Noname manuscript No.
(will be inserted by the editor)

Combining Machine Learning and Semantic Features in the Classification of Corporate Disclosures

Stefan Evert · Philipp Heinrich · Klaus Henselmann · Ulrich Rabenstein · Elisabeth Scherr · Martin Schmitt · Lutz Schröder

January 17, 2019

Abstract We investigate an approach to improving statistical text classification by combining machine learners with an ontology-based identification of domain-specific topic categories. We apply this approach to ad hoc disclosures by public companies. This form of obligatory publicity concerns all information that might affect the stock price; relevant topic categories are governed by stringent regulations. Our goal is to classify disclosures according to their effect on stock prices (negative, neutral, positive). In the study reported here, we combine natural language parsing with a formal background ontology to recognize disclosures concerning particular topics from a prescribed list. The semantic analysis identifies some of these topics with reasonable accuracy. We then demonstrate that machine learners benefit from the additional ontology-based information when predicting the cumulative abnormal return attributed to the disclosure at hand.

Keywords Ontology, Machine Learning, Corporate Disclosures, NLP

Stefan Evert and Philipp Heinrich (corresponding author)
Department Germanistik und Komparatistik, FAU Erlangen-Nürnberg
Bismarckstr. 6, 91054 Erlangen
Tel.: +49 9131 85 29250
E-mail: philipp.heinrich@fau.de

Klaus Henselmann and Elisabeth Scherr
School of Business and Economics, FAU Erlangen-Nürnberg

Ulrich Rabenstein and Lutz Schröder
Dept. of Computer Science, FAU Erlangen-Nürnberg

Martin Schmitt
Center for Information and Language Processing, LMU München

1 Introduction

Stock prices commonly react to information about important company events such as merger agreements, shifts in important personnel, or grants of material patents. Therefore, in order to avoid market distortions through transactions based on insider knowledge, most regulators world-wide require public companies to publish, without delay, any internal information that might affect the stock price. Nevertheless, it can be observed that markets do not include all the information that is publicly available, which can be explained by the limited information processing capability of humans (Bloomfield 2002). Intelligent decision support systems may therefore help improve capital market efficiency.

With the amount of electronically available information rising, there is increasing interest in developing new means of assessing the semantics of corporate disclosures in order to better handle the high information load. Prior research successfully explores the use of textual analysis to predict stock performance (Bollen et al 2010; Verchow 2011; Jegadeesh and Wu 2013; Ding et al 2015), often based on big data sources such as Twitter trends. Although our use case involves the prediction of an econometric variable as well, accurate prediction of stock prices is *not* our main goal. The broader aim of our work is to extract hidden information from financial texts; we therefore refrain from involving additional data sets such as social media to enhance performance.

In the study reported here, we aim to improve the performance of a statistical text classifier by integrating knowledge retrieved from the text by an ontology-based reasoner. Our use case is the prediction of stock market reactions after the publication of corporate events according to German law. In so-called *ad hoc disclosures*, companies have to report any important event that might affect the stock price. Although the relevant types of events are essentially predefined by law,[1] this information is not indicated explicitly in the disclosures and instead needs to be derived from the textual content.

Ad hoc disclosures are a suitable object of the investigation of combining machine learning with ontological reasoning for two reasons: Firstly, these disclosures are supposed to provide information relevant to the stock price and thus offer a straightforward task for machine learning and evaluation (the prediction of stock prices from text). Secondly, companies have an incentive to downplay negative events and hide them between the lines. Aiming to reveal hidden indicators in the ad hoc disclosures, we focus entirely on their textual content and, as indicated above, do not make use of external information from social media or other sources. This makes the prediction task fairly hard, and it is thus somewhat surprising that our trained classifiers do provide an effective trading strategy. Except for Verchow (2011), who uses only very basic computational linguistic methodology in his analysis of capital market efficiency and thus does not achieve high predictive accuracy, we are not aware of any prior work that attempts stock market prediction from ad hoc disclosures.

2 Methodology

We proceed to discuss the methodology of our analysis. The material is structured as follows: In Section 2.1 we give an overview of our corpus, the target variable, and the associated prediction task. We then briefly introduce the relevant machine learning techniques and their evaluation in Section 2.2. We motivate the usefulness of an ontology and outline our ontology design in Section 2.3. Section 2.4 concludes the methodological part by explaining the different ways of integrating machine learners with ontological features; we refer to statistical classifiers and their

[1] See the guideline issued by the Federal Financial Supervisory Authority (BaFin) (2009) for a list of possible price-sensitive events.

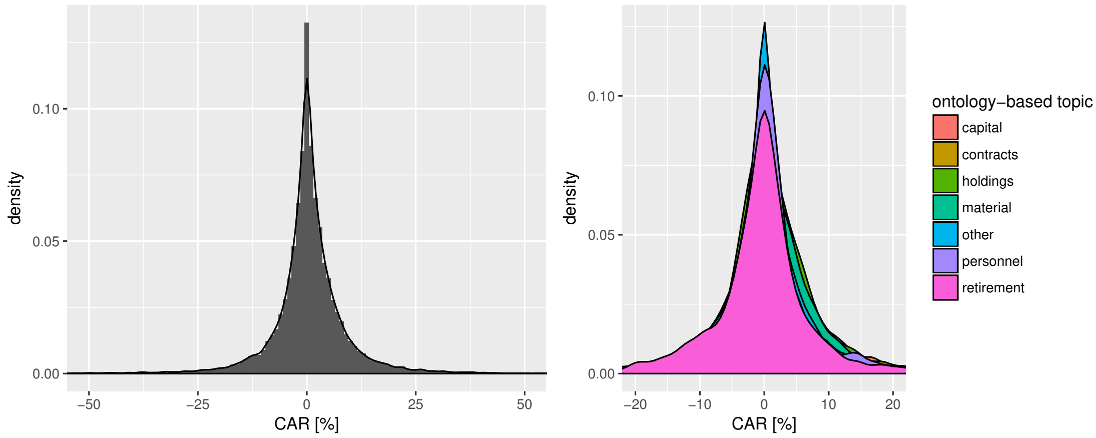

Fig. 1 Distribution of the target variable CAR in the corpus. Left hand panel: distribution in the whole corpus (excluding outliers with $|CAR_{it}| > 50$); right hand panel: distribution sorted by topic categories (excluding outliers with $|CAR_{it}| > 20$).

ontologically enhanced versions more generally as *task solvers*. In Section 2.4 we also present further evaluation techniques for the combination of machine learning (ML) with ontological information.

2.1 Data and Prediction Tasks

Corpus We use a sample of announcements of corporate events provided by the DGAP service of the Equity Story AG. Our sample selection starts with over 80,000 mandatory announcements of material events that have been disseminated via the DGAP between mid-1996 and mid-2012. We restrict our analyses to those disclosures that are machine-readable and written in English[2]. Due to these constraints, text-based deduplication and further restrictions concerning availability of metadata (see the following paragraph), we obtain a final corpus of 28,409 documents (*textual units*) such as the following example:

> Montabaur, December 31, 2001. Michael Scheeren, CFO of United Internet AG and with the company for 11 years, will retire from his position on the Executive Board as of December 31, 2001. It is planned that he will replace Mr. Hans-Peter Bachmann on the Supervisory Board from January 1, 2002. Scheeren will retain his close ties to the Group as he remains Chairman of the Supervisory Boards of AdLINK AG, 1&1 Internet AG and twenty4help AG. He will also represent United Internet AG on the Supervisory Boards of GMX AG, jobpilot AG and NTplus AG. Mr. Norbert Lang has been named as successor for Michael Scheeren. Lang has been with United Internet since 1994. After first heading the financial department, he joined the United Internet Executive Board one year ago.

Target variable For each ad hoc disclosure i, we measure its effect on the stock market using an event study following prior literature (Strong 1992; McWilliams and Siegel 1997; Corrado 2011). In particular, the market model is used to calculate the market-adjusted stock return surrounding the disclosure date t of the material event:

$$AR_{it} = R_{it} - E(R_{it}) = R_{it} - (\hat{\alpha}_i + \hat{\beta}_i \cdot R_{Mt}) \tag{1}$$

[2] German law requires the material event disclosures to be in German, in another accepted language or in English depending on specific criteria.

Daily market-adjusted returns or abnormal returns (AR_{it}) are calculated as the deviation between the observed stock return of each individual company (R_{it}) and the expected stock return ($E(R_{it})$). We use the return of the CDAX index[3] as a proxy for the market return and estimate $E(R_{it})$ by regressing a historic series of observed daily stock returns (R_{it}) on the corresponding daily market returns (R_{Mt}) using ordinary least squares (OLS) estimation. The estimation period starts 6 days ($t-6$) and spans up to 155 days ($t-155$) prior to the event date. The estimated intercept ($\hat{\alpha}_i$) and slope ($\hat{\beta}_i$) of the OLS model are then inserted into equation (1) to calculate the abnormal return (AR_{it}).

We use daily return index data from Thomson Reuters Datastream, which is adjusted for capital events (e.g., dividends, stock splits); daily returns are calculated as logarithmic returns (i.e. as $\log(V_f/V_i)$ where V_i is the initial and V_f the final value; use of this quantity is standard to ensure symmetry).[4] In order to account for the fact that part of the information relating to the event is priced early or late, we use an event window of three trading days. Hence, the cumulative abnormal return (CAR_{it}) surrounding each event announcement date (t) is calculated as the sum of the abnormal returns between one day prior ($t-1$) and one day after ($t+1$) the disclosure of the event. The distribution of the target variable is heavy-tailed, slightly skewed, and concentrates around 0 (cf. Figure 1).

Prediction tasks Although the target variable is metric, we abstain from regression analysis for two reasons: Firstly, the data shows heavy tails, which makes it difficult for regressors to find suitable weights. Secondly, it is far more important in practice to distinguish between positive and negative reactions than to predict the exact degree of the reaction. We hence set ourselves the prediction task of recognizing negative, neutral and positive responses based on a ternary categorization (1/3 of disclosures each); the categories are constructed by means of the respective quantiles of the empirical distribution of CAR.

Since these artificially created categories are hard to distinguish for machine learners – especially if the true CAR value is close to a category boundary – we also analyse the performance of the task solvers in a slightly modified prediction task with more clear-cut categories, i.e. ternary categorization into well-separated categories (20% of most negative and most positive reactions and the 20% closest to the median). We thus refer to the first one of these tasks as the *difficult* prediction task and to its modified version as the *easy* one (see Table 1 for an overview).

2.2 ML Classification

Our ML classifiers for solving the prediction tasks in Table 1 are primarily based on bag-of-words feature sets. After heuristic deletion of boilerplate footers and headers, removal of stop words, e-mail addresses, URLs, punctuation and numbers, as well as lower-casing and lemmatization as provided by the Stanford CoreNLP suite (Manning et al 2014) (cf. section 2.3), the resulting feature vocabulary contains $n_{voc} = 18{,}496$ lemmas which appear at least five times.

We additionally extract document-based sentiment polarity (in the form of numerical sentiment scores ranging from -1 to $+1$) using a simplified version of the SentiKLUE algorithm (Evert et al 2014). Furthermore, a set of stylometric features is generated with the help of the Multidimensional Analysis Tagger (Nini 2015). These 107 features were proposed by Biber (1988) for the multivariate analysis of text registers along dimensions such as the expression of narrative

[3] The *Composite DAX* (CDAX) is a stock market index based on German stocks that are listed in the General Standard or Prime Standard market segments, see `http://deutsche-boerse.com/dbg-en/about-us/services/know-how/glossary/glossary-article/CDAX/2560202`.

[4] Non-trading days are excluded.

	negative	neutral	positive	corpus
difficult	9,479	9,466	9464	28,409
material	*1,977*	*1,592*	*2,123*	*5,692 (20.0%)*
contracts	*1,354*	*1,547*	*1,797*	*4,698 (16.5%)*
capital	*1,098*	*906*	*959*	*2,963 (10.4%)*
retirement	*414*	*341*	*293*	*1,048 (3.7%)*
personnel	*283*	*317*	*260*	*860 (3.0%)*
holdings	*67*	*65*	*83*	*215 (0.7%)*
other	*4,286*	*4,698*	*3,949*	*12,933 (45.5%)*
easy	5,687	5,695	5,679	17,061 (60%)
material	*1,286*	*878*	*1,266*	*3,430 (20.1%)*
contracts	*743*	*914*	*1,122*	*2,779 (16.2%)*
capital	*633*	*521*	*592*	*1,746 (10.2%)*
retirement	*267*	*207*	*167*	*641 (3.8%)*
personnel	*162*	*186*	*147*	*495 (2.9%)*
holdings	*31*	*43*	*50*	*124 (0.7%))*
other	*2,565*	*2,946*	*2,335*	*7,846 (46.0%)*

Table 1 The two prediction tasks to be solved by the machine learning classifiers and their combinations with ontological features. The *easy* prediction task is a slight modification of the *difficult* one, involving more sharply separated categories. The rows in italics show the number of disclosures concerned with the respective topic category (cf. section 2.3).

and non-narrative concerns, the opposition of abstract versus non-abstraction information, or the overt expression of persuasion.

We also experiment with latent semantic indexing (LSI; Landauer et al 1998) in order to replace the very sparse bag-of-words model with a denser representation and make it easier to learn feature weights on relatively small training sets. LSI works by performing a truncated singular-value decomposition of the log(tf.idf)-weighted document-term-matrix. We preserve 100 components with the highest singular values for our experiments.

We present results for Logistic Regression (MaxEnt) with $\ell1$-penalty tuned by 10-fold cross-validation on the training set (for implementation details see Pedregosa et al 2011). Other machine learning algorithms such as Support Vector Machines and Multinomial Naïve Bayes as well as a modification of the latter as used by Verchow (2011) yielded similar results and are omitted here for the sake of brevity and clarity.

2.2.1 Evaluation of the ML approach

We use accuracy in 10-fold stratified cross-validation (90% training, 10% test data) as a performance measure and compute 95% confidence intervals for the mean accuracy across all 10 folds (based on a normal approximation). Since we have equally-sized categories and stratify the class distribution in the cross-validation,[5] a random baseline classifier achieves an accuracy of $1/3 = 33.3\%$ in our ternary classification tasks.

For the comparison of different classifiers – or, more precisely, different features matrices used by the same classifier – we use the McNemar test, which tests marginal homogeneity paired samples. It is applied to the 2×2 dichotomous contingency table of the classifiers, where the dichotomy is provided by a logical value indicating whether the prediction result is correct.

In order to demonstrate the practical usefulness of our ML approach, we also evaluate the machine learning classifier by means of a simple trading strategy: (1) *buy* if the ML approach predicts category *positive*, (2) *short-sell* if it yields *negative*, and (3) *hold* if the result is *neutral*.

[5] That is to say: all categories contain equal numbers of disclosures in each fold of the cross-validation.

A scalar performance measure is given by the sum of all individual net gains of CAR values.[6] In this setting, we use constant classifiers that always make the same decision as baselines.

2.3 Ontological Feature Extraction

Our idea is to use the semantic *topic categories* (not to be confused with the categorization of stock market reaction as per Section 2.2) that regulate the emission of disclosures in the first place in order to improve the ML classifiers. Recall that the disclosures are sent out for very specific reasons, but these are not explicitly mentioned in the text of a disclosure or in the associated metadata. Although the boundaries between different topic categories are somewhat fuzzy, most of the disclosures are sent out for one particular reason: manual analysis of a sample of 1,000 disclosures showed that only about 15% fall into more than one topic category.

2.3.1 Motivation for ontological feature extraction

The background information about the initial reason to send out the disclosures is valuable and provides a different sort of knowledge than the sort of "semantic information" that can be retrieved from the text itself by unsupervised learning (e.g. automatic clustering of the disclosures). Techniques such as Latent Dirichlet Allocation (LDA) or Latent Semantic Indexing are often found to be helpful in text classification because they reduce the high-dimensional bag-of-words feature space (with n_{voc} = 18,496 dimensions in our case) to a comparatively small number of *latent semantic* dimensions. Machine learners are expected to perform better because information is packaged more densely into the latent features and a smaller number of parameters needs to be trained. However, exploratory tests showed that our ML classifier does not benefit from such dimensionality reduction techniques, as can be seen in the ablation tests below.

We might also use topic modelling to pre-cluster the disclosures into meaningful categories. The result of an LDA is a vector for each document comprising the probabilities with which each of the topics has contributed to the creation of the document. The "mean latency" of a topic is thus the average probability of that topic across all documents. Each topic, in turn, has be interpreted in terms of word (or, in our case: lemma) lists; a priori, it seems unlikely that any unsupervised technique will yield stable and unambiguous results. Feuerriegel et al (2015), e. g., uses an LDA to gain 40 clusters which he then maps bijectively onto the pre-defined set of topics regulating their emission according to Federal Financial Supervisory Authority (BaFin) (2009). In a similar way, we use the LDA implementation as provided by Pedregosa et al (2011) on the tf.idf-weighted lemma vectors of all texts with standard parameters (20 passes). The most prominent topic in our corpus (according to our LDA model) with a mean latency of almost 25%, is made up of rather generic lemmas such as

> *product, service, technology, lead, position, agreement, new, future, work, solution, system, customer, provide, production, subsidiary, focus, ceo, industry, develop,* and *management.*

The topic above could e.g. be interpreted as *future products and contracts* or alike, yet there are clearly ambiguous and noisy terms such as *ceo*, *industry*, etc., which make the interpretation very speculative. The second most prominent topic with more than 19% mean latency contains the following lemmas:

> *earnings, previous, tax, ebit, figure, quarter, compare, profit, revenue, net, positive, income, first, month, rise, increase, operating, ebitda, result, fiscal.*

[6] The trading strategy rests on the assumption that we can buy or sell the shares after the material event and thus indeed collect the net gain of CAR values.

This topic points towards quarterly reports. However, neither of the topics point towards a clearly recognizable reason for the emission of a disclosure according to BaFin. Furthermore, the distribution of LDA topics on, e.g., the subset of disclosures that inform about retirements (see the following subsection) is almost identical to their distribution on the full corpus. The topics that can be retrieved from an LDA analysis thus do not help in recognizing particular topics that can be identified manually.

Therefore, we develop a formal ontology to retrieve meaningful semantic features. Since this is expensive with regards to implemenation effort, we concentrate on a few frequent and particularly interesting topic categories, namely

1. capital changes including adjustments to a company's capital (*capital*)
2. conclusion, amendment or termination of particularly important contractual relationships including cooperation agreements (*contracts*)
3. acquisition or disposal of major holdings (*holdings*)
4. material changes in results of financial statements or interim reports compared with previous results or market forecasts (*material*)
5. unexpected changes in key positions held within the company concerning, e.g., the chairman of the board of management, chairman of the supervisory board; or the resignation of the company's founder (*personnel*)
6. disclosures concerned with the retirement of key personnel (*retirement*) – this is, in fact, a subset of the topic category *personnel*, which we retrieve with very high precision and recall.

Any disclosure that cannot reasonably be assigned to any of the topic categories above by our ontology is classified as *other*. For the distribution of the topic categories in the corpus see Table 1. There is a low yet noteworthy association between topic categories and target variable, see the right hand side of Figure 1; *retirement* disclosures e. g. yield predominantely negative CAR values.

2.3.2 NLP pre-processing

The first step in operationalizing the corporate texts is a preprocessing stage in which disclosures are analysed using various off-the-shelf natural language processing techniques, including part-of-speech tagging, morphological analysis, named entity recognition, syntactic parsing, coreference resolution and word sense disambiguation.

The Stanford CoreNLP suite (Manning et al 2014) offers publicly available tools for the first five tasks. They are part of a pipeline architecture, i.e. every component can access the results of the previous components. For specifics on the POS tagger, see Toutanova and Manning (2000); the tagger uses the Penn Treebank tag set[7]. The NER software implements linear chain Conditional Random Field sequence models, see Finkel et al (2005); the entity categories are PERSON, ORGANIZATION, LOCATION and MISC. Furthermore, numerical entities are detected and classified into categories such as MONEY, NUMBER, DATE, TIME, DURATION and SET. We use the deterministic coreference resolution system described in Lee et al (2013). For word sense disambiguation, we use the algorithm described in Banerjee and Pedersen (2002) and the sense inventory of the lexical semantic database WordNet (Miller 1995).

In the ontological representation, the disambiguated words are mapped to WordNet concepts (*synsets*[8]). The ontology consists of three components:[9]

[7] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[8] *Synsets* are sets of synonyms representing a lexical semantic concept or word sense.
[9] *ABox* and *TBox* are the assertion and terminological components of the ontology, respectively.

– A TBox capturing relations among concepts, essentially obtained by extracting relevant information from WordNet for the terms encountered in the text.
– A manually maintained TBox capturing domain-specific background knowledge.
– An ABox recording the content of the parsed disclosures, generated from the NLP results.

We discuss these parts in more detail below.

2.3.3 Ontology creation from NLP results

We first describe the automatically generated parts of the ontology. It has to be emphasized that this ontology is not learned in any sense; rather, the procedure is essentially aimed at transforming linguistically analysed texts into the Web Ontology Language (OWL), additionally taking into account lexical semantic information from WordNet. WordNet information is provided in terms of a chain of subclass or subproperty inclusions connecting the word form actually appearing in a text to its synset identified by the word sense disambiguation module. E.g. for the word form *leaves* (possibly indicating a retirement event) this takes the following shape (Listing 1):

```
ObjectProperty: leave

ObjectProperty: leaves
  SubPropertyOf: leave

ObjectProperty: 2383440_Leave_depart_pull_up_stakes
  SubPropertyOf: leave
```

Listing 1 OWL representation of WordNet information for word form *leaves*.

The last, most specific object property relates to the synset corresponding to the relevant sense of *leaves*. It is composed of the synset's unique WordNet identifier (2383440), followed by the list of all synonyms in the set (to ensure human readability).

In this sense, we largely follow what Lüngen et al (2012) call the *class model* in the representation of WordNet content, i.e. we model synsets as classes and the hyponymy relation as subclass inclusion. As pointed out in *op. cit.*, this makes the modelling of other relationships between synsets, such as meronymy, slightly less straightforward; these are more naturally dealt with in the *instance model* where synsets become individuals in the ontology and hence can be directly related by arbitrary roles. We accommodate such relations in our model by implicitly eliminating occurrences of such individuals. For example, consider the meronymy relation, which in WordNet relates, e.g., 13104059_tree with 13128003_crown. We replace the (in our modelling, impossible) direct relation between these two synsets with an axiom regarding their inhabitants, specifically with

```
Class: 13104059_tree
    SubClassOf: hasPart some 13128003_crown
```

As indicated above, the NLP results are transformed into an ABox. The default procedure is to map subjects and objects of sentences, identified by the dependency analysis, to individuals in the ABox, whereas the verbs connecting subjects and objects become object properties. For prepositional objects, the preposition is made part of the object property, which is then named in the form $\langle$verb$\rangle$_$\langle$preposition$\rangle$. If the auxiliary verb *will* is detected in connection with the verb (e.g. if the disclosure states that the CEO *will* resign rather than that he has already resigned), the object property is named announced_$\langle$X$\rangle$, where X is the original name of the object property, and marked as being a subproperty of both announced and X. Subjects and objects receive as a type the concept generated from their synset according to the word sense disambiguation, and receive as facts their mutual relationship as specified by the synset of the verb. For example,

Individual: John_Doe
  Types: Person
  Facts: 2383440_leave_depart_pull_up_stakes company

Individual: company
  Types: 8058098_Company

Listing 2 ABox representation of sentence *John Doe leaves the company.*

Individual: John_Doe
  Types: Person, CFO
  Facts: 2383440_leave_depart_pull_up_stakes company.
    CFO_of company

Listing 3 Extended ABox for sentence *John Doe leaves the company.*

the sentence *John Doe leaves the company* with the syntactic dependency analysis in Figure 2 is translated into the ABox depicted in Listing 2.

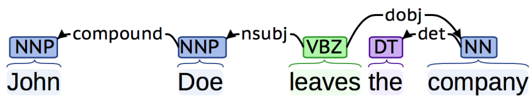

Fig. 2 Dependency parse of *John Doe leaves the company.*

Note that each syntactic dependency connects only two words. For compound nouns, the rightmost noun is regarded as the head noun, and the other component nouns are linked to the head noun via a *compound* relation. Compound nouns have to be recomposed from the syntactic dependencies, which results in the individual *John Doe* rather than just *Doe*. Coreferences are resolved while creating the ontology, so facts referring to a pronoun are attached to the corresponding individual.

In this case, the types of the individuals are inferred from named entity recognition (Person) and morphological analysis (Company). Appositions are also used to infer types: the dependent of an apposition determines an additional type for its governor, and triples describing the dependent are assigned to the governor. Prepositional triples are prefixed by the dependent of the apposition. For instance, from the phrase *John Doe, CFO of the company*, one obtains the dependency relations

$$\text{appos}(\mathit{Doe}, \mathit{CFO}) \quad \text{and} \quad \text{of}(\mathit{CFO}, \mathit{company}),$$

which extend the knowledge about John Doe in the way depicted in Listing 3.

The previous examples always contained the main piece of information, e.g. on someone leaving a company, or doing something in general, in a subject-predicate-object-like structure. A sentence like “He announced the retirement of John Doe” does not fit into this pattern. Therefore our system uses derivational relations from WordNet to transform triples like *of(retirement, John Doe)* into a subject-predicate-object structure, *retire(John Doe, dummy)*. The dummy individual is needed because the intransitive verb *retire* (from which *retirement* is derived) does not take an object. This type of normalization simplifies querying the assertional knowledge parsed from the text in subsequent steps.

```
Class: 9916601_chief_financial_officer_cfo
    EquivalentTo: works_on some Cfo_position
    SubClassOf: works_on exactly 1 Executive_board_position

Class: Cfo_leave1
    EquivalentTo: leave some Cfo_position,
                  Cfo and leave some Executive_board_position

Class: Cfo_leave2
    EquivalentTo: Cfo and (leave some Executive_board),
                  leave  some Cfo_position

Class: leave3
    EquivalentTo: (have some (Contract and expire some owl:Thing)),
    SubClassOf: leave some Position

Class: leave4
    EquivalentTo: agree some (Termination and (of some Mandate)),
    SubClassOf: leave some Position

Class: leave5
    EquivalentTo: submit some Resignation,
    SubClassOf: leave some Position
```

Listing 4 Excerpt from the background ontology.

2.3.4 Alternative ABox model

Alternatively, we can populate the ABox with all chunks parsed from the text, and relate these individuals by their syntactic dependencies (and additionally by connecting prepositions). This leads to an ABox that reflects the structure of the actual text more closely and refrains from (partly speculative) extraction of putative facts. We then build some infrastructure over the object properties to ease the axiomatics and queries; in particular, we introduce an object property hasFrameElement that subsumes essentially all syntactic dependencies. This approach necessitates a slight variation of the way background knowledge is phrased; we discuss these issues further below.

2.3.5 Background knowledge

The system is supported by a static, manually maintained background ontology capturing general and domain knowledge that is not explicit in the text of the disclosures. Some of the relevant facts are quite simple, e.g. that stepping down is a form of leaving and that "Executive Board" and "Management Board" are synonyms. Other axioms are more interesting and capture combinations of standard jargon with basic knowledge of the domain. E.g. at the domain-specific level we include axioms saying that CFOs work on exactly one executive board position, and that they retire from their CFO position iff they retire from their executive board position.[10] At a less specific level there are axioms saying that, e.g., letting your contract expire, agreeing to the termination of your mandate, and submitting your resignation all amount to leaving your current position. The formulation of statements such as these is illustrated in Listing 4.

[10] The universal validity of these axioms may be debatable but since OWL does not incorporate default reasoning, there appears to be no realistic way to ensure stricter accuracy.

2.3.6 Background knowledge over the alternative ABox model

Besides the philosophical distinction between having chunks of text rather than real-world invididuals as the putative inhabitants of classes in the ontology, the main technical difference encountered in the alternative ABox model described above is that verbs are modelled as individuals rather than object properties. This enables a more flexible axiomatization of verbs – while for object properties, one essentially has only hierarchy axioms, (classes of) individuals can be axiomatized using the full power of OWL class expressions. We give two examples, both taken from the background ontology on the topic of changes in contractual relationships (*contracts*). The following class definition captures a wide-spread way of expressing the conclusion of new contractual agreements:

```
Class: agreement_event
    EquivalentTo: Agree and has_frame_element some
            (Legal_document_agreement or 6770875_condition)

Class: Legal_document_agreement
    EquivalentTo: 6770275_agreement or 6479665_legal_document
```

That is, an agreement_event is an occurrence of the lemma *agree* that has a syntactic dependency with a lemma from at least one of three prescribed synsets. A more involved example concerns general contractual changes:

```
Class: Contractual_changes_frame
    EquivalentTo: Something_changes_frame and Concerns_company_contract

Class: Something_changes_frame
    EquivalentTo: 2427103_establish or 2608347_begin
            or 2609764_end or 126264_change or 209943_termination

Class: Concerns_company_contract
    EquivalentTo: (has_frame_element some (Legal_document_agreement
                and Concerns_companies))
        or ((has_frame_element some Legal_document_agreement)
                and Concerns_companies)
        or (has_frame_element some 6770875_condition)

Class: Concerns companies
    EquivalentTo: (preposition some NER_organization)
            or (concerns some Partnership)
```

In words, a contractual change is an occurrence of a lemma that belongs to one of a number of specified synsets indicating temporal change and concerns company contracts. The latter is defined as either being syntactically connected to a legal document or agreement concerning a company, or concerning the company and being syntactically connected to a legal document or agreement, or being syntactically connected to an occurrence of a lemma belonging to the synset *condition* (6770875). Finally, 'concerning a company' is defined by two alternatives, one of which requires being connected via a preposition to an organization previously identified as a named entity.

2.3.7 Querying

With the ontology in place, we can now detect disclosures concerning the respective topics. Most of the time this just means that we query for a topic number recognized by the ontology:

```
SELECT DISTINCT ?person ?leave ?object WHERE{
    ?person ?leave ?object.
    ?person a :Person.
    ?leave rdfs:subPropertyOf :leave.
    FILTER NOT EXISTS{ ?person ?leave2 ?object.
        ?leave2 rdfs:subPropertyOf ?leave.
        FILTER NOT EXISTS{?leave2 owl:equivalentProperty ?leave. }}}
```

Listing 5 Query detecting retirements

```
SELECT DISTINCT ?event ?topic_number WHERE {
    ?event a :Topic_of_interest.
    ?event :has_topic_number ?topic_number.
    }
```

Here, the association of topics with topic numbers is managed via the ontology: e.g. the topic number 13 'contractual change' is governed by the following axioms (among others):

```
Class: Contract_event
    SubClassOf:
        Topic_of_interest,
        has_topic_number value 13

Class: agreement_event
    EquivalentTo : Agree and has_frame_element some
        (Legal_document_agreement or 6770875_condition)
    SubClassOf:
        Contract_event
```

The first subclass axiom just associates contract change events with the mentioned topic number; the second subclass axiom identifies a particular type of contractual changes, events regarding agreements (with the definition as already shown on p. 11).

In the sample case of retirement of key personnel, we have a more fine-grained modelling that allows retrieving also data regarding the concrete instance of the topic, in this case the retiring person and the position retired from. The corresponding query is shown in Listing 5. The filter statements serve to eliminate multiple results that differ only in the value of the *?leave* variable, i.e. use different subproperties of *leave* but refer to the same person and position. That is, the query is set up in such a way as to return only the triples with the most specific object property as the instantiation for *?leave*. In case the query returns any result, the ad hoc disclosure is marked as containing a message about a retirement. In case the instantiation of *?leave* is a subproperty of *announced*, the disclosure is additionally annotated as being (only) an announcement (this feature is not shown in Listing 5).

2.3.8 Evaluation of the ontological approach

The ontology has been tested against a manual classification of a subset of 1,000 disclosures. Ontological detection of topics was most successful on a topic that is slightly more specific than the *changes in key personnel* topic in the list of prescribed topics (*personnel*), the above-mentioned topic of *retirement of key personnel* and announcenemts thereof. Recognition of this topic was tested on a set of 300 messages containing any inflected form of the words *leave* or *retire*. The disclosures were categorized manually as retirement (178 messages) or non-retirement (122 messages). The low baseline accuracy of 59.3% shows that the mere occurrence of the keywords *leave* and *retire* is not a reliable predictor. Our algorithm obtained recall and precision values

topic category	precision	recall
holdings	42.9%	27.8%
capital	34.9%	47.5%
material	93.5%	56.1%
contracts	23.8%	59.1%
personnel	75.0%	54.6%
retirements	97.0%	90.4%

Table 2 Evaluation of the ontological approach in terms of precision and recall. The *retirements* topic category has been modelled more thoroughly than the other ones, and can be recognized with fairly high precision and recall.

of 90.4% and 97% for retirement events, respectively.[11] Regarding the additional property of retirements being only announced rather than already realized, 75 of 139 messages were successfully identified as (only) announcing at least one retirement, and 6 were falsely classified as (mere) announcements (recall 54%, precision 92.5%). It is, of course, not entirely surprising that automated detection of the event ("retirement") as such works better than automated detection of the much more abstract question of its factuality.

We subsequently used the alternative ABox model described above to recognize the topics listed at the beginning of the section. Overall, precision is fairly low for the very generally modelled topic categories *holdings*, *capital*, and *contracts*, ranging between 23.8% and 42.9%. It is substantially higher for the topic categories *personnel* (75.0%) and *material* (93.5%), although recall is low throughout all topic categories, ranging from 27.8% (*holdings*) to 59.1% (*contracts*), see Table 2. On the other hand, overall accuracy was well above 50%, and it turns out that the ontological features are helpful in the classification despite their relatively low recall.

2.4 Integration Methods

We now turn to equipping the ML classifiers with the ontologically extracted topic features in order to improve their performance. Our idea is that the ontological information about the types of material events that regulate the dissemination of the disclosures in the first place can be used for splitting the overall problem into smaller sub-problems: A machine learner trained solely on disclosures about a certain topic such as *retirements* or *capital* is confronted with an easier task than a system that does not have any information about the reasons for the dissemination of the disclosures at hand; just as a human expert confronted with very specific disclosures has an easier task than someone who is confronted with an unstructured bulk of disclosures.

Our first combination of ML and ontology is by means of adding a single ontological feature to the document-lemma feature matrix (**single ontological feature**). However, since a single ontological feature can easily be overseen amongst other features, we experiment with a separation of the various topic vocabularies: If a disclosure is recognized by the ontological model as being about a *retirement*, for instance, the string `retirement` is appended to each lemma in the text (**separate vocabulary**).

This method has the disadvantage that the ML classifier cannot generalize information about the general meaning of lemmas (e.g. *risk* or *losses*) gathered from the much larger remainder of the corpus to the retirement disclosures. It is likely to underperform in this setting because it is effectively restricted to a small training corpus. We thus consider a third combination method that mirrors the retirement vocabulary (**mirrored vocabulary**): All disclosures now

[11] 161 of the 178 true retirement messages were detected correctly by the algorithm (true positives) while 5 disclosures were incorrectly marked as retirements (false positives).

retain the original lemmas, but are complemented with an *additional* category-specific vocabulary. In the example disclosure on p. 3, lemmas such as $\mathit{Montabaur}_{\text{retirement}}$, $\mathit{CFO}_{\text{retirement}}$, and $\mathit{company}_{\text{retirement}}$ are added without deleting the original lemmas.

To put it in other words, the reasoning behind separate and mirrored vocabularies is as follows: A *single* ontological feature might not be recognized efficiently by a machine learner. A *separate* vocabulary, moreover, discriminates against disclosures that belong to small topic categories, since the machine learner cannot exploit features from the bigger part of the training corpus; as a result, the amount of training data for lemmas in the class-specific vocabulary is drastically reduced. The setting of separate vocabularies for each class is equivalent to training a different machine learner on each topic category. Last but not least, the ontological features are weight *adjustments* in the case of the mirrored vocabulary: Here the machine learner can learn features both on the specific topic and from the whole corpus and can then exploit this knowledge for all disclosures.

Including the basic feature matrix (**bag of words**) and the one with reduced dimensions by means of an SVD (**LSI**), there are thus five feature matrices and their combinations with the sentiment and stylistic features that can be used for prediction (see Table 3 for an overview).

feature matrix	*dimensionality*
LSI	100
bag of words	18,496
single ontological features	18,503
separate vocabulary	39,017
mirrored vocabulary	57,513

Table 3 The five basic different feature matrices used for prediction.

2.4.1 Evaluation of integration methods

Since the integration methods essentially differ only in the feature used, they can be compared to the original classifiers (and their respective baselines) in a straightforward manner by means of accuracy in 10-fold cross-validation. Moreover, for retirement disclosures, another baseline is readily at hand. Since the retirement feature is weakly yet significantly associated with the target variable, retirement disclosures can be classified *ontologically*: The greater part of retirement disclosures lead to a negative stock market reaction (cf. Table 1), so that the ontology already outperforms the baseline by assigning the category *negative* to all retirement disclosures.

3 Results and Discussion

There are two kinds of effects to be analysed: Firstly, the effect of ontological information on the prediction quality of task solvers can be quantified. Secondly, one can observe how the feature weights are affected by the additional information, which gives interesting insights into the different usage of language in particular discourse topic domains.

3.1 Prediction Results

The complete results in terms of accuracy can be found in Table 4 for the difficult and the easy prediction task. Comparing the different prediction tasks with one another, one can unsurprisingly see that performance is higher in the case of clear-cut categories. Generally, the machine learners consistently outperform the random baseline (approximately 1/3).

features	**difficult**	**easy**
baseline	0.3337 ± 0.0002	0.3338 ± 0.0005
LSI	0.3959 ± 0.0158	0.4320 ± 0.0212
+ stylistics	0.3941 ± 0.0148	0.4208 ± 0.0231
+ sentiment	0.4240 ± 0.0136	0.4596 ± 0.0264
+ stylistics + sentiment	0.3944 ± 0.0144	0.4218 ± 0.0215
bag of words	0.4389 ± 0.0166	0.4765 ± 0.0309
+ stylistics	0.4382 ± 0.0186	0.4793 ± 0.0281
+ sentiment	0.4388 ± 0.0160	0.4776 ± 0.0304
+ stylistics + sentiment	0.4388 ± 0.0185	0.4790 ± 0.0300
single ontological feature	0.4401 ± 0.0161	0.4763 ± 0.0316
+ stylistics	0.4383 ± 0.0188	0.4797 ± 0.0297
+ sentiment	0.4399 ± 0.0156	0.4769 ± 0.0309
+ stylistics + sentiment	0.4384 ± 0.0180	0.4793 ± 0.0301
separate vocabulary	0.4241 ± 0.0230	0.4582 ± 0.0256
+ stylistics	0.4272 ± 0.0215	0.4615 ± 0.0279
+ sentiment	0.4242 ± 0.0232	0.4591 ± 0.0261
+ stylistics + sentiment	0.4276 ± 0.0204	0.4623 ± 0.0279
mirrored vocabulary	0.4413 ± 0.0238	0.4821 ± 0.0308
+ stylistics	**0.4432** ± 0.0211	**0.4840** ± 0.0271
+ sentiment	0.4419 ± 0.0232	0.4828 ± 0.0300
+ stylistics + sentiment	0.4429 ± 0.0204	0.4837 ± 0.0286

Table 4 Performance (mean accuracy and 95% confidence interval) of the Machine Learner in the different prediction tasks on the whole corpus using different feature matrices. The naïve baseline (majority classifier) is given by $1/3 = 33.3\%$.

For instance, MaxEnt with the bag of words as feature set significantly outperforms the 1/3 baseline in the easy prediction task with an accuracy of **47.65%** (± **3.09%**). Moreover, Table 5 shows that we obtain substantial net gain when using the machine learner with any feature matrix and in any task and applying the trading strategy outlined in section 2.2 (except for the LSI features without any additional information): the accumulated continuous returns of the task solver result in a mean profit of more than 0.2% per disclosure and a considerable increase of our start capital despite the simple approach.

It is worth mentioning that Latent Semantic Indexing leads to significant performance losses compared to the bag of words. It is also in this setting that a combination with sentiment features yields huge gains, presumably because this (mostly lexical) information is not present in the latent LSI dimensions. In all the other settings, enriching the features sests with stylistic and/or sentiment features does not lead to a significant improvement.

Comparing the different feature sets with one another, we see in Table 4 that our hypotheses from Section 2.4 are confirmed: Adding a single ontological feature does not change the results compared to the bag of words significantly; separating the vocabularies from one another, on the other hand, has the anticipated adverse effect. The machine learner is not able to generalize from the rather small training corpora.

Furthermore, mirroring the vocabulary indeed shows significantly better results compared to the simple bag of words approach (and the LSI and separate vocabulary approaches). A McNemar test shows e.g. a significant p-value of $p = 0.039$ when comparing the mirrored vocabulary in combination with Biber's stylistic features to the simple bag-of-words approach (including stylistic features) in the easy prediction task.

features	difficult	easy
baseline	-1409.58 ± 1137.06	0.00 ± 0.00
LSI	-492.14 ± 971.85	218.44 ± 1087.04
+ stylistics	1016.77 ± 1318.74	1217.99 ± 716.53
+ sentiment	1662.78 ± 1427.92	1927.79 ± 1284.49
+ stylistics + sentiment	1100.12 ± 950.66	1245.85 ± 763.55
bag of words	3082.89 ± 1154.76	2968.24 ± 896.75
+ stylistics	3080.50 ± 1145.08	3029.92 ± 860.70
+ sentiment	3082.23 ± 1154.12	2973.58 ± 944.87
+ stylistics + sentiment	**3090.17** ± 1133.78	3013.33 ± 947.50
single ontological feature	3094.65 ± 1287.55	2936.11 ± 1133.62
+ stylistics	3046.95 ± 1148.55	3021.04 ± 1005.07
+ sentiment	3074.04 ± 1216.37	2983.95 ± 1061.87
+ stylistics + sentiment	3090.68 ± 1158.80	3037.53 ± 1060.41
seperate vocabulary	2238.75 ± 898.62	2293.81 ± 1074.88
+ stylistics	2425.94 ± 1080.95	2414.40 ± 1162.22
+ sentiment	2255.96 ± 868.49	2288.05 ± 1138.08
+ stylistics + sentiment	2460.95 ± 1072.04	2501.86 ± 1119.19
mirrored vocabulary	2936.67 ± 1135.44	3101.66 ± 1080.74
+ stylistics	2960.48 ± 1054.11	3164.38 ± 936.71
+ sentiment	2959.08 ± 1102.65	3117.88 ± 1059.12
+ stylistics + sentiment	2963.26 ± 1001.52	**3157.96** ± 922.07

Table 5 Performance in terms of economic profit according to the trading strategy outlined in Section 2.2 the Machine Learner could realize using different feature matrices. The naïve baseline (majority classifier) is given by a trading strategy that always only buys if it encounters predominantly *positive* CAR values in the training corpus, and short-sells if it encounters predominantly *negative* ones.

3.2 Feature Weight Analysis

Which lemmas change their feature weights? In the present setting, we identified lemmas whose feature weights are substantially different in *retirement* disclosures than in the topic category *other* (which can be seen from the separate vocabulary), or which obtain a relatively high ‘adjustment’ weight in the mirrored retirement vocabulary (the best-performing feature set). Table 6 shows such lemmas based on feature weights for the category *positive*. Results for category *negative* are omitted since they show similar patterns.

For example, the lemmas *exceed* (1.293 for category *positive* in the vanilla bag of words) and *improvement* (0.708 in the bow) are generally associated with a positive CAR response. However, the mirrored vocabulary reduces these weights in retirement disclosures by -0.019 for *exceed* and -0.014 for *improvement*, showing that they imply a different outcome for this event type (see section 2.4 for an explanation of weight adjustments).[12] The relatively small adjustement is probably due to the low overall proportion of retirements, and the effect becomes much clearer with a separate vocabulary: the feature weights on retirement disclosures are -0.021 (*exceed*) and -0.018 (*improvement*), respectively, showing that these lemmas no longer indicate a positive stock market reaction. Similarly, the lemma *insolvency* is generally associated with a negative reaction, but indicates a positive response when used in retirement disclosures.

[12] The setting with the single feature is omitted here because its lemma feature weights are almost identical to the vanilla bag of words. Recall that we just add a single feature indicating the ontological category in this feature set, so it cannot account for the differences in language use between retirements and other messages that we are interested in here.

lemma	vanilla bow	separate voc.		mirrored voc.	
		other	*retirement*	*other*	*retirement*
exceed	1.293	1.293	-0.021	1.293	-0.019
fall	-0.864	-0.842	-0.034	-0.855	-0.027
career	0.090	-0.033	0.115	0.044	0.089
improvement	0.708	0.696	-0.018	0.700	-0.014
rise	0.612	0.616	-0.024	0.614	-0.023
weak	-0.769	-0.766	-0.012	-0.769	-0.009
lower	-1.022	-1.012	-0.041	-1.018	-0.028
positive	1.149	1.130	-0.007	1.137	-0.015
insolvency	-0.386	-0.447	0.081	-0.417	0.059

Table 6 Lemmas whose feature weights for category *positive* are substantially different in *retirement* disclosures from their vanilla feature weights.

4 Conclusion

Machine learners are used in many prediction tasks of computational linguistics. We have combined a semantics-based approach to recognition of message content with a machine-learning classification of documents, specifically of corporate disclosures according to their effect on the stock price.

Machine learners benefit from ontological information since it enables them to deal with more specific range of language use. The core idea tested in our feasibility study is that words are used more consistently within the specific domain of the topic pre-defined by the ontology. The effect on prediction accuracy is small but consistent.

Future work will be aimed partly at refining the ontological approach to improve its precision and recall (both already above 90% on one target feature, retirements, but not achieving comparable performance for more general topics). Moreover, we will strive to develop new methods for exploiting the subjective use of language in different domains in order to improve prediction accuracy.

References

Banerjee S, Pedersen T (2002) An adapted Lesk algorithm for word sense disambiguation using WordNet. In: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, Springer-Verlag, London, UK, CICLing '02, pp 136–145, URL http://dl.acm.org/citation.cfm?id=647344.724142

Biber D (1988) Variation across Speech and Writing. Cambridge University Press, Cambridge

Bloomfield R (2002) The 'incomplete revelation hypothesis' and financial reporting. Accounting Horizons 16:233–243

Bollen J, Mao H, Zeng XJ (2010) Twitter mood predicts the stock market. Journal of Computational Science 2(1):1–8, URL http://arxiv.org/pdf/1010.3003v1.pdf

Corrado C (2011) Event studies: A methodology review. Accounting and Finance 51:207–234

Ding X, Zhang Y, Liu T, Duan J (2015) Deep learning for event-driven stock prediction. In: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (ICJAI), pp 2327–2333, URL http://ijcai.org/papers15/Papers/IJCAI15-329.pdf

Evert S, Proisl T, Greiner P, Kabashi B (2014) SentiKLUE: Updating a polarity classifier in 48 hours. In: Nakov P, Zesch T (eds) Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Association for Computational Linguistics, Dublin, pp 551–555, URL http://www.aclweb.org/anthology/S14-2096

Federal Financial Supervisory Authority (BaFin) (2009) Issuer guidelines

Feuerriegel S, Ratku A, Neumann D (2015) Which News Disclosures Matter? News Reception Compared Across Topics Extracted from the Latent Dirichlet Allocation. News Reception Compared Across Topics Extracted from the Latent Dirichlet Allocation (February 13, 2015) URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2564603

Finkel JR, Grenager T, Manning C (2005) Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '05, pp 363–370, DOI 10.3115/1219840.1219885, URL https://doi.org/10.3115/1219840.1219885

Jegadeesh N, Wu D (2013) Word power: A new approach for content analysis. J Financial Economics 110:712–729

Landauer TK, Foltz PW, Laham D (1998) An introduction to latent semantic analysis. Discourse Processes 25(2-3):259–284, DOI 10.1080/01638539809545028

Lee H, Chang A, Peirsman Y, Chambers N, Surdeanu M, Jurafsky D (2013) Deterministic coreference resolution based on entity-centric, precision-ranked rules. Comput Linguist 39(4):885–916, DOI 10.1162/COLI_a_00152, URL http://dx.doi.org/10.1162/COLI_a_00152

Lüngen H, Beißwenger M, Selzam B, Storrer A (2012) Modelling and processing wordnets in OWL. In: Mehler A, Kühnberger K, Lobin H, Lüngen H, Storrer A, Witt A (eds) Modeling, Learning, and Processing of Text Technological Data Structures, Studies in Computational Intelligence, vol 370, Springer, pp 347–376

Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp 55–60, URL http://www.aclweb.org/anthology/P/P14/P14-5010

McWilliams A, Siegel D (1997) Event studies in management research: Theoretical and empirical issues. Academy of Management J 40:626–657

Miller GA (1995) WordNet: A lexical database for English. Communications of the ACM 38(11):39

Nini A (2015) Multidimensional Analysis Tagger (version 1.3). URL http://sites.google.com/site/multidimensionaltagger

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12:2825–2830, URL http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Strong N (1992) Modelling abnormal returns: A review article. J Business Finance & Accounting 19:533–553

Toutanova K, Manning CD (2000) Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13, Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '00, pp 63–70, DOI 10.3115/1117794.1117802, URL https://doi.org/10.3115/1117794.1117802

Verchow T (2011) Ad-hoc-Publizität und Kapitalmarkteffizienz: Eine untersuchung basierend auf der textanalyse von ad-hoc-mitteilungen. PhD thesis, Ulm University, Faculty of Mathematics and Economics