Combining ML and Semantic Features in the Classification of Corporate Disclosures

Stefan Evert[1], **Philipp Heinrich**[1], Klaus Henselmann[2],
Ulrich Rabenstein[3], Elisabeth Scherr[2], and Lutz Schröder[3]
[1] *Corpus Linguistics Group* — [2] *School of Business and Economics* — [3] *Dept. of Computer Science*
Friedrich-Alexander University of Erlangen-Nuremberg
LACompLing — Stockholm — August 16, 2017

Outline

Introduction

Motivation

- goal: extract hidden information from financial texts
- data basis: ad hoc disclosures and their effect on stock market performance
- methodology: extract semantic feature and feed it to a machine learner
- idea: after extracting the overt features, ML can learn hidden features

Related work

- Bollen et al. (2010) mine big data (e.g. twitter) for stock market prediction
- Ding et al. (2015) use events extracted from news to predict stock market performance
- Verchow (2011) analyses capital market efficiency using ad hoc announcements
- Feuerriegel et al. (2015) perform sentiment analysis and topic modelling of ad hoc announcements incl. stock price prediction

Motivation

- goal: extract hidden information from financial texts
- data basis: ad hoc disclosures and their effect on stock market performance
- methodology: extract semantic feature and feed it to a machine learner
- idea: after extracting the overt features, ML can learn hidden features

Related work

- Bollen et al. (2010) mine big data (e.g. twitter) for stock market prediction
- Ding et al. (2015) use events extracted from news to predict stock market performance
- Verchow (2011) analyses capital market efficiency using ad hoc announcements
- Feuerriegel et al. (2015) perform sentiment analysis and topic modelling of ad hoc announcements incl. stock price prediction

Predicting Stock Prices from Text

Ad hoc announcements

- Federal Financial Supervisory Authority (BaFin) (2009) regulates emission and lists potentially price-sensitive events

Event categories (reasons for the emission of ad hoc announcements)

- strategic corporate decisions
- corporate actions
- company- and business-specific information
- legal and political events
- human resources measures
- miscellaneous

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.

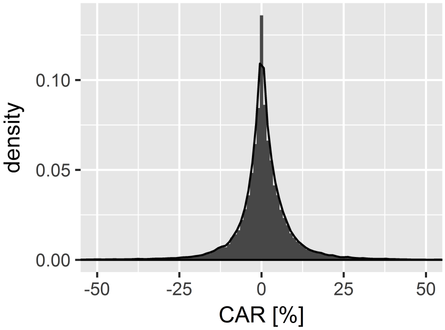
Corpus

- source: DGAP service of Equity Story AG
- sample period: mid-1996 until mid-2012
- 28,287 pre-selected texts (English, machine-readable, meta-data)

Abnormal Return

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

Cumulative Abnormal Return (three-days-window)



Prediction Tasks

- heavy-tailed data $\rightarrow$ regression is difficult
- in practice: distinguishing classes more important than predicting exact degree of reaction
- classes defined by empirical quantiles of CAR
- ternary categorization practical and feasible
- easy version: distinctive categories

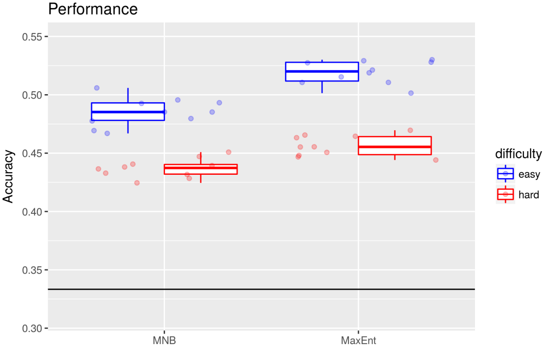
	negative	neutral	positive	corpus size
hard	9,433	9,436	9,418	28,287 (100%)
easy	5,661	5,645	5,648	16,954 (60%)

ML Classification

- pre-processing of disclosures:
  - deletion of:
    - boilerplate footers & headers
    - stop-words
    - e-mail addresses & URLs
    - punctuationmarks
  - lemmatization
  - lower-casing
- features:
  - document-term-matrix
    - hard: 28,287 documents $\times$ 32,401 lemmas
    - easy: 16,954 documents $\times$ 30,585 lemmas
  - tf.idf weighting (learned on training data, applied to test data)
- classifiers:
  - MNB
  - MaxEnt (ℓ_1 tuned on training data set)

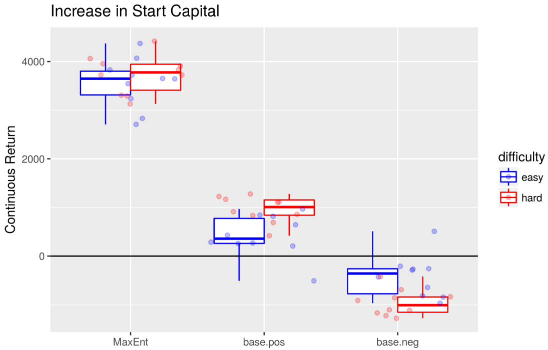
Evaluation: Accuracy in 10-fold Stratified Crossvalidation

Accuracy	MNB	MaxEnt	baseline
hard	43.7% (±1.6%)	**45.6%** (±1.8%)	33.3%
easy	48.5% (±2.4%)	**51.9%** (±1.9%)	33.3%



Evaluation: Trading Strategy

ML predicts *positive*	$\rightarrow$	*buy*
ML predicts *negative*	$\rightarrow$	*sell short*
ML predicts *neutral*	$\rightarrow$	*hold*



Ontological Feature Extraction

Event categories (reasons for the emission of ad hoc announcements)

1. strategic corporate decisions
2. corporate actions
3. company- and business-specific information
4. legal and political events
5. human resources measures
   - 5.1 change in personnel
     - 5.1.1 suggestions for appointments
     - 5.1.2 extensions to the supervisory and management boads
     - 5.1.3 extension of contracts
     - 5.1.4 key personnel turnover (ca. 5%)
   - 5.2 announcements of forced redundancies
6. miscellaneous

Background

- disclosures are sent out for very specific reasons
  - Federal Financial Supervisory Authority (BaFin) (2009)
- event categories are somewhat fuzzy, but mostly straightforward
  - about 15% of announcements cannot be categorized unambiguously
- event categories are not mentioned in the text
  - development of a **formal ontology for detecting emission reason**

Ontology for detecting *retirement disclosures*

1. automatically generated TBox
   - captures relations among concepts (using WordNet)
2. automatically generated ABox
   - records the content of parsed disclosures
3. manually maintained TBox
   - captures domain-specific background knowledge

NLP pre-processing

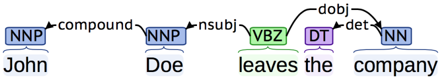
- pre-processing via Stanford CoreNLP:
  - PoS-tagging
  - morphological analysis (lemmatization)
  - syntactic parsing
  - NER
  - coreference resolution
- WSD via Lesk algorithm and WordNet (Banerjee and Pedersen, 2002)

automatically generated TBox

- transformation of NLP results into Web Ontology Language (OWL)
- use of lexical semantic information from WordNet
  - mapping of surface realizations to respective synsets

automatically generated ABox

- mappings:
  - subjects and objects → individuals
  - connecting verbs → object properties
  - prepositional verbs: preposition → part of object property
- additionals:
  - detection of announcements (*will retire*)
  - resolution of compound nouns
  - inference of types: NER, appositions, morphological analysis, ...



Individual: John_Doe
    Types: Person
    Facts: 2383440_leave_depart_pull_up_stakes company

Individual: company
    Types: 8058098_Company

manually maintained TBox (background knowledge)

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

Types of background knowledge

- general and domain-specific knowledge
- examples:
  - "stepping down" = "leaving"
  - "letting contract expire" = "leaving current position"
  - "Executive Board" = "Management Board"
  - CFOs work on exactly one executive position

Querying

```
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. }}}
```

Types of background knowledge

- general and domain-specific knowledge
- examples:
  - “stepping down” = “leaving”
  - “letting contract expire” = “leaving current position”
  - “Executive Board” = “Management Board”
  - CFOs work on exactly one executive position

Querying

```
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. }}}
```

Evaluation

		ontological classes		
		ret.	*non-ret.*	*total*
manual classes	*ret.*	161 (TP)	17 (FN)	178
	non-ret.	5 (FP)	117 (TN)	122
	total	166	134	300

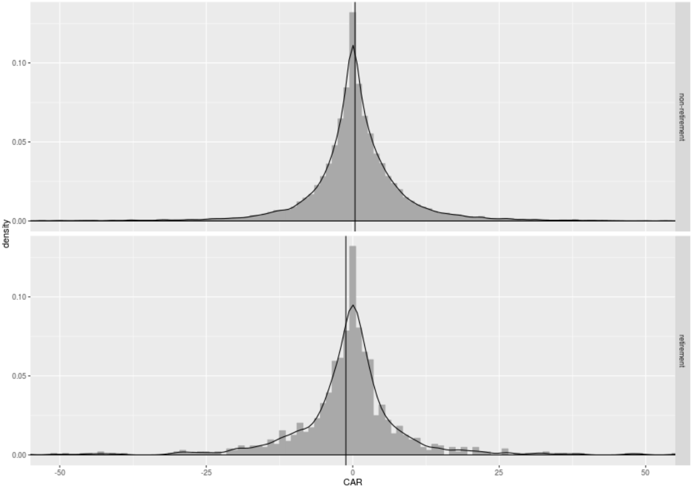
- manual evaluation:
  - 300 disclosures containing *leave* or *retire*
  - baseline accuracy: 59.3% (178 / 300)
  - recall: **90.4%** (161 / 178) — precision: **97%** (161 / 166)
- subsequent run on whole corpus
  - 1,046 / 28,287 (ca. 3.7%)
  - 639 / 16,954 (ca. 3.8%)

Evaluation

		ontological classes		
		ret.	*non-ret.*	*total*
manual classes	*ret.*	161 (TP)	17 (FN)	178
	non-ret.	5 (FP)	117 (TN)	122
	total	166	134	300

- manual evaluation:
  - 300 disclosures containing *leave* or *retire*
  - baseline accuracy: 59.3% (178 / 300)
  - recall: **90.4%** (161 / 178) — precision: **97%** (161 / 166)
- subsequent run on whole corpus
  - 1,046 / 28,287 (ca. 3.7%)
  - 639 / 16,954 (ca. 3.8%)

Association of Semantic Feature and Target Variable



Integrating the Semantic Feature into the DTM

idea

- equip ML with ontologically extracted retirement feature
- split the overall problem into smaller sub-problems
- ML focussing on retirement disclosures has an easier task

methodology

1. add a single "retirement" feature to feature matrix
2. separate vocabularies of retirement disclosures and non-retirements
3. mirror vocabulary of retirement disclosures

evaluation

- comparison straightforward (accuracy in 10-fold stratified cross-validation)
- additionally: *ontological* classification (retirement disclosures are predominantly neagative) $\rightarrow$ ontological baseline outperforms initial baseline by predicting category *negative* for all retirement disclosures

idea

- equip ML with ontologically extracted retirement feature
- split the overall problem into smaller sub-problems
- ML focussing on retirement disclosures has an easier task

methodology

1. add a single “retirement” feature to feature matrix
2. separate vocabularies of retirement disclosures and non-retirements
3. mirror vocabulary of retirement disclosures

evaluation

- comparison straightforward (accuracy in 10-fold stratified cross-validation)
- additionally: *ontological* classification (retirement disclosures are predominantly neagative) $\rightarrow$ ontological baseline outperforms initial baseline by predicting category *negative* for all retirement disclosures

idea

- equip ML with ontologically extracted retirement feature
- split the overall problem into smaller sub-problems
- ML focussing on retirement disclosures has an easier task

methodology

1. add a single “retirement” feature to feature matrix
2. separate vocabularies of retirement disclosures and non-retirements
3. mirror vocabulary of retirement disclosures

evaluation

- comparison straightforward (accuracy in 10-fold stratified cross-validation)
- additionally: *ontological* classification (retirement disclosures are predominantly neagative) $\rightarrow$ ontological baseline outperforms initial baseline by predicting category *negative* for all retirement disclosures

Feature Matrices

	description	n_{voc} (hard)
FM1	vanilla feature matrix without retirement feature	32,401
FM2	FM1 with a single *additional retirement feature*	32,402
FM3	FM1 with a *separate vocabulary* for the ret. disclosures	37,652
FM4	FM1 with a *mirrored vocabulary* for the ret. disclosures	38,762

Prediction Tasks

	negative	neutral	positive	corpus size
hard	9,433	9,436	9,418	28,287 (100%)
retirements	*413*	*341*	*292*	*1046 (3.7%)*
easy	5,661	5,645	5,648	16,954 (60%)
retirements	*267*	*205*	*167*	*639 (3.8%)*

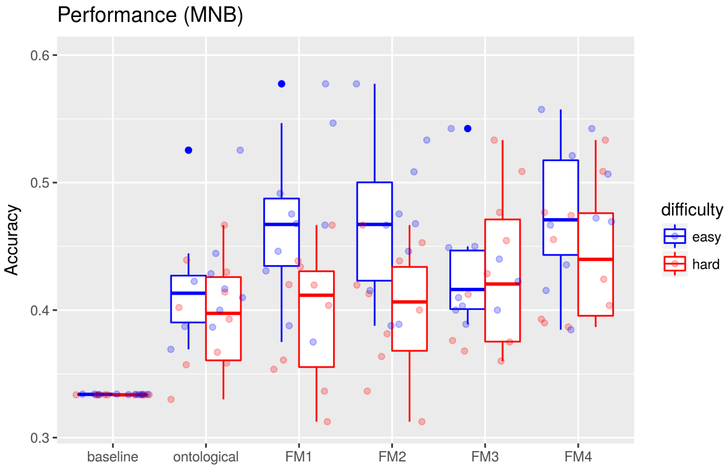
Results and Discussion

Prediction results

hard	*full*		*retirements*	
	MNB	MaxEnt	MNB	MaxEnt
FM1	.437 (±.016)	.456 (±.018)	.395 (±.101)	.426 (±.092)
FM2	.437 (±.016)	.456 (±.016)	.395 (±.101)	.426 (±.092)
FM3	.437 (±.015)	.456 (±.019)	.429 (±.124)	.414 (±.091)
FM4	.439 (±.014)	**.459** (±.025)	.445 (±.106)	**.450** (±.128)
baseline	$^1/_3 = .333$		.396 (±.086)	

easy	*full*		*retirements*	
	MNB	MaxEnt	MNB	MaxEnt
FM1	.485 (±.024)	**.519** (±.019)	.467 (±.126)	.479 (±.115)
FM2	.485 (±.024)	.518 (±.014)	.467 (±.123)	.481 (±.117)
FM3	.482 (±.022)	**.519** (±.021)	.431 (±.090)	.470 (±.098)
FM4	.486 (±.022)	**.519** (±.018)	.477 (±.111)	**.500** (±.092)
baseline	$^1/_3 = .333$		.419 (±.087)	

Performance Comparison



Feature Weight Analysis (MaxEnt, hard, category *positive*)

lemma	FM1	FM3 non-ret.	FM3 ret.	FM4 non-ret.	FM4 ret.
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

Conclusion

- combination of semantics-based approach (ontology) with ML classification on bag-of-lemmas (formal features)
- MLs benefit from ontological information
  - more specific realm of language use
  - hypothesis: words are used more consistently with the specific domain of retirement disclosures
- effect is consistent, yet not statistically significant

Future Work

- refine ontological approach
- broaden ontological categories
- how to exploit subjective use of language in different domains?

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

Thanks for listening.

Any questions?

References

Satanjeev Banerjee and Ted Pedersen. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 136–145, London, UK, 2002. Springer-Verlag. ISBN 3-540-43219-1. URL `http://dl.acm.org/citation.cfm?id=647344.724142`.

Johann Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, October 2010. URL `http://arxiv.org/pdf/1010.3003v1.pdf`.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (ICJAI)*, pages 2327–2333, 2015. URL `http://ijcai.org/papers15/Papers/IJCAI15-329.pdf`.

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

Stefan Feuerriegel, Antal Ratku, and Dirk Neumann. 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)*, 2015. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2564603`.

Thomas Verchow. *Ad-hoc-Publizität und Kapitalmarkteffizienz: Eine Untersuchung basierend auf der Textanalyse von Ad-hoc-Mitteilungen*. PhD thesis, Ulm University, Faculty of Mathematics and Economics, 2011.