information Technology	MSFT	4
Netflix Inc.	Information Technology	NFLX	9
Newmont Mining Corporation	Materials	NEM	7
Procter & Gamble	Consumer Staples	PG	6
The Walt Disney Company	Consumer Discretionary	DIS	13
Twenty-First Century Fox Class B	Consumer Discretionary	FOX	5
Wal-Mart Stores	Consumer Staples	WMT	8
Wells Fargo	Financials	WFC	4

spread and to avoid that a company, news event, sector or topic was over-represented in the data. Avoiding topical specialisation ensures the data is general enough to represent business news as a text genre.

The initial crawl was restricted to all companies that are historically and currently present in the Standard & Poor's 500 (S&P 500) Index. The S&P 500 was chosen over other indices because it contains companies which are commonly reported and has more sector diversification than other indices such as the DJIA or the technology-heavy NASDAQ. Made up of 500 of the most widely traded stocks in

the U.S., it represents about 80% of the total value of U.S. stock markets. In general, the S&P 500 index thus gives a good indication of movement in the U.S. marketplace as a whole. Additionally, technical data on stock performance and metrics is readily available for stocks listed in the S&P 500, which also enables stock performance and market research. We scraped the linked article pages and their metadata (such as tags, author, publisher source, etc.) daily in the period of June 2016 to May 2017.

The selection of 30 companies was made based on sectorial diversification and frequency in the crawled data. This selection procedure provides a balance between the need for sector diversification and representation of industries in the business news genre. All included companies are listed in Table 1 alongside their GICS-codified industries. For each company, a final random selection of articles was made based on temporal spread of all available articles. Using temporal binning of publication dates, we ensure that reported topics are different over time because articles that are published at the same time often report on the same topic.

To extract the textual content from the article web-page, three article content extraction tools were used: (1) a self-written rule-based parser, the (2) rule-based Goose (Plush, 2014) and (3) machine learning-based BoilerPipe (Kohlschütter et al., 2010). The best output for each article was manually chosen in an article selection phase and missing text was manually corrected.

Articles that did not meet the following quality and relevancy criteria were removed from the selection: (a) company-specific focus, (b) topical diversity: articles focusing on a earlier reported event in the corpus are removed, (c) pertinence and novelty (only articles relevant to the present state of the market were selected), and (d) human language production, leading to the exclusion of robot-written and templated articles.

This method ensures a representative, topically diverse, and temporally spread corpus enabling both text-based information extraction as well as market research.

4 Annotation definitions and procedure

In this section, we define the event annotations as well as the annotation process. While an event can be defined as a specific real-world occurrence involving participants, our focus is on economic events, i.e. *‘textually reported real-world occurrences, actions, relations, and situations involving companies and firms’*. Since we could not rely on annotation guidelines from previous work, nor on publicly available datasets, we decided to seek inspiration in the rich body of previous work on event annotation from the ACE/ERE (Song et al., 2015) competitions, for which detailed guidelines and annotations are available. For the annotation of economic events we thus draw on the operationalization of event structures from the DEFT Rich ERE annotation guidelines (Linguistic Data Consortium, 2016), and more specifically the technical reports on annotating events (Linguistic Data Consortium, 2015b) and arguments (Linguistic Data Consortium, 2015a).

Fig. 2 Examples of annotated economic event schemata with argument roles. Boldface indicates event trigger spans which can be discontinuous

There is one important difference with ACE/ERE datasets: We do not tag predefined entity types and relations. Unlike ACE/ERE, argument role definitions are hence not restricted by entity types, i.e. any text span that describes a relevant event argument within the event scope is tagged.

Events are limited to a *type* and *subtype* from our event typology, which we will further discuss in Sect. 4.1. A sentence can contain multiple events and event argument links can cross sentence boundaries. The full event annotation consists of the event trigger, argument span, and attributes (type, subtype, modality, negation) in running text (i.e. an event mention in ACE/ERE terminology). Figure 2 shows examples of annotated events with attributes.

The event trigger is the minimal span of text (a single word or a small phrase) that most succinctly expresses the occurrence of an event. Generally, we think of the trigger as the word that most strongly refers to an event. Event triggers are annotated at the token-level and discontinuous token-spans are allowed. In Fig. 2, "sell" is a single-word trigger while "sales ... stall" and "revenues ... rise" are discontinuous multi-token triggers. Annotators are instructed to keep the trigger as small as possible while maintaining the core lexical semantics of the event.

Event arguments are the participating entities that are involved in the event by filling a certain prototypical semantic role. *Arguments fill a number of roles specific to an event type*. Examples are the person or company performing an action, the amount of capital involved, or some other piece of information, like the time or place where the event happens. It is possible that none of the participants or only a select few are tagged. We distinguish two types of event arguments:

1. *Participant arguments*: these are the text spans describing a Participant that is involved in the event. For each event type/subtype there is a specific set of Participant roles to be filled. We only tag Participant arguments that occur within the event mention scope for an event trigger. The event mention scope is defined as the span of a document from the first trigger of an event to the next trigger you see for the same, coreferent event. This avoids double tagging of arguments for co-referent events.
2. *Filler arguments*: Filler arguments are non-central arguments in the event but provide descriptive and discriminative information. They denote the time, place, capital (amount of money), etc. attached to an event. Filler arguments differ from Participant arguments in two ways: (i) they are not bound by event mention scope and can be tagged anywhere in the document; (ii) they have to be full noun phrases (NPs), unlike Participants that can be anaphoric pronouns.

We also annotated two features pertaining to the *factuality* of the event along two axes: Modality with values Certain vs. Other and Polarity with values Affirmed vs. Negated. *Polarity* captures whether an event mention is affirmed (the event actually took place) or negated (the event did not happen). *Modality* captures the degree of certainty (i.e., epistemic modality) about the event being the case as expressed by the author or another source. Automatic processing of event factuality is necessary for downstream applications such as knowledge-base population or market modelling so that events that are speculative or did not take place in the real-world can be filtered out.³

Additionally, two types of relations are also annotated: *Event coreference* indicates whether an event intuitively refers to the same real-world event. It is required that type and subtype are identical and that the location and temporal scope match. For event coreference, we directly inherited Rich ERE's concept of "event hoppers". A hopper is a group of annotated events that semantically refer to the same event occurrence. This differs from strict event coreference in ACE and Light ERE in which arguments also have to corefer strictly by exact matching of all event features and argument semantics. We also annotate the *Canonical referent* of a pronominal or anaphoric noun phrase. When participant arguments are realized as a pronoun ("it") or an anaphoric noun phrase ("the company", "the stock"), annotators are instructed to draw a link to the nearest canonical nominal referent

³ ERE events have one feature *Realis* with three values: Actual (asserted), Generic (habitual), and Other (future, hypothetical, negated, uncertain, etc.). "Generic" realis was infrequent in initial trails and was not annotated in our corpus.

Fig. 3 Diagram illustrating iterative refinement of the event typology

(containing the name of the company, person, or the stock). This link is useful for producing a dataset variant in which all arguments are nominal phrases (names).

4.1 Event typology creation

The event typology specifies which types of events we annotate and the semantic argument roles in which they are involved. The aim of defining an event typology for economic event detection is to obtain adequate coverage of the factual content of economic news articles. In designing such a typology, we have to balance relevance from an investor point-of-view, generality of event types across industries, and coverage. Our event type selection methodology was similar to FrameNet’s, where types start with an annotator’s intuition and are gradually defined and grouped based on corpus examples (cf. Ruppenhofer et al., 2016, p. 12 for details on frame construction and type selection).

We iteratively developed the event typology on a subcorpus in close collaboration with a financial domain expert and in accordance with the prior state-of-the-art. The Stock Sonar project (Feldman et al., 2011b) expert-created event typology identifies 8 event types, viz. “Legal”, “Analyst Recommendation”, “Financial”, “Stock Price Change”, “Deals”, “Mergers and Acquisitions”, “Partnerships”, “Product”, and “Employment”. Boudoukh et al. (2019) identified the following 18 event categories based on Capital-IQ types⁴ and a cross-section from academic event studies: “business trends, capital returns, CSR/brand, deals, earnings factors, employment, facility, financial, financing, forecast, general, investment, legal,

⁴ Capital IQ is a technology and financial services company that acts as the research division of Standard and Poor’s.

mergers & acquisitions, product, ratings, stocks, and stock holdings". The event typology we arrived at by gradual refinement has a high overlap with this work. The Hogenboom et al. (2013) rule-based SPEED event detection system differentiates between ten different financial events (announcements regarding "CEOs", "presidents", "products", "competitors", "partners", "subsidiaries", "share values", "revenues", "profits", and "losses"). The PULS business intelligence system of Du et al. (2016) detects 15 event types pre-categorized as "positive" or "negative". Positive business events include "acquisition", "investment", "order marketing", "product launch", "merger". Negative events are "bankruptcy", "lawsuit", "business closing", "layoff", "product recall", "accident". The SentiFM project for sentiment analysis and business event detection defined 10 events: "Buy rating", "Debt", "Dividend", "Merger&Acquisition", "Profit", "Quarterly results", "Sales volume", "Share repurchase", "Target price", and "Turnover". The ACE/ERE event typology defines 8 types and 33 subtypes of which one event type, namely "Business" with 4 subtypes ("Start-org", "Merge-org", "Declare-bankruptcy" and "End-org") is also relevant to the economic domain. These types however lack descriptiveness in capturing the information commonly conveyed in business-specific news and are too coarse to be of any use for business analysis (Fig. 3).

These existing event typologies were taken as a basis for our typology. We joined similar types into a seed list of event types and on a random subsample of 40 crawled articles, each belonging to different companies, we counted the presence of these events. Additionally, we labeled special interest events and news pegs⁵ that were not in the seed typology. We counted the frequency of these types, subtypes and participants in the subcorpus and adjusted the typology by adding new event types or removing infrequent events. Relevant participant roles were also added in this process of iterative refinement. After this iterative process, we ended up with a typology consisting of 18 types and 42 subtypes of company-specific economic events as listed in Table 2. For a full description of the types and subtypes, we refer to the guidelines (Jacobs, 2020b).

The types "Profit/loss" (aka "Earnings"), "Revenue", "Expense", "Sales Volume" represent major metrics in financial reporting on the income statement. We found these types the most common generally reported financial metric in business news. Other commonly reported metrics such as debt ratios or cash flow statements as well as highly industry-specific indicators were collected under the "Financial Report" event type. Asset performance event types include "Rating" and "Security Value". These involve analysts giving investor advice and discussions of a company's stock value. "Macroeconomics" is a broad category pertaining to events that do not involve decisions or interactions of specific firms. While this category is not company-specific, it was devised to capture highly common discussions of sector trends, economy-wide phenomena, and governmental policy in news.

⁵ "News pegs" are the real-world noteworthy events that trigger a journalist to write an article [definition from Upadhyay et al. (2016)].

Table 2 Event types and subtypes

Type	Subtypes
CSR/brand	
Deal	
Dividend	Payment, yield raise, reduction, stable
Employment	Start, end, compensation
Expense	Increase, decrease, stable
Facility	Open, close
FinancialReport	Beat, miss, stable
Financing	
Investment	
Legal	Proceeding, conviction/settlement, acquit, appeal
Macroeconomics	
Merger/acquisition	
Product/service	Launch, cancellation/recall, trail
Profit/loss	Increase, decrease, stable
Rating	Buy/outperform, sell/underperform, hold, upgrade, downgrade, maintain, price target
Revenue	Increase, decrease, stable
SalesVolume	Increase, decrease, stable
SecurityValue	Increase, decrease, stable

4.2 Annotation process

We used the online text annotation platform WebAnno v4.5 for all our annotations. Three annotators were trained under supervision. All annotators have professional linguistic training in English equivalent to C2-level and all annotators had taken an undergraduate-level course on economics. We also trained the annotators in the terminology and financial concepts related to the events. An independent annotator corrected first pass annotations on a subsample of 104 (40%) documents. Potentially problematic documents were prioritized for correction by searching for often made annotation mistakes: e.g., error-prone event types, unusual event-to-article-length ratio or a high amount of weakly-realized events.

5 Inter-annotator agreement and annotator performance

To determine the feasibility of the annotation task, we conducted an inter-annotator agreement study on a subsample of 30 documents pertaining to 10 different companies. Each document was fully annotated by the three annotators. Our task demands a large degree of interpretation as events are abstract semantic categories (the event type, subtype and arguments of an event have the potential to be ambiguous) and as such an agreement metric should highlight the strength of

consensus among annotators. We applied two approaches to determine annotation agreement: (a) Chance-corrected span-aligned agreement of event token spans, (b) Event nugget scoring to capture performance of annotators on the full event structure compared to an adjudicated reference.

5.1 Span-aligned agreement

Event annotations are made at the token level and token spans can also be discontinuous. In line with Lee and Sun (2019), who examined agreement for span annotations in medical text, we computed agreement metrics by using overlapping span alignment. This allows us to determine agreement on the labels of events, while the exact span overlap is not of interest. Events can be considered semantic categories which are not bound to specific syntactic or lexical boundary rules, leading to quite some variation in boundaries of token spans as exemplified in Fig. 4. Figure 4 examples 1–3 show typical examples of annotations where boundaries do not match exactly, but the core event trigger remains valid. Applying exact span agreement in which spans only match at the boundaries would underestimate the agreement of span annotations.

Other work in computational linguistics with span annotations perform agreement scoring at the sentence-level (Pavlopoulos and Androutsopoulos, 2014; Pontiki et al., 2016) or clause level (Thet et al., 2010) and thus circumvent the issue of alignment. These approaches, however, produce larger units and lose granularity resulting in higher agreement scores due to not discounting unmatched spans that are in the same sentence/clause. Our span-alignment approach presents a more rigorous attempt at unitizing spans for computing agreement.

For calculating span-aligned agreement, we first unitized the text by matching overlapping spans of annotations into alignment groups and we subsequently computed agreement metrics of those groups. In other work where overlapping span alignment is used (Lee & Sun, 2019; Liu et al., 2015), this is done pair-wise between annotators, with the final agreement score being the mean of each annotator pair. However averaging of pair-wise scores does not allow calculating the mean expected agreement distribution of labels across all annotators necessary for Krippendorff's α . Hence, we made alignment groups for all three annotators at once. When one annotator had multiple candidates that overlapped with other annotators, one annotation had to be selected. This selection was made based on the dice-coefficient (equivalent to F_1 -score) of overlapping tokens. The other non-matched annotations were then placed in their own alignment group. For example, in Sentence 4 in Fig. 4, annotator 3 has made a self-overlapping annotation "*increasing commodity prices*" and "*increasing commodity*". The selection is made by taking the mean dice overlap score of all other annotations which results in alignment of "*increasing commodity prices*" with the spans of the other annotators. The span "*commodity prices*" was then placed in its own alignment group with no match. Example 5 is a similar case with a discontinuous annotation of annotator 2 showing that our alignment approach handles discontinuity correctly. Annotator 2 has made a discontinuous annotation "*upgraded . . . to outperform*" and annotator 3 has made two different annotations "*upgraded*" and "*outperform*". "*upgraded*" by

Fig. 4 Examples from the corpus of non-exact span overlap. Double lines in example 4. indicate that the annotator has made an overlapping annotation

annotator 1 and 3 and the discontinuous “*upgraded ... to outperform*” will be placed in the same alignment group while “*outperform*” is unmatched as desired.

To measure the amount of agreement that is above chance across the three annotators, we used the chance-corrected agreement measures Fleiss’ kappa (κ), Krippendorff’s alpha (α), and Gwet’s AC1 on the aligned span annotations. Fleiss’ Kappa (κ) (Fleiss, 1971) is a chance-corrected agreement measure that generalizes Scott’s Pi to any number of annotators. Krippendorff (2004) argues for the use of Alpha (α) in favour of other measure like Kappa because of its independence to the number of annotators and its robustness against imperfect data. AC1 was introduced by Gwet (Gwet, 2001) to provide a more robust measure of inter-annotator reliability than Fleiss’ Kappa. AC1 can overcome Kappa’s sensitivity to trait prevalence and annotator’s classification probabilities (i.e., marginal probabilities) (Gwet, 2002). The interpretation of AC1 is similar to generalized kappa, which is used to assess inter-annotator reliability when there are multiple annotators. We used the R package irrCACC⁶ to compute these metrics and their statistical properties.

Table 3 shows the agreement analysis with coefficients, their standard error, the 95% confidence interval (C.I.) and p -values. We benchmark these coefficients using the cumulative membership probabilities within the agreement ranges as set out by Landis and Koch (1977): $\kappa < 0 = \text{poor agreement}$, $0 \leq \kappa \leq 0.2 = \text{slight agreement}$, $0.2 < \kappa \leq 0.4 = \text{fair agreement}$, $0.4 < \kappa \leq 0.6 = \text{moderate agreement}$, $0.6 < \kappa \leq 0.8 = \text{substantial agreement}$, $0.8 < \kappa \leq 1 = \text{almost perfect agreement}$. As recommended by Gwet (2014), we set the benchmark membership probability cut-off point at 95%.⁷

Agreement on the event main type is “almost perfect” for all metrics with $\kappa = 0.88$, $\alpha = 0.87$, and $AC1 = 0.89$ indicating that annotators are highly consistent in assigning the 18 event type categories. For event subtype labeling, “substantial” agreement is obtained on all metrics with $\kappa = 0.81$, $\alpha = 0.81$, and $AC1 = 0.82$. The

⁶ <https://CRAN.R-project.org/package=irrCAC>.

⁷ Due to the cumulative membership probability cut-off, membership to a benchmark range can turn out lower than the computed coefficient score.

Table 3 Span-aligned agreement scores on events for type, full type (type and subtype), negation, and modality

Metric	Coeff	StdErr	95% C.I.	<i>p</i> -value
Main type				
Fleiss' κ	0.877	0.025	(0.828, 0.927)	<0.0001
Krippendorff's α	0.874	0.012	(0.85, 0.897)	<0.0001
Gwet's AC1	0.885	0.025	(0.836, 0.934)	<0.0001
Full type (w/subtype)				
Fleiss' κ	0.813	0.025	(0.764, 0.861)	<0.0001
Krippendorff's α	0.809	0.014	(0.783, 0.836)	<0.0001
Gwet's AC1	0.824	0.025	(0.776, 0.872)	<0.0001
Negation				
Fleiss' κ	0.852	0.070	(0.715, 0.99)	<0.0001
Krippendorff's α	0.780	0.094	(0.596, 0.963)	<0.0001
Gwet's AC1	0.993	0.025	(0.944, 1)	<0.0001
Modality				
Fleiss' κ	0.483	0.046	(0.392, 0.574)	<0.0001
Krippendorff's α	0.459	0.041	(0.379, 0.539)	<0.0001
Gwet's AC1	0.835	0.027	(0.782, 0.887)	<0.0001

most common mistake we see in the data is when annotators do not assign a subtype. These high agreement scores for both the type and subtype labels showcase that the event typology is sufficiently discriminative. The negation label obtains a “substantial” agreement score using Fleiss' κ ($\kappa = 0.85$) and Krippendorff's α ($\alpha = 0.78$) and an “almost perfect” agreement according to AC1 ($AC1 = 0.99$). Annotators thus have little problems with determining whether events lie within a negated context. Event modality falls in the “moderate” agreement range for κ ($= 0.48$) and α ($= 0.46$), but a “substantial” agreement is obtained for AC1 ($= 0.84$). This indicates that determining whether an event is presented as being uncertain remains difficult for annotators. The different AC1 scores for the negation and modality annotations can be explained by the prevalence bias of the κ and α metrics which AC1 attempts to mitigate since both categories are highly skewed. While AC1 provides “substantial” agreement on modality, we err on the side of caution and recommend refining the modality labels in a secondary revision pass, before using this dataset for factuality processing. Overall, we can state with high certainty that substantial agreement is obtained on events, with the exception of modality. This indicates that the guidelines and typology produce consistent event annotations of high quality.

5.2 Event nugget performance analysis

While token-level metrics can provide information on the difficulty of identifying token span extents in text, they do not relay useful information on the combined performance of event span annotation and its attributes (i.e., event type, event subtype, polarity, modality).

The second approach is not strictly a chance-corrected agreement measure to capture the degree to which annotators agree on annotation units, but this also seeks to capture the quality of the annotations. The specific event nugget scoring method integrates token span-overlap and attribute accuracy. This method was included because it is also used in the ACE/ERE project to assess annotation quality. The Liu et al. (2015) event nugget scorer, developed on DEFT Rich ERE data for the TAC-KBP Event Nugget Extraction task (Mitamura et al., 2015b), produces a span similarity score for a pair of event mentions between 0 and 1. This span score is the Dice coefficient over token ids (equivalent to F_1 -score for token overlap). Each system mention is mapped to the reference by selecting the maximum dice overlap score. This captures the performance of annotators on identifying the correct token spans. Subsequently, attribute label accuracy is computed for matched events. The attributes used by Mitamura et al. (2015b) in their assessment of event nugget annotation in DEFT ERE were limited to “Type” and “Realis”. We can obtain a corresponding value for Rich ERE “Realis” by combining values for “Negation” and “Modality” from our annotations.

The authors compared the nugget scores of the annotators on each other and scored each annotator separately on a reference set. In line with Mitamura et al. (2015b), we created an adjudicated set (ADJ) through correcting and combining all annotations by an independent judge. In their analysis, they also compared their Light ERE annotation scores with earlier ACE2005 annotations. Both of those datasets involved two annotators, while our SENTiVENT dataset involves three. For comparison, we listed the reported scores of the original work (Mitamura et al., 2015b) next to the nugget score performance on our data in Table 4.

Our annotators seem to perform worse on the identification of span boundaries compared to work in ACE and ERE, whereas the accuracy scores of attributes, are on par or better. By having three annotators instead of two more potential disagreement on span boundaries was introduced.

For future annotations, we suggest mitigating span identification issues by reducing the complexity of the event annotation task by splitting up trigger span identification and attribute assignment into two completely separate phases: an event trigger phase in which all potentially relevant events of business news are tagged, followed by an attribute phase in which attributes and arguments are tagged. In this manner, the cognitive overload of annotators that need to memorize rules for span identification and attribute assignment is lessened. The disadvantage being that annotation time and cost will increase substantially.

Overall, similarly to the chance-corrected IAA study of aligned spans, we observe that the event attributes scores are highly satisfactory, although the identification of event spans shows room for improvement. We believe the consistency study shows that the dataset quality is adequate for data-driven machine learning techniques.

Table 4 Nugget scorer performance of annotators vs. adjudicated reference (scores are averaged across annotators)

	SENTiVENT (ours)	TAC KBP/ERE	ACE2005
Span	57.0	83.8	80.6
Type	77.3	78	73
Subtype	68.2	–	–
Realis	77.6	74.5	70.5
Negation	87.3	–	–
Modality	78.0	–	–

Table 5 Basic corpus statistics for the suggested split of the development training set and the adjudicated test set

	Dev	Test	Total
Documents	258	30	288
Events	5212	982	6194
Sentences	6156	727	6883
Tokens	153,272	17,126	170,398
Participant arguments	8746	1,779	10,525
Filler arguments	2658	517	3175
Discontiguous triggers	716	171	887
Canonical referent links	760	179	939
Co-reference links	1143	218	1361

6 Corpus analysis and statistics

In total, our annotators labeled 258 training and 30 evaluation documents, which makes our corpus comparable in size to the largest of the monolingual TAC-KBP corpora. The standardized evaluation (holdout test) set of 30 documents stems from the adjudicated reference set in the inter-annotator agreement study. In the 2016 TAC-KBP event nugget track, which implements Rich ERE event guidelines, the annotated training corpus sizes were 200 documents for English, 200,000 words for Chinese and 120,000 for Spanish (Mitamura et al., 2016b) with an evaluation set size of 60 documents for each language.

We examined the distribution of annotation units as well as some statistical properties which might be relevant for classification (see Table 5). Considering the event type and subtype distribution, as visualised in Fig. 5, we can observe that the Macroeconomic class was the most frequently used event annotation. We introduced this class as a low-specificity category to capture all non-company specific events such as market trends, market-share, competition, regulatory issues and governmental policy, etc.

6.1 Lexical richness of events

A known difficulty in event processing and other semantically-oriented NLP tasks (a.k.a. “deep semantics”) is the large lexical richness of categories (Van de Kauter

Fig. 5 Frequency of event categories in SENTiVENT English corpus

et al., 2015a). Because events are abstract semantic categories, they can be expressed in a large variety of lexicalizations. To quantify the lexical variation within our event type categories, we computed several lexical richness metrics for each event type for the event trigger.

Lexical richness measures give an idea of which classes and event types could cause difficulty when building classifiers that rely on lexical features such as the transformer models we used in the pilot study. Lexical diversity is a measure of how many different words are used in a stretch of text. In its simplest form, lexical diversity is measured through the calculation of type-token ratio (TTR) where the counts of unique words are divided by their occurrences. TTR, however, has an inverse relationship with sample size, introducing a bias when comparing texts of different lengths. As shown in (McCarthy & Jarvis, 2010, 2007), Maas (Mass, 1972), Measure of Textual Lexical Diversity (MTLD) (McCarthy, 2005) and Hypergeometric distribution diversity (HDD) (McCarthy & Jarvis, 2007) correct for this bias and are more robust measures of lexical richness than TTR-based measures such as CTTR (Guiraud, 1959), RTTR (Carroll, 1964), MSSTR (Johnson, 1944), MATTR (Covington & McFall, 2010), Herdan (Herdan, 1964), Somers (Somers, 1966), and Dugast (Dugast, 1978).

We computed these robust non-TTR-based measures using the LexicalRichness Python module (Shun, 2019) as well as lexical entropy as Shannon entropy. We pre-processed the text by stemming and removing stopwords using the Porter2/ SnowballStemmer algorithm in NLTK. Because the event types text lengths differ, we only considered HDD, Maas and MTLD in our discussion because they are substantially less sensitive to differences in text length as recommended in McCarthy and Jarvis (2010).

Figure 6 shows the mean-normalized measures of lexical richness per event type for event triggers. Note that the Maas measure inversely measures lexical richness. This shows that “CSR/Brand” and “Macroeconomics” exhibit the most lexically

Fig. 6 Lexical richness measures (mean-normalized) of triggers for all event types. Ordering is from least to most lexically diverse

rich triggers. That “FinancialReport” and “Macroeconomics” show large amounts of lexical variety is due to them being conceived as event categories that encompass a large amount of distinct real-world possibilities. Event triggers of “Merger/Acquisition”, “Investment”, “Dividend”, “Revenue”, “Expense”, and the financial metric type (“SalesVolume”) are unsurprisingly not very lexically diverse. These events are typically narrowly defined and inspection of the corpus shows that they are not often the target of creative language use such as metonymy or metaphor. We expect that these types to perform well in trigger classification.

7 Event detection pilot study

As validation for this dataset, we performed a set of event detection experiments for determining the type of economic event in text. Event detection is a simpler task than event extraction for which this dataset was designed: instead of extracting events, their attributes (type, subtype, factuality) and arguments, event detection consists in assigning an event type to a given piece of text. For this pilot study, we conceived event type detection as a multilabel sentence-level classification problem,

and tested the fine-tuning of several neural transformer-based language models, which are currently state-of-the-art in text classification.

7.1 Method

The field of NLP has transitioned from developing task-specific models to fine-tuning approaches based on large general-purpose language models (Howard & Ruder, 2018; Peters et al., 2018). Pre-trained models of this type are based on the transformer architecture (Vaswani et al., 2017). Currently, the most widely-used model of this type is BERT (Devlin et al., 2019) and its many variants. Fine-tuning BERT-like models has resulted in state-of-the-art performance in many NLP tasks.

We experimented with fine-tuning pre-trained transformer models for English using the “huggingface/transformers” PyTorch library which provides models in their repository (Wolf et al., 2019). We used the provided default hyper-parameters, tokenizers, and configuration for all models. To add sentence classification capability to the language models, a sequence classification head was added to the original transformer architecture. The batch size was set to 8 instances and sequence length to 128 for all models. The only hyperparameter set in cross-validation grid-search was the amount of training epochs ($e = \{4, 8, 16\}$). The best hyper-parametrization was chosen for each architecture by macro-averaged F_1 -score over 10 folds in cross-validation. Macro-averaged F_1 -score was chosen because we consider each class equally important despite severe class imbalance. We also perform a held-out test in which the model is trained on the dev set and tested on the adjudicated test set.

We provide baselines in the form of “dummy” classifiers: (a) the “majority-class baseline” always predicts the most frequent label, (b) the “stratified baseline” generates random predictions by respecting the training set class distribution, and (c) the “uniform baseline” generates predictions uniformly at random.

7.2 Pre-trained models

We performed text classification using pre-trained BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM (Conneau and Lample 2019), and XLNet (Yang et al., 2019) models.

BERT is an attention-based auto-encoding sequence-to-sequence model using two unsupervised task objectives. The first task is word masking or Masked Language Model (MLM), where the model has to guess which word is masked in its position in the text. The second task is Next Sentence Prediction (NSP) performed by predicting if two sentences are subsequent in the corpus, or if they are randomly sampled from the corpus. The specific English BERT model used is the “bert-base-cased” in which capitalization is preserved as provided with the original paper.⁸

The XLM model we used is based on the work of Conneau and Lample (2019). We only used the monolingual MLM English model. The main difference with BERT is that XLM uses Byte-Pair Encoding (BPE) tokens as input instead of

⁸ <https://github.com/google-research/bert>.

WordPiece tokens and drops the NSP objective. The difference between BPE and WordPiece lies in the way the token symbol pairs are chosen for adding to the vocabulary. Instead of relying on the frequency of the pairs, WordPiece chooses the one which maximises the likelihood of the training data. These changes have been shown to improve performance on text-classification tasks. The specific model we used is “xlm-mlm-en-2048”.

The RoBERTa model (Liu et al., 2019) improved over BERT by dropping next sentence prediction and using only the MLM task on multiple sentences instead of single sentences. The authors argue that while NSP was intended to learn inter-sentence coherence, it actually learned topic similarity because of the random sampling of sentences in negative instances. The specific RoBERTa models used are the “roberta-base” and “roberta-large” based model provided alongside the original paper.⁹

XLNet is a permutation language model that combines strengths of autoregressive and auto-encoding modeling approaches: Permutation language models are trained to predict tokens given preceding context like a traditional unidirectional language model, but instead of predicting the tokens in sequential order, it predicts tokens in a random order sampling from both the left and right context. XLNet incorporates two key ideas from the TransformerXL architecture (Dai et al., 2019): relative positional embeddings and the recurrence mechanism. In combination with the permutation objective these techniques effectively capture bidirectional context while avoiding the independence assumption and pretrain-finetune discrepancy caused by the use of masked tokens in BERT. The specific XLNet models used are the “xlnet-base” and “xlnet-large” based models released alongside the original work.¹⁰

7.3 Results

Table 6 shows the macro-averaged precision, recall, and F_1 -scores for the highest-scoring hyper-parametrization of each model. Results show that RoBERTa (large 335M parameter variant, trained for 16 epochs) obtained the best results with a 60.1% (*std* 3%) macro-averaged F_1 -score on cross-validation and 59.1% on holdout. Not shown in Table 6, RoBERTa-Large obtained a micro-averaged F_1 -score of 62.1% (*std* 2.7) in cross-validation and 63.7% on holdout, and a ROC-AUC of 0.81 (*std* 0.02) in cross-validation and 0.90 on holdout. RoBERTa’s smaller base variant (parameters) and XLNet (base variant) also perform similarly while the other models trail behind. The other models are not as stable in cross-validation with a relatively large standard deviation in scores across folds. The stable results between the cross-validated development set and test set indicates the adequacy of the pre-determined train-test split for future benchmarking purposes.

⁹ <https://github.com/pytorch/fairseq/tree/master/examples/roberta>.

¹⁰ <https://github.com/zihangdai/xlnet/>.

Table 6 10-Fold cross-validation and held-out precision (P), recall (R), F_1 -score (macro-averaged) for sentence-level event detection on 18 classes

Model	Cross-validation			Holdout		
	P	R	F_1	P	R	F_1
Majority-class baseline	.005 $\pm$.00	.056 $\pm$.00	.010 $\pm$.00	.004	.056	.008
Stratified baseline	.039 $\pm$.01	.037 $\pm$.01	.036 $\pm$.01	.065	.043	.045
Uniform baseline	.040 $\pm$.01	.499 $\pm$.04	.071 $\pm$.01	.058	.489	.097
XLM	.422 $\pm$.28	.414 $\pm$.27	.404 $\pm$.27	.688	.489	.543
BERT-Large	.551 $\pm$.19	.527 $\pm$.18	.520 $\pm$.18	.721	.512	.579
XLNet-Large	.535 $\pm$.19	.558 $\pm$.19	.528 $\pm$.18	.670	.552	.576
XLNet-Base	.575 $\pm$.06	.610 $\pm$.04	.572 $\pm$.03	.737	.535	.585
RoBERTa-Base	.590 $\pm$.06	.626 $\pm$.04	.587 $\pm$.02	.686	.539	.590
RoBERTa-Large	.604 $\pm$.07	.639 $\pm$.04	.601 $\pm$.03	.683	.564	.591

The best-scoring hyper-parametrization for each model is shown

Boldface indicates highest score across models

Fig. 7 Macro-averaged precision, recall, F_1 -scores in cross-validation per type of the best system. (Only F_1 -score is labeled.)

Table 7 Qualitative error analysis of miss-classification types

Description	Freq	
Weak lexical cues: Absence of words that are common financial terminology overtly denoting an event type. Often due to anaphoric reference or highly general terms used.	63.6%	
Example	True label	Prediction
- <i>Within oncology , Imbruvica and Darzalex have seen solid growth of late , and we expect this trend to continue in 2018 .</i>	Revenue	Fin.Report
- <i>The bank ' s executives said Wednesday that the 2018 outlook is bright .</i>	Fin.Report	None
- <i>Chevron , however , would appear to have the edge in terms of taking those proved reserves and proving itself .</i>	Macroecon.	None
Plausible new label: The classifier predicts a label that is not in the set of true labels, but can reasonably be assigned to the instance. These cases happen usually when reference annotations failed to annotate the type because it was less salient/relevant.	12.41% of all miss-classification 34.6% of all miss-classification where a new label is assigned	
Example	True label	Prediction
- <i>Bank of America Earnings Hurt by Tax-Related Charge</i>	Profit/Loss	Legal, Profit/Loss
- <i>Another parameter to compare debt is the total debt-to-total capital ratio .</i>	None	Financing
- <i>FBR analyst Lucas Pipes updated his estimates for the company after its 3Q17 results and updated 2017 guidance.</i>	None	Fin.Report
Specialized context: High usage of specialized or idiomatic language such as industry/product-specific terminology in the sentence or proverbs.	17.7%	
Example	True label	Prediction
- <i>The biggest differences between the prosaic case and the Buck Rogers case lie in how long it might take Apple to double in value .</i>	Fin.Report	None
- <i>Newmont is pursuing greenfield exploration around the world and has added production with nine mine expansions in five years , as well as the purchase of the Cripple Creek & Victor mine in Colorado .</i>	Facility	Merger/Acq.
- <i>Meanwhile , PBMs that are part of UnitedHealth Group and Humana Inc. will be more shielded , Gupta said.</i>	Macroecon.	None
Ambiguous trigger: The event trigger words are ambiguous and commonly associated with other categories.	19.4%	
Example	True label	Prediction
- <i>Should Boeing Buy GE Aviation ?</i>	Merger/Acq.	Rating
- <i>Bank of America Earnings Hurt by Tax-Related Charge.</i>	Profit/Loss	Legal, Profit/Loss
- <i>Total spending in the App Store was approximately \$ 15 billion.</i>	SalesVolume	Expense

Description	Freq	
Weak lexical cues: Absence of words that are common financial terminology overtly denoting an event type. Often due to anaphoric reference or highly general terms used.	63.6%	
Example	True label	Prediction
- <i>Within oncology , Imbruvica and Darzalex have seen solid growth of late , and we expect this trend to continue in 2018 .</i>	Revenue	Fin.Report
- <i>The bank ' s executives said Wednesday that the 2018 outlook is bright .</i>	Fin.Report	None
- <i>Chevron , however , would appear to have the edge in terms of taking those proved reserves and proving itself .</i>	Macroecon.	None
Plausible new label: The classifier predicts a label that is not in the set of true labels, but can reasonably be assigned to the instance. These cases happen usually when reference annotations failed to annotate the type because it was less salient/relevant.	12.41% of all miss-classification 34.6% of all miss-classification where a new label is assigned	
Example	True label	Prediction
- <i>Bank of America Earnings Hurt by Tax-Related Charge</i>	Profit/Loss	Legal, Profit/Loss
- <i>Another parameter to compare debt is the total debt-to-total capital ratio .</i>	None	Financing
- <i>FBR analyst Lucas Pipes updated his estimates for the company after its 3Q17 results and updated 2017 guidance.</i>	None	Fin.Report
Specialized context: High usage of specialized or idiomatic language such as industry/product-specific terminology in the sentence or proverbs.	17.7%	
Example	True label	Prediction
- <i>The biggest differences between the prosaic case and the Buck Rogers case lie in how long it might take Apple to double in value .</i>	Fin.Report	None
- <i>Newmont is pursuing greenfield exploration around the world and has added production with nine mine expansions in five years , as well as the purchase of the Cripple Creek & Victor mine in Colorado .</i>	Facility	Merger/Acq.
- <i>Meanwhile , PBMs that are part of UnitedHealth Group and Humana Inc. will be more shielded , Gupte said.</i>	Macroecon.	None
Ambiguous trigger: The event trigger words are ambiguous and commonly associated with other categories.	19.4%	
Example	True label	Prediction
- <i>Should Boeing Buy GE Aviation ?</i>	Merger/Acq.	Rating
- <i>Bank of America Earnings Hurt by Tax-Related Charge.</i>	Profit/Loss	Legal, Profit/Loss
- <i>Total spending in the App Store was approximately \$ 15 billion.</i>	SalesVolume	Expense

7.4 Discussion and error analysis

The results of the best performing model were further analysed to gain more insights into the event type classification, the results of which are displayed in Fig. 7. At the left-hand side of the figure, the event type “CSR/Brand” (n = 365, 5.9% of all events) yields the lowest scores at a 33% F1-score with markedly low recall. “Financing” (n = 141, 2.3%), “FinancialReport” (n = 417 6.7%), “Deal” (n = 230 3.7%), “Investment” (n = 139, 2.2%), “Macroeconomics” (n = 827, 13.3%) have fairly low F1-score of around 50%. This is likely due to a combination of class infrequency and lexical variation observed by the corresponding lexical diversity in the event triggers.

We also performed a qualitative error analysis on miss-classifications in the output. We manually annotated a randomly selected sample of half the instances (169/335) where a classification error was made. This type of analysis allows to identify classification difficulties in context. The types of miss-classifications, examples and their frequency is given in Table 7.

Miss-classifications were most often caused by the absence of strong lexical cues (“weak cues”) which are lexical terms that strongly reference an event type e.g. “Hold” for “Rating”, “acquire” for “Acquisition”. This is due to anaphoric or less-specific references that are not captured within the sentence-level context and require document-level processing to resolve co-reference, which is lacking in our pilot classifier. Another type of error was caused by an ambiguous event trigger, e.g. “Buy” as a “Buy rating” or for “Merger/Acquisition”, “Spending” “SalesVolume” (customer spending) vs. “Expense”. We believe that both the weak cues and ambiguous triggers could be mitigated by machine learning methods which take into account the larger document-level context. Furthermore, entity and event co-reference processing would likely need to be incorporated. Some errors also occurred in a highly lexically specialized context with a lot of industry-specific terms or idiomatic expressions such as proverbs/sayings. The uniqueness of these expressions make it unlikely that these words or similar contexts were seen in the training set. This is a difficult problem to resolve with machine learning methods that are lexically-based.

Finally, in some instances a label outside the true label set was predicted that can plausibly be assigned. This usually happened when the plausible event type was not annotated because it was not deemed relevant, but strong lexical cues for that event were present. In rare cases, our evaluation set should be revised with these plausible types.

8 Data availability and replication

The annotation scheme guidelines, pilot study source code, sentence-level replication data, and trained models are publicly available at Jacobs (2020a). The original dataset will be freely downloadable at the end of the SENTiVENT project through this repository. Up until then, the fully annotated corpus is available on request for academic research purposes.

9 Conclusion and future work

Event extraction is a productive field of research and a required step in many data-driven tasks in which factual information is needed to capture changes in the real-world. Various general domain fine-grained event extraction corpora are freely available but no economically focused corpus exists. Currently, this resource scarcity limits the development of information extraction for economic and financial applications. The aim of the research presented in this paper was to construct a detailed, manually annotated, high-quality gold standard dataset for event extraction in the economic domain.

We present an annotation scheme of company-specific events in English economic news articles. A representative corpus was crawled and annotated with an iteratively developed economic event typology with 18 types and 64 subtypes. This resulted in around 6200 annotated events in 288 documents with event triggers, participant arguments, event co-reference, and event attributes such as type, subtype, negation, and modality. Agreement was generally substantial and annotator performance was adequate, indicating that the annotation scheme produces consistent event annotations of high quality. In an event classification pilot study, satisfactory results were obtained with a macro-averaged F_1 -score of 59% validating the dataset for machine learning purposes. This dataset provides a rich resource on events as training data for supervised machine learning for economic and financial applications.

In future work, we will apply recent methods for fine-grained event extraction based on neural attention-based methods. Extracting full event schemata remains a highly challenging task even on well-established datasets such as ACE2005. As in the current state-of-the-art event extraction methods, we will experiment with semi-supervised data generation methods.¹¹ This dataset provides a strong supervised seed set which should aid such approaches. We also plan on moving from purely data-driven to hybrid machine learning pipelines that integrate existing knowledge-based semantic resources such as DBpedia (Auer et al., 2007) and economic ontologies (Fani & Bagheri, 2015; Kakkonen & Mufti, 2011; Lösch & Nikitina, 2009).

In currently ongoing work, we are adding “investor sentiment” annotations to this corpus. We will add a “positive”, “neutral”, “negative” polarity attribute on top of events as well as separate sentiment expression annotations with their targets. These sentiment annotations allow the joint processing of “common-sense” sentiment of events and we will investigate how extracted event schemata can be used upstream from aspect-based sentiment analysis.

Acknowledgements This research has been carried out as part of a Ph.D. fellowship on the SENTiVENT project, funded by the Research Foundation—Flanders (cf. *supra*).

¹¹ During the decision time of this work, experimental work on fine-grained extraction was published in Jacobs and Hoste (2020).

Author contributions Not applicable.

Data availability Source code and pilot study experiment data is publicly available at Jacobs (2020a). Full annotated dataset will be made available after the end of the SENTiVENT project.

Code availability (software application or custom code): cf. supra.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., & Ellis, J. (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*, pp. 45–53.
- Arendarenko, E., & Kakkonen, T. (2012). Ontology-based information and event extraction for business intelligence. In: *Artificial Intelligence: Methodology, Systems, and Applications, Springer, Lecture Notes in Computer Science*, vol. 7557, pp. 89–102.
- Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In K. Aberer, K. S. Choi, N. Noy, D. Allemang, K. I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, & P. Cudré-Mauroux (Eds.), *The semantic web* (pp. 722–735). Berlin, Heidelberg: Springer.
- Ben Ami, Z., & Feldman, R. (2017). Event-based trading: Building superior trading strategies with state-of-the-art information extraction tools. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2907600, SSRN Working Paper 2907600.
- Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). *Text mining for central banks*. Available at SSRN 2624811.
- Bies, A., Song, Z., Getman, J., Ellis, J., Mott, J., Strassel, S., Palmer, M., Mitamura, T., Freedman, M., Ji, H., & Gorman, T. O. (2016). A comparison of event representations in DEFT. In: *Proceedings of the Fourth Workshop on Events*, pp. 27–36.
- Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. P. (2016). *Information, trading, and volatility: Evidence from firm-specific news*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2193667, SSRN Working Paper 2193667.
- Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2019). Information, trading, and volatility: Evidence from firm-specific news. *The Review of Financial Studies*, 32(3), 992–1033.
- Carroll, J. B. (1964). *Language and thought*. Hoboken: Prentice-Hall.
- Cavar, D., & Josefy, M. (2018). Mapping deep nlp to knowledge graphs: An enhanced approach to analyzing corporate filings with regulators. In: *Proceedings of the The first financial narrative processing workshop (FNP 2018)*.
- Chen, D., Zou, Y., Harimoto, K., Bao, R., Ren, X., & Sun, X. (2019). Incorporating fine-grained events in stock movement prediction. In: *Proceedings of the Second Workshop on Economics and Natural*

- Language Processing, Association for Computational Linguistics, Hong Kong*, pp. 31–40, <https://doi.org/10.18653/v1/D19-5105>. <https://www.aclweb.org/anthology/D19-5105>
- Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. In: *Advances in Neural Information Processing Systems*, pp. 7057–7067.
- Consortium, L. D. (2005). Ace (automatic content extraction) English annotation guidelines for events version 5.4.3. <https://www ldc.upenn.edu/sites/www ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf>
- Covington, M. A., & McFall, J. D. (2010). Cutting the gordian knot: The moving-average type-token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2), 94–100.
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy*, pp. 2978–2988, <https://doi.org/10.18653/v1/P19-1285>, URL <https://www.aclweb.org/anthology/P19-1285>
- Day, M., & Lee, C. (2016). Deep learning for financial sentiment analysis on finance news providers. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1127–1134.
- De Clercq, O., Lefever, E., Jacobs, G., Carpels, T., & Hoste, V. (2017). Towards an integrated pipeline for aspect-based sentiment analysis in various domains. In: *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 136–142.
- Del Corro, L., & Hoffart, J. (2020). Unsupervised Extraction of Market Moving Events with Neural Attention. arXiv e-prints [arXiv:2001.09466](https://arxiv.org/abs/2001.09466)
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186.
- Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. In: *Proceedings of the 24th International Conference on Artificial Intelligence, AAAI Press, IJCAI'15*, pp. 2327–2333. <http://dl.acm.org/citation.cfm?id=2832415.2832572>
- Ding, X., Zhang, Y., Liu, T., & Duan, J. (2016). Knowledge-driven event embedding for stock prediction. In: *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers*, pp. 2133–2142.
- Doddington, G. R., Mitchell, A., Przybicki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The automatic content extraction (ace) program—tasks, data, and evaluation. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.
- Du, M., Pivovarov, L., & Yangarber, R. (2016). PULS: natural language processing for business intelligence. In: *Proceedings of the 2016 Workshop on Human Language Technology*, pp. 1–8.
- Dugast, D. (1978). Sur quoi se fonde la notion d'étendue théorique du vocabulaire. *Français (Le) Moderne Paris*, 46(1), 25–32.
- Ein-Dor, L., Gera, A., Toledo-Ronen, O., Halfon, A., Sznajder, B., Dankin, L., Bilu, Y., Katz, Y., & Slonim, N. (2019). Financial Event Extraction Using Wikipedia-Based Weak Supervision. In: *Proceedings of the Second Workshop on Economics and Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA*, pp. 10–15, <https://doi.org/10.18653/v1/D19-5102>, [arXiv:1911.10783](https://arxiv.org/abs/1911.10783)
- Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *The Journal of Finance*, 48(5), 1749–1778.
- Fani, H., & Bagheri, E. (2015). An Ontology for Describing Security Events. *The 27th International Conference on Software Engineering and Knowledge Engineering, SEKE 2015, Wyndham Pittsburgh University Center, Pittsburgh, PA, USA, July 6-8, 2015*, pp. 455–460. <https://doi.org/10.18293/SEKE2015-101>
- Feldman, R., Rosenfeld, B., Bar-Haim, R., & Fresko, M. (2011a). The stock sonar sentiment analysis of stocks based on a hybrid approach. In: *Twenty-Third IAAI Conference*.
- Feldman, R., Rosenfeld, B., Bar-Haim, R., & Fresko, M. (2011b). *The Stock Sonar—Sentiment Analysis of Stocks Based on a Hybrid Approach*. Iaaai, pp. 1642–1647. <http://www.aaai.org/ocs/index.php/IAAI/IAAI-11/paper/view/3506/4024>
- Feuerriegel, S., & Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, 112, 88–97.

- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378.
- Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels—An empirical analysis. *Tourism Management*, 61, 43–54.
- Gémar, G., & Jiménez-Quintero, J. A. (2015). Text mining social media for competitive analysis. *Tourism & Management Studies*, 11(1), 84–90.
- Guiraud, P. (1959). *Problèmes et méthodes de la statistique linguistique*. Dordrecht: D. Reidel Publishing Company.
- Gwet, K. L. (2001). *Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters*. Gaithersburg, MD: STATAXIS Publishing Company.
- Gwet, K. L. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-rater Reliability Assessment Series*, 2(1), 9.
- Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Oxford: Advanced Analytics LLC.
- Han, S., Hao, X., & Huang, H. (2018). An event-extraction approach for business analysis from online Chinese news. *Electronic Commerce Research and Applications*, 28, 244–260. <https://doi.org/10.1016/j.elerap.2018.02.006>
- Herdan, G. (1964). *Quantitative linguistics*. Oxford: Butterworth.
- Hogenboom, A., Hogenboom, F., Frasincar, F., Schouten, K., & Van Der Meer, O. (2013). Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications*, 64(1), 27–52. <https://doi.org/10.1007/s11042-012-1122-0>
- Hogenboom, F., de Winter, M., Frasincar, F., & Kaymak, U. (2015). A news event-driven approach for the historical value at risk method. *Expert Systems with Applications*, 42(10), 4667–4675.
- Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339.
- Hu, Z., Liu, W., Bian, J., Liu, X., Liu, T. Y. (2018). Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In: *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 261–269.
- Jacobs, G. (2020a). Data for SENTiVENT: Enabling Supervised Information Extraction of Company-specific Events in Economic and Financial News. <https://doi.org/10.17605/OSF.IO/8JEC2>, <https://osf.io/8jec2/>
- Jacobs, G. (2020b). SENTiVENT Event Annotation Guidelines v1.1. Tech. rep., Ghent University. <https://doi.org/10.17605/OSF.IO/XQZW5>, <https://osf.io/xqzw5>
- Jacobs, G., & Hoste, V. (2020). Extracting fine-grained economic events from business news. In: *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation, COLING, Barcelona, Spain (Online)*, pp. 235–245, <https://www.aclweb.org/anthology/2020-fnp-1.36>
- Jacobs, G., Lefever, E., & Hoste, V. (2018). Economic event detection in company-specific news text. In: *Proceedings of the First Workshop on Economics and Natural Language Processing*, pp. 1–10.
- Johnson, W. (1944). I. A program of research. *Psychological Monographs*, 56(2), 1.
- Kakkonen, T., & Mufti, T. (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—i-KNOW '11*, p. 1. <https://doi.org/10.1145/2024288.2024318>, <http://dl.acm.org/citation.cfm?doid=2024288.2024318>
- Karami, A., Bennett, L. S., & He, X. (2018). Mining public opinion about economic issues: Twitter and the us presidential election. *International Journal of Strategic Decision Sciences (IJSDS)*, 9(1), 18–28.
- Khedr, A. E., Yaseen, N., et al. (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*, 9(7), 22.
- Kohlschütter, C., Fankhauser, P., & Nejdli, W. (2010). Boilerplate detection using shallow text features. In: *Proceedings of the third ACM international conference on Web search and data mining*, ACM, pp. 441–450.
- Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3), 411–433.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

- Lee, G. E., & Sun, A. (2019). A study on agreement in Pico span annotations. In: *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1, pp. 1149–1152, <https://doi.org/10.1145/3331184.3331352>, arXiv:1904.09557
- Lefever, E., & Hoste, V. (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)*, European Language Resources Association (ELRA), Portorož, Slovenia, pp. 330–335, <https://www.aclweb.org/anthology/L16-1051>
- Linguistic Data Consortium. (2015a). *DEFT Rich ERE Annotation Guidelines: Argument Filler V2.3. Tech. rep., Linguistic Data Consortium.*
- Linguistic Data Consortium. (2015b). *DEFT Rich ERE Annotation Guidelines: Events V3.0. Tech. rep., Linguistic Data Consortium.*
- Linguistic Data Consortium. (2016). *Rich ERE Annotation Guidelines Overview V4.2. Tech. rep., Linguistic Data Consortium.* Retrieved September, 9, 2018, from http://www.nist.gov/tac/2016/KBP/guidelines/summary_rich_ere_v4.2.pdf
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
- Liu, Z., Mitamura, T., & Hovy, E. (2015). Evaluation Algorithms for Event Nugget Detection : A Pilot Study. In: *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pp. 53–57. <http://www.aclweb.org/anthology/W/W15/W15-0807.pdf>
- Lösch, U., & Nikitina, N. (2009). The newsEvents ontology: an ontology for describing business events. In: *Proceedings of the 2009 International Conference on Ontology Patterns-Volume 516*, CEUR-WS. org, pp. 187–193.
- MacKinlay, A. C. (1997). Event studies in economics and finance. *Journal of Economic Literature*, 35(1), 13–39.
- Mass, H. D. (1972). Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, 2(8), 73.
- McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld). *Dissertation Abstracts International*, 66(12)
- McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–488.
- McCarthy, P. M., & Jarvis, S. (2010). MTLT, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. <https://doi.org/10.3758/BRM.42.2.381>
- Mian, G. M., & Sankaraguruswamy, S. (2012). investor sentiment and stock market response to earnings news. *The Accounting Review*, 87(4), 1357–1384.
- Mitamura, T., Liu, Z., & Hovy, E. H. (2015a). Overview of tac kbp 2015 event nugget track. In: *Proceedings of the 2015 Text Analysis Conference*. https://tac.nist.gov/publications/2015/additional.papers/TAC2015.KBP_Event_Nugget_overview.proceedings.pdf
- Mitamura, T., Liu, Z., & Hovy, E. (2016a). Overview of tac-kbp 2016 event nugget track. In: *Proceedings of the 2016 Text Analysis Conference*. https://tac.nist.gov/publications/2016/additional.papers/TAC2016.KBP_Event_Nugget_overview.proceedings.pdf
- Mitamura, T., Liu, Z., & Hovy, E. (2016b). *Overview of TAC-KBP 2016 Event Nugget Track. Tac Kbp 2016.*
- Mitamura, T., Liu, Z., & Hovy, E. H. (2017). Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. In: *Proceedings of the 2017 Text Analysis Conference*. https://tac.nist.gov/publications/2017/additional.papers/TAC2017.KBP_Event_Nugget_overview.proceedings.pdf
- Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., & Strassel, S. (2015b). Event nugget annotation: Processes and issues. In: *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, Association for Computational Linguistics*, Denver, Colorado, pp. 66–76. <https://doi.org/10.3115/v1/W15-0809>, <https://www.aclweb.org/anthology/W15-0809>
- Nardo, M., Petracco-Giudici, M., & Naltsidis, M. (2016). Walking down wall street with a tablet: A survey of stock market predictions using the web. *Journal of Economic Surveys*, 30(2), 356–369.

- Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653–7670.
- Nguyen, T. H., & Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1354–1364.
- Oberlechner, T., & Hocking, S. (2004). Information sources, news, and rumors in financial markets: Insights into the foreign exchange market. *Journal of Economic Psychology*, 25(3), 407–424.
- Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73, 125–144.
- Pavlopoulos, J., & Androutsopoulos, I. (2014). Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method. In: *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pp. 44–52.
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237.
- Plush, J. (2014). *Goose: Article extractor*. <https://github.com/GravityLabs/goose>
- Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al. (2016). Semeval-2016 task 5: Aspect based sentiment analysis. In: *10th International Workshop on Semantic Evaluation (SemEval 2016)*.
- Qian, Y., Deng, X., Ye, Q., Ma, B., & Yuan, H. (2019). On detecting business event from the headlines and leads of massive online news articles. *Information Processing and Management*. <https://doi.org/10.1016/j.ipm.2019.102086>
- Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: A review. *Journal of Research in Interactive Marketing*. <https://doi.org/10.1108/JRIM-05-2017-0030>
- Rönnqvist, S., & Sarlin, P. (2017). Bank distress in the news: Describing events through deep learning. *Neurocomputing*, 264, 57–70. <https://doi.org/10.1016/j.neucom.2016.12.110>, arXiv:1603.05670
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., & Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. ICSI UC Berkeley: Tech. rep.
- Shun, S. Y. (2019). *Lexicalrichness python module*. <https://github.com/LSYS/LexicalRichness>
- Somers, H. (1966). Statistical methods in literary analysis. In J. Leeds (Ed.), *The computer and literary style* (pp. 128–140). Kent, OH: Kent.
- Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., & Ma, X. (2015). From light to rich ERE: annotation of entities, relations, and events. In: *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pp. 89–98.
- 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.
- Thet, T. T., Na, J. C., & Khoo, C. S. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6), 823–848.
- Tobback, E., Naudts, H., Daelemans, W., Junqué de Fortuny, E., & Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. *International Journal of Forecasting*, 34(2), 355–365. <https://doi.org/10.1016/j.ijforecast.2016.08.006>. <http://www.sciencedirect.com/science/article/pii/S016920701630098X>
- Upadhyay, S., Christodoulopoulos, C., & Roth, D. (2016). “making the news”: Identifying noteworthy events in news articles. In: *Proceedings of the Fourth Workshop on Events*, pp. 1–7.
- Van de Kauter, M., Breesch, D., & Hoste, V. (2015). Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 42(11), 4999–5010. <https://doi.org/10.1016/j.eswa.2015.02.007>
- Van de Kauter, M., Desmet, B., & Hoste, V. (2015). The good, the bad and the implicit: A comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation*, 49(3), 685–720. <https://doi.org/10.1007/s10579-015-9297-4>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In: *Proceedings of the Advances in neural information processing systems*, pp. 5998–6008.

- Walker, C., S. Strassel, Medero, J., & Maeda, K. (2006). *Ace 2005 multilingual training corpus*. <https://catalog.ldc.upenn.edu/LDC2006T06>
- Wei, L., Li, G., Li, J., & Zhu, X. (2019). Bank risk aggregation with forward-looking textual risk disclosures. *The North American Journal of Economics and Finance*, 50, 101016.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). *Huggingface's transformers: State-of-the-art natural language processing*. ArXiv [arXiv:abs/1910.03771](https://arxiv.org/abs/1910.03771)
- Xie, B., Passonneau, R. J., Wu, L., & Creamer, G. G. (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)*, Association for Computational Linguistics, pp. 873–883. <http://www.aclweb.org/anthology/P13-1086>
- Xu, Y., & Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1970–1979.
- Yang, H., Chen, Y., Liu, K., Xiao, Y., & Zhao, J. (2018). Dcfec: A document-level chinese financial event extraction system based on automatically labeled training data. In: *Proceedings of ACL 2018, System Demonstrations*, pp. 50–55.
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *Xlnet: Generalized autoregressive pretraining for language understanding*. [arXiv:1906.08237](https://arxiv.org/abs/1906.08237)
- Zhang, X., Qu, S., Huang, J., Fang, B., & Yu, P. (2018). Stock market prediction via multi-source multiple instance learning. *IEEE Access*, 6, 50720–50728.
- Zhou, D., Chen, L., & He, Y. (2015). An unsupervised framework of exploring events on twitter: Filtering, extraction and categorization. In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence*.