A Corpus of Corporate Annual and Social Responsibility Reports: 280 Million Tokens of Balanced Organizational Writing

Sebastian G. M. Händschke[2] * **Sven Buechel**[1] * **Jan Goldenstein**[2]
Philipp Poschmann[2] **Tinghui Duan**[1] **Peter Walgenbach**[2] **Udo Hahn**[1]

[1] Jena University Language and Information Engineering (JULIE) Lab
http://www.julielab.de

[2] School of Economics and Business Administration
http://www.orga.uni-jena.de

Friedrich-Schiller-Universität Jena, Jena, Germany

Abstract

We introduce JOCo, a novel text corpus for NLP analytics in the field of economics, business and management. This corpus is composed of corporate annual and social responsibility reports of the top 30 US, UK and German companies in the major (DJIA, FTSE 100, DAX), middle-sized (S&P 500, FTSE 250, MDAX) and technology (NASDAQ, FTSE AIM 100, TECDAX) stock indices, respectively. Altogether, this adds up to 5,000 reports from 270 companies headquartered in three of the world's most important economies. The corpus spans a time frame from 2000 up to 2015 and contains, in total, 282M tokens. We also feature JOCo in a small-scale experiment to demonstrate its potential for NLP-fueled studies in economics, business and management research.

1 Introduction

A crucial prerequisite in today's NLP research is the availability of large amounts of language data. National reference corpora such as the ANC for American English (Ide and Suderman, 2004), the BNC for British English (Burnard, 2000), and the DeReKo for German (Kupietz and Lüngen, 2014) assemble a collection of language data with a focus on ordinary language use covering a wide range of genres (e.g., newspaper articles, technical writing and popular fiction, letters, transcripts of court or parliament speeches, etc.). Corpora exclusively focusing on newspaper articles have been particularly influential for the development of syntactic and semantic methodologies in NLP research (e.g., Penn TreeBank (Marcus et al., 1993) or Penn PropBank (Palmer et al., 2005) for the English language).

Turning to more specialized, mostly scientific, domains these general language resources can only be reused at the cost of substantial performance penalties due to characteristic sublanguage phenomena in those domains. For the biomedical domain, e.g., these negative effects can be shown for the whole range of low-level (sentence splitting, tokenization (Tomanek et al., 2007; Griffis et al., 2016)) up to high-level tasks (such as syntactic analysis (Laippala et al., 2014; Jiang et al., 2015)). As a consequence, these specialized fields of NLP research have created their own resource infrastructure in terms of domain-specific lexicons and corpora for syntactic and semantic processing.

The rapidly increasing number of publications using text analytics for economics, business, and management (for surveys, cf. Lu et al. (2010); Goldenstein et al. (2015); Kumar and Ravi (2016)) indicates the emergence of an entirely new application domain for NLP systems (see Section 2). At first sight, one might argue that domain-specific corpora such as the Penn TreeBank are sufficient since they already contain economy-related language data. Yet, as these resources assemble only excerpts from newspaper articles, at second sight, such resources turn out to be biased. Newspaper articles reflect *journalists'* interpretations and do not necessarily directly transport the attitudes and views of *economic actors*, such as an individual (consumer) or business corporations (Simon, 1991).

This shortcoming can be alleviated if one targets the economic actors' verbal communication behavior directly on various media channels. Our choice is to focus on annual reports (AR) and corporate social responsibility reports (CSRR) of major business corporations in Western economies.

* These authors contributed equally to this work.

Proceedings of the First Workshop on Economics and Natural Language Processing, pages 20–31
Melbourne, Australia, July 20, 2018.

Altogether these documents comprise 282M tokens and reflect the unfiltered views of these commercial enterprises *and* their embedding in the social and regulatory system in market-driven societies. Viewing enterprises as social actors with their own goals, their legal, social and other responsibilities becomes increasingly relevant for both the explanation and prediction of economic and organizational phenomena, as well as for economics, management and organization science, in general (King et al., 2010; Bromley and Sharkey, 2017). While the raw data set we assembled can be used for scientific purposes only, we also offer an embedding model trained on it which is available without any legal restrictions.[1]

2 Related Work

The ties between NLP, economics, management, and organization science have evolved around different types of economic actors and roles they play in an economic setting. One stream of work deals with NLP-based *customer* analytics by profiling customers, tracking their product/company preferences, screening customer reviews, etc. (Archak et al., 2011; Ikeda et al., 2013; Zhang and Pennacchiotti, 2013; Stavrianou and Brun, 2015; Yang et al., 2015; Sakaki et al., 2016; Pekar and Binner, 2017). Another stream is concerned with NLP-based *product* analytics, e.g., based on (social) media monitoring, summarizing reviews, or identifying (deceptive/fake) product descriptions or reviews (Mukherjee et al., 2012; Feng et al., 2012; Wang and Ester, 2014; Tsunoda et al., 2015; Fang and Zhan, 2015; Kessler et al., 2015; Imada et al., 2016; Chen et al., 2016; Pryzant et al., 2017).

Yet, the main thrust of work is devoted to NLP-based financial *(stock) market* analytics, e.g., analyzing companies' market performance indicators (trend prediction, performance forecasting, volatility prediction, etc.) and verbal statements related to market performance, competitors or future perspectives (Schumaker and Chen, 2009; Kogan et al., 2009; Nassirtoussi et al., 2014; Li et al., 2014; Qiu and Srinivasan, 2014; Kazemian et al., 2014; de Fortuny et al., 2014; Ammann et al., 2014; Wang and Hua, 2014; Nguyen and Shirai, 2015; Luss and d'Aspremont, 2015; Ding et al., 2015; Liu et al., 2015; Feuerriegel and Prendinger, 2016; Rekabsaz et al., 2017; Xing et al., 2018; Li et al., 2018).

This external market view is complemented by NLP-based *organization/enterprise* analytics, e.g., social role taking, risk prediction, fraud analysis, market share analytics, etc. (Goel et al., 2010; Hájek and Olej, 2015; Buechel et al., 2016; Goel and Uzuner, 2016; El-Haj et al., 2016; Tsai and Wang, 2017), including *competitive* or *business intelligence* services based on NLP tooling (Chaudhuri et al., 2011; Chung, 2014).

From a methodological perspective, the social interactions between these actors—customers, enterprises, and political/juridical authorities—have been studied in terms of *sentiments* they bring to bear (Van De Kauter et al., 2015). Evidence is collected from consumers' and enterprises' verbal behavior and their communication about products and services, e.g., via social media (Chen et al., 2014; Si et al., 2014; Liu, 2015; Alshahrani et al., 2018). This research is complemented by studies related to *reputation, expertise, credibility and trust* models for agents in the economic process (as traders, sellers, advertisers) based on mining communication traces and recommendation legacy data, including fake ad/review recognition (Bar-Haim et al., 2011; Brown, 2012; Mukherjee et al., 2012; Rechenthin et al., 2013; Tang and Chen, 2014; Žnidaršič et al., 2018).

System-wise, specialized types of search engines have been developed, for instance, *enterprise search engines* (e-commerce, e-marketing) or *consumer search engines*, market monitors, product/service recommender systems (Vandic et al., 2017; Trotman et al., 2017). This also includes *customer-supplier interaction platforms* (e.g., portals, helps desks, newsgroups) and transaction support systems based on natural language communication (including business chat bots) (Cui et al., 2017; Altinok, 2018). Specialized modes of *information extraction* and text mining in economic domains, e.g., temporal event or transaction mining have also been explored (Tao et al., 2015; Lefever and Hoste, 2016; Ding et al., 2016), as well as *information aggregation* from single sources (e.g., review summaries, automatic threading) (Gerani et al., 2014).

The language resources behind these activities include specialized *lexicons* (Loughran and McDonald, 2011) and *ontologies* for economics (Leibniz Information Centre for Economics, 2014), the adaptation or acquisition of lexicons for economic NLP (Xie et al., 2013; Moore et al.,

[1] www.orga.uni-jena.de/orga/en/Corpus.html

2016; Oliveira et al., 2016; Chen et al., 2018), *corpora* and annotations policies (guidelines, metadata schemata, etc.) for economic NLP concerned with domain-specific text genres (business reports, auditing documents, product reviews, economic newswire, social media posts or blogs, business letters, legislation documents, etc.) (Flickinger et al., 2012; Takala et al., 2014; Kessler and Kuhn, 2014; Asooja et al., 2015; Schön et al., 2018) , and dedicated *tools* for economic NLP (e.g., NER taggers, sublanguage parsers, pipelines for processing economic discourse) (Schumaker and Chen, 2009; Feldman et al., 2011; Hogenboom et al., 2013; Kessler and Kuhn, 2013; Lee et al., 2014; Malo et al., 2014; Weichselbraun et al., 2015; Lefever and Hoste, 2016; Ding et al., 2016; El-Haj et al., 2018).

Pioneering efforts in considering texts originally produced by enterprises as a basis for economic NLP were made by Kloptchenko et al. (2004) who used sentiments in enterprises' quarterly reports as a predictor for stock market prices. Later Kogan et al. (2009) came up with the influential *10-K Corpus*, a collection of 54,379 ARs from 10,492 different, publically traded companies covering a time interval from 1996 up to 2006. This seminal resource is a cornerstone of economic corpus development and our work is meant to complement it with current and more diverse language data.

3 Corpus Description

The corpus we here introduce consists of ARs and CSRRs from companies in the United States, the United Kingdom and Germany. An *AR* is a comprehensive report published yearly by publicly-listed corporations on their activities and financial performance of the past year. ARs provide information for current and prospective shareholders, the governmental and regulatory bodies, the stock exchanges, as well as all other stakeholders (Neu et al., 1998; Yuthas et al., 2002). A *CSRR* is a regular report published by a company or an organization about the economic, environmental and social impacts caused by its activities (Dahlsrud, 2008; Chen and Bouvain, 2009; Fifka, 2013). CSRRs also present the organization's values and governance model, and reveal the link between its strategy and its commitment to the organization's environment and a sustainable global economy (Du et al., 2010; Aguinis and Glavas, 2012).

With regard to the popular 10-K corpus (Kogan et al., 2009), the data set we present is significantly smaller in size (both in terms of tokens and companies). However, the 10-K corpus only covers ARs, while we also include CSRRs allowing a wider view on organizational communication traces. Also, the 10-K corpus only includes reports up to the year 2006, whereas our work incorporates documents as recent as 2015. Additionally, the 10-K corpus is only based on the 10-k forms mandated by the Securities Exchange Commission (SEC) in the US. Nonetheless, US corporations' ARs contain the same information as required by the 10-k forms and much more. Furthermore, ARs are a genre of reports diffused globally (Rutherford, 2005; Meyer and Höllerer, 2010). Hence, the choice of ARs as a backbone for our corpus allows for a careful international sampling strategy balancing different kinds of corporations from different countries. This property makes our corpus particularly well suited for deeper economic investigations with respect to cross-index, cross-industry and cross-country comparisons.

3.1 Selection of Raw Data

ARs as well as CSRRs are considered relevant for our corpus based on two main criteria, namely the company that issued them and the year they report about. We selected companies in a step-wise process, first selecting the countries of origin and then the stock indices they were listed in.

Regarding the selection of countries, we chose the US, the UK and Germany, because altogether their total GDP makes up for 30% of the WGDP (as of 2014), thus representing a relevant portion of the global economy. For each of these three countries, 90 companies where selected for inclusion in our corpus. We first took the 30 most intensively traded and most highly valued corporations of the American Dow Jones Industrial Average (DIJA), the British Financial Times Stock Exchange (FTSE 100) and the German Stock Index (DAX; "Deutscher Aktienindex"). Next, we added reports of middle-sized companies (30 per country) and technology companies (again 30 per country) for a total of 270 companies in our sample. Middle-sized companies were selected from the S&P500, the FTSE 250 and the MDAX, whereas tech firms were chosen from the NASDAQ, the FTSE AIM 100 and the TECDAX indices for the US, the UK and Germany, respectively. We se-

Index	Annual Reports			Corporate Social Responsibility Reps			Total		
	Tokens	Sentences	Reps	Tokens	Sentences	Reps	Tokens	Sentences	Reps
DIJA	27,139,371	864,724	458	7,168,558	253,564	239	34,307,929	1,118,288	697
S&P500	23,914,717	780,372	335	2,902,234	101,707	113	26,816,951	882,079	448
NASDAQ	24,937,589	737,156	342	896,070	32,769	58	25,833,659	769,925	400
FTSE 100	47,086,382	1,458,637	452	8,913,870	322,565	278	56,000,252	1,781,202	730
FTSE 250	20,654,093	619,239	472	1,657,327	56,052	86	22,311,420	675,291	558
FTSE AIM 100	15,878,972	477,245	426	207,220	7,746	30	16,086,192	484,991	456
DAX	45,170,200	1,535,016	469	9,646,971	362,162	254	54,817,171	1,897,178	723
MDAX	23,198,101	786,189	366	3,193,350	116,437	93	26,391,451	902,626	459
TechDAX	19,083,290	654,875	350	203,393	8,076	15	19,286,683	662,951	365
Total	247,062,715	7,913,453	3,670	34,788,993	1,261,078	1,166	281,851,708	9,174,531	4,836

Table 1: Numbers of tokens, sentences and reports relative to stock index and report category.

economy		*growth*		*tax*		*leadership*		*sustainable*	
recession	.70	grow	.66	taxes	.73	leaders	.66	sustainably	.64
economies	.69	double-digit	.64	taxation	.71	excellence	.57	sustainability	.64
upswing	.68	strong	.63	deferred	.65	reinforce	.56	environmentally	.56
upturn	.67	organic	.60	non-deductible	.61	leader	.55	stewardship	.56
gdp	.66	profitable	.60	carryforwards	.57	competencies	.55	low-carbon	.54

Table 2: Sample word embeddings illustrated by their five nearest neighbors based on cosine similarity.

lected each corporation from the three countries so that they matched the corresponding two counterparts with respect to industry segment, sales and trading volumes.

Lastly, we let the time span of our corpus range between the years 2000 and 2015. Each report (AR and CSRR) from one of the 270 companies in the previously defined sample that addresses one of these years was included in the corpus, if possible (see also the following Subsection 3.2). The year 2000 was chosen as a starting point because of, first, the burst of the dotcom-bubble and, second, the upcoming of CSRRs. Further details regarding our sampling strategy are provided in the README file of our corpus distribution.

3.2 Data Acquisition and Cleansing

The reports determined in this way were collected by three student assistants from the Business and Management Department by downloading the reports in PDF format from the companies' websites. In some cases, especially for documents from the early 2000s, reports were not available for downloading. The students (and, if necessary, one of the authors) then requested the documents directly from the respective investor relations department via email. The following metadata were recorded: *report type* (either AR or CSRR), *reference year* of the report[2] (as given on the title page), *company* of origin, and *stock index*.

[2] In some cases, and in particular with regard to CSRR, sometimes multiple consecutive years were indicated. In these cases, only the first year is considered as reference year.

We used the pdf2text software by `glyphandcog.com` to extract plain text from the collected PDF files. In general, this software extracts text with sufficient quality. However, the final result depends heavily on the layout and style of the input files. For this reason, the resulting plain text files were iteratively refined in a rule-based fashion. This post-processing included restoring of the original text structure of headings and paragraphs, deleting superfluous line breaks and hyphenation, page numbers and (rarely occurring) odd character sequences, as well as remnants of structured data, such as tables. This post-processing strategy yielded a mostly clean corpus of raw textual data only, i.e., preserving the running text of the original PDF files as good as possible while at the same time stripping off all irrelevant non-linguistic data.

3.3 Corpus Analysis

After corpus construction, we used `NLTK.org` tools (Bird, 2006) for counting tokens and sentences for all of the reports. The results, summarized for each stock index, are depicted in Table 1. In total, our corpus comprises almost 5,000 reports, summing up to 282M tokens (9M sentences). This constitutes a substantial collection of textual data (for comparison, the BNC, ANC, and DEREKO contain 100M, 15M, and 42B tokens, respectively). The vast majority of the data set consists of ARs (247M tokens vs. 35M tokens from CSRRs). American, British and German corporations are properly represented in the data set,

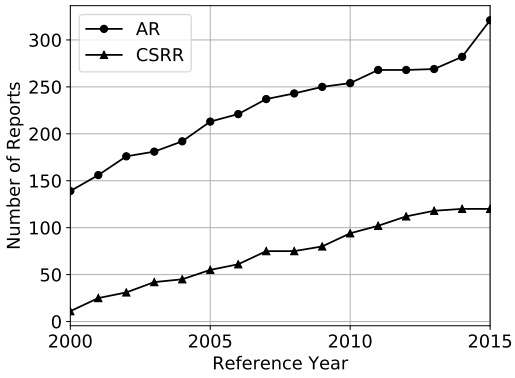

Figure 1: Distribution of reports over time.

i.e., for each of these countries, their three indices add up to about 90M tokens.

Figure 1 depicts the growth curves for ARs as well as CSRRs. As can be seen, for both ARs and CSRRs, the number of reports increases over time. This graph also reflects the fact that documents become harder to acquire the older they are, as we have experienced during data collection. Note that we could only collect a marginal number of CSRRs for the year 2000 (11). This is due to the fact, that their issuance became wide-spread only in this and the following years, as discussed above.

3.4 Word Embeddings

The distribution of the plain text data of JOCo is restricted by Intellectual Property Rights (IPR) regulations. As a substitute, we train word embeddings using the `FastText.cc` toolkit (Bojanowski et al., 2017) to capture the distributional semantics of economic jargon. As a prerequisite, the corpus was tokenized using NLTK and case-folded. Only words with frequency $\geq$ 50 were modeled. Subword information was *not* taken into account. The latter two decision were taken to decrease the number of artifacts stemming from the PDF conversion in our final embedding model.

To illustrate the semantics captured in this way, Table 2 lists sample entries of our embedding model together with their five nearest neighbors. As can be seen, the results reveal high face validity: *"growth"*, e.g., exhibits strong reference to its economic meaning (such as in *"double-digit growth"* or *"organic growth"*) but does not refer to biological growth which may have been indicated by neighbors like *"plant"* or *"hormones"*.

4 Effects of Organizational Emotions

To demonstrate the potential of the JOCo corpus, we investigate the interaction of linguistic signals from corporations and their market performance. We focus on emotions expressed in ARs since the interplay of organizational cognition, character, and emotions is becoming a hot topic in organization science (Albrow, 1992; King, 2015; Buechel et al., 2016; Händschke et al., 2017). We conducted this work on a subsample of the corpus covering British and German firms only and their ARs from 2008 to 2015 to allow for European comparability. Financial and accounting metadata were retrieved from AMADEUS,[3] a database that holds data of European firms (except for banks and insurance companies).

In the regression analysis, we employ the generalized estimating equations (GEE) method (Liang and Zeger, 1986), a time series model that handles repeating observations over time. In our case we use its multivariate linear regression variant (see the Appendix for details). The dependent variable 'performance' is operationalized as *Return on Equity (ROE)*, lagged by one year to allow for causality. Following the established psychological VAD model of emotions (Bradley and Lang, 1994), the independent explanatory variables are three dimensions of espoused organizational emotions—*Valence*, *Arousal*, and *Dominance*. These three dimensions are measured individually for each AR using the open-source tool JEmAS[4] (Buechel and Hahn, 2016) that yields a value for each of the dimensions per firm per year. Due to the high correlation between dominance and valence, the latter variable was dropped from the model to prevent biasing of the estimators (cf. the correlation matrix given in the Appendix, Table 3). Control variables are the corporation's size (in terms of employees and assets, both logarithmized),[5] operational profitability (sales per employee and sales per assets) and country of origin measured with a dummy variable where Germany is coded as '1'.

For our full model (Model III in Table 4), we find that Arousal has a significant ($p < .001$) negative effect on ROE, meaning that a company performs better, the calmer it communicates. However, this effect is more pronounced for British companies since the interaction term be-

[3] `https://amadeus.bvdinfo.com/`
[4] `https://github.com/JULIELab/JEmAS`
[5] All other metric variables have been standardized.

tween Arousal and country (GER) shows a significant ($p < .001$) positive effect. Thus, our results suggest that espoused organizational emotionality correlates with performance, yet the nature of this interaction is country-dependent. Accordingly, our findings point towards the existence of a distinct organizational character (King, 2015) and emotionality (Albrow, 1992), and thus render support viewing organizations as social actors (King et al., 2010; Bromley and Sharkey, 2017). This piece of evidence might have far-reaching implications for the organizations' role and responsibility in society (Beyer et al., 2014).

5 Conclusion

We introduced JOCo, a novel text corpus for NLP analytics in the field of economics, business and management. This corpus comprises ARs and CSRRs of 270 publicly traded corporations in the US, UK and Germany from 2000 to 2015. Altogether, we assembled roughly up to 5,000 reports and, in total, 282M tokens (9M sentences). By design, JOCo carefully balances various characteristics allowing cross-index, cross-industry, and cross-country comparisons and, thus, enables informed prospective applications in business research and economics, for which we provided a first, yet preliminary example.

Acknowledgments

We thank the anonymous reviewers for their helpful comments, as well as our research assistants, esp. Nadine Halli, for their effort in assembling the primary documents.

References

Herman Aguinis and Ante Glavas. 2012. What we know and don't know about corporate social responsibility: A review and research agenda. *Journal of Management*, 38(4):932–968.

Martin Albrow. 1992. Sine ira et studio—or do organizations have feelings? *Organization Studies*, 13(3):313–329.

Mohammed Alshahrani, Fuxi Zhu, Mohammed Alghaili, Eshrag Refaee, and Mervat Bamiah. 2018. BORSAH: An Arabic sentiment financial tweets corpus. In *FNP 2018 — Proceedings of the 1st Financial Narrative Processing Workshop @ LREC 2018. Miyazaki, Japan, 7 May 2018*, pages 17–22.

Duygu Altinok. 2018. An ontology-based dialogue management system for banking and finance dialogue systems. In *FNP 2018 — Proceedings of the 1st Financial Narrative Processing Workshop @ LREC 2018. Miyazaki, Japan, 7 May 2018*, pages 1–9.

Manuel Ammann, Roman Frey, and Michael Verhofen. 2014. Do newspaper articles predict aggregate stock returns? *Journal of Behavioral Finance*, 15(3):195–213.

Nikolay Archak, Anindya Ghose, and Panagiotis G. Ipeirotis. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8):1485–1509.

Kartik Asooja, Georgeta Bordea, Gabriela Vulcu, Leona O'Brien, Angelina Espinoza, Elie Abi-Lahoud, Paul Buitelaar, and Tom Butler. 2015. Semantic annotation of finance regulatory text using multilabel classification. In *LeDA-SWAn 2015 — Proceedings of the 2015 International Workshop on Legal Domain and Semantic Web Applications @ ESWC 2015. Portorož, Slovenia, June 1, 2015*.

Roy Bar-Haim, Elad Dinur, Ronen Feldman, Moshe Fresko, and Guy Goldstein. 2011. Identifying and following expert investors in stock microblogs. In *EMNLP 2011 — Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, U.K., 27-31 July 2011*, pages 1310–1319.

Susan Beyer, Stephan Bohn, Toni Grünheid, Sebastian Händschke, Raluca Kerekes, Jonas Müller, and Peter Walgenbach. 2014. Wofür übernehmen Unternehmungen Verantwortung? Und wie kommunizieren sie ihre Verantwortungsübernahme? *Zeitschrift für Wirtschafts- und Unternehmensethik*, 15(1):57–80.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *COLING-ACL 2006 — Proceedings of the 21st International Conference on Computational Linguistics & 44th Annual Meeting of the Association for Computational Linguistics: Interactive Presentation Sessions. Sydney, New South Wales, Australia, 17-18 July 2006*, pages 69–72.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5(1):135–146.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Patricia Bromley and Amanda Sharkey. 2017. Casting call: The expanding nature of actorhood in US firms, 1960–2010. *Accounting, Organizations and Society*, 59:3–20.

Eric D. Brown. 2012. Will Twitter make you a better investor? A look at sentiment, user reputation and their effect on the stock market. In *SAIS 2012 —*

Proceedings of the Southern Association for Information Systems Conference. Atlanta, Georgia, USA, March 23-24, 2012, pages 36–42.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem: Dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 — Proceedings of the 22nd European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS 2016). The Hague, The Netherlands, August 29 - September 2, 2016*, number 285 in Frontiers in Artificial Intelligence and Applications, pages 1114–1122, Amsterdam, Berlin, Washington, D.C. IOS Press.

Sven Buechel, Udo Hahn, Jan Goldenstein, Sebastian G. M. Händschke, and Peter Walgenbach. 2016. Do enterprises have emotions? In *WASSA 2016 — Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ NAACL-HLT 2016. San Diego, California, USA, June 16, 2016*, pages 147–153.

Lou Burnard. 2000. User Reference Guide for the British National Corpus. Technical report, British National Corpus Consortium, Humanities Computing Unit, Oxford University Computing Services, Oxford University, Oxford, U.K.

Surajit Chaudhuri, Umeshwar Dayal, and Vivek Narasayya. 2011. An overview of business intelligence technology. *Communications of the ACM*, 54(8):88–98.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. NTUSD-FIN: A market sentiment dictionary for financial social media data applications. In *FNP 2018 — Proceedings of the 1st Financial Narrative Processing Workshop @ LREC 2018. Miyazaki, Japan, 7 May 2018*, pages 37–43.

Hailiang Chen, Prabuddha De, Yu (Jeffrey) Hu, and Byoung-Hyoun Hwang. 2014. Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5):1367–1403.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *EMNLP 2016 — Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA, November 1-5, 2016*, pages 1650–1659.

Stephen Chen and Petra Bouvain. 2009. Is corporate responsibility converging? A comparison of corporate responsibility reporting in the USA, UK, Australia, and Germany. *Journal of Business Ethics*, 87(1):299–317.

Wingyan Chung. 2014. BIZPRO: Extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management*, 34(2):272–284.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SUPERAGENT: A customer service chatbot for e-commerce websites. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Vancouver, British Columbia, Canada, August 1, 2017*, pages 97–102.

Alexander Dahlsrud. 2008. How corporate social responsibility is defined: An analysis of 37 definitions. *Corporate Social Responsibility and Environmental Management*, 15(1):1–13.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *IJCAI '15 — Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, July 25-31, 2015*, pages 2327–2333.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2016. Knowledge-driven event embedding for stock prediction. In *COLING 2016 — Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, December 11-16, 2016*, pages 2133–2142.

Shuili Du, Chitrabhan B Bhattacharya, and Sankar Sen. 2010. Maximizing business returns to corporate social responsibility (CSR): The role of CSR communication. *International Journal of Management Reviews*, 12(1):8–19.

The Leibniz Information Centre for Economics. 2014. STW thesaurus for economics. Technical report.

Mahmoud El-Haj, Paul Rayson, Paulo Alves, and Steven Young. 2018. Towards a multilingual financial narrative processing system. In *FNP 2018 — Proceedings of the 1st Financial Narrative Processing Workshop @ LREC 2018. Miyazaki, Japan, 7 May 2018*, pages 52–58.

Mahmoud El-Haj, Paul Rayson, Steven Young, Andrew Moore, Martin Walker, Thomas Schleicher, and Vasiliki Athanasakou. 2016. Learning tone and attribution for financial text mining. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 1820–1825.

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data*, 2:#5.

Ronen Feldman, Benjamin Rosenfeld, Roy Bar-Haim, and Moshe Fresko. 2011. The STOCK SONAR: Sentiment analysis of stocks based on a hybrid approach. In *AAAI-IAAI-EAAI '11 — Proceedings of the 25th AAAI Conference on Artificial Intelligence & 23rd Conference on Innovative Applications of Artificial Intelligence & 2nd Symposium on Educational Advances in Artificial Intelligence. San Francisco, California, USA, August 7-11, 2011*, pages 1642–1647.

Song Feng, Longfei Xing, Anupam Gogar, and Yejin Choi. 2012. Distributional footprints of deceptive product reviews. In *ICWSM 2012 — Proceedings of the 6th International AAAI Conference on Weblogs and Social Media. Dublin, Ireland, June 4-7, 2012*, pages 98–105.

Stefan Feuerriegel and Helmut Prendinger. 2016. News-based trading strategies. *Decision Support Systems*, 90:65–74.

Matthias S. Fifka. 2013. Corporate responsibility reporting and its determinants in comparative perspective. A review of the empirical literature and a meta-analysis. *Business Strategy and the Environment*, 22(1):1–35.

Daniel P. Flickinger, Yi Zhang, and Valia Kordoni. 2012. DEEPBANK: A dynamically annotated treebank of the Wall Street Journal. In *TLT '11 — Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories. Lisbon, Portugal, 30 November - 1 December 2012*, pages 85–96.

Enric Junqué de Fortuny, Tom De Smedt, David Martens, and Walter Daelemans. 2014. Evaluating and understanding text-based stock price prediction models. *Information Processing & Management*, 50(2):426–441.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, October 25-29, 2014*, pages 1602–1613.

Sunita Goel, Jagdish Gangolly, Sue R. Faerman, and Özlem Uzuner. 2010. Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting*, 7:25–46.

Sunita Goel and Özlem Uzuner. 2016. Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239.

Jan Goldenstein, Philipp Poschmann, and Sebastian G. M. Händschke. 2015. Linguistic analysis: The study of textual data in management and organization studies with NLP. In *Academy of Management Proceedings*, volume 2015, page 10882. Academy of Management, Briarcliff Manor, NY.

Denis Griffis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert M. Lai. 2016. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. In *Proceedings of the AMIA 2016 Joint Summits on Translational Science. San Francisco, California, USA, March 21-24, 2016*, pages 88–97.

Petr Hájek and Vladimír Olej. 2015. Word categorization of corporate annual reports for bankruptcy prediction by machine learning methods. In *Text, Speech, and Dialogue. TSD 2015 — Proceedings of the 18th International Conference on Text, Speech, and Dialogue. Pilsen, Czech Republic, September 14-17, 2015*, number 9302 in Lecture Notes in Computer Science (LNCS), pages 122–130, Berlin. Springer.

Sebastian GM Händschke, Jan Goldenstein, and Peter Walgenbach. 2017. Cognitive isomorphism: Effects of management ideas as filters of organizational cognition. In *Academy of Management Proceedings*, volume 2017, page 14435. Academy of Management, Briarcliff Manor, NY.

Alexander Hogenboom, Frederik Hogenboom, Flavius Frasincar, Kim Schouten, and Otto van der Meer. 2013. Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications*, 64(1):27–52.

Nancy C. Ide and Keith Suderman. 2004. The American National Corpus First Release. In *LREC 2004 — Proceedings of the 4th International Conference on Language Resources and Evaluation. In Memory of Antonio Zampolli. Lisbon, Portugal, 24-30 May, 2004*, pages 1681–1684.

Kazushi Ikeda, Gen Hattori, Chihiro Ono, Hideki Asoh, and Teruo Higashino. 2013. Twitter user profiling based on text and community mining for market analysis. *Knowledge-Based Systems*, 51:35–47.

Takakazu Imada, Yusuke Inoue, Lei Chen, Syunya Doi, Tian Nie, Chen Zhao, Takehito Utsuro, and Yasuhide Kawada. 2016. Analyzing time series changes of correlation between market share and concerns on companies measured through search engine suggests. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 1917–1923.

Min Jiang, Yang Huang, Jung-Wei Fan, Buzhou Tang, Joshua C. Denny, and Hua Xu. 2015. Parsing clinical text: How good are the state-of-the-art parsers? *BMC Medical Informatics and Decision Making*, 15(Suppl 1):S2.

Siavash Kazemian, Shunan Zhao, and Gerald Penn. 2014. Evaluating sentiment analysis evaluation: A case study in securities trading. In *WASSA 2014 — Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ ACL 2014. Baltimore, Maryland, USA, June 27, 2014*, pages 119–127.

Wiltrud Kessler, Roman Klinger, and Jonas Kuhn. 2015. Towards opinion mining from reviews for the prediction of product rankings. In *WASSA 2015 — Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP 2015. Lisbon, Portugal, 17 September 2015*, pages 51–57.

Wiltrud Kessler and Jonas Kuhn. 2013. Detection of product comparisons: How far does an out-of-the-box semantic role labeling system take you? In *EMNLP 2013 — Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA, 18-21 October 2013*, pages 1892–1897.

Wiltrud Kessler and Jonas Kuhn. 2014. A corpus of comparisons in product reviews. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 2242–2248.

Brayden G. King. 2015. Organizational actors, character, and Selznicks theory of organizations. In Matthew S. Kraatz, editor, *Institutions and Ideals: Philip Selznicks Legacy for Organizational Studies*, pages 149–174. Emerald Group.

Brayden G. King, Teppo Felin, and David A. Whetten. 2010. Finding the organization in organizational theory. A meta-theory of the organization as a social actor. *Organization Science*, 21(1):290–305.

Antonina Kloptchenko, Tomas Eklund, Barbro Back, Jonas Karlsson, Hannu Vanharanta, and Ari Visa. 2004. Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance and Management*, 12(1):29–41.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *NAACL-HLT 2009 — Human Language Technologies: Proceedings of the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Boulder, Colorado, USA, May 31 - June 5, 2009*, volume 1, pages 272–280.

B. Shravan Kumar and Vadlamani Ravi. 2016. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114:128–147.

Marc Kupietz and Harald Lüngen. 2014. Recent developments in DeReKo. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 2378–2385.

Veronika Laippala, Timo Viljanen, Antti Airola, Jenna Kanerva, Sanna Salanterä, Tapio Salakoski, and Filip Ginter. 2014. Statistical parsing of varieties of clinical Finnish. *Artificial Intelligence in Medicine*, 61(3):131–136.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Daniel Jurafsky. 2014. On the importance of text analysis for stock price prediction. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 1170–1175.

Els Lefever and Véronique Hoste. 2016. A classification-based approach to economic event detection in Dutch news text. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 330–335.

Qing Li, Yan Chen, Jun Wang, Yuanzhu Chen, and Hsinchun Chen. 2018. Web media and stock markets : A survey and future directions from a big data perspective. *IEEE Transactions on Knowledge and Data Engineering*, 30(2):381–399.

Qing Li, TieJun Wang, Ping Li, Ling Liu, Qixu Gong, and Yuanzhu Chen. 2014. The effect of news and public mood on stock movements. *Information Sciences*, 278:826–840.

Kung-Yee Liang and Scott L Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, New York, NY.

Ling Liu, Jing Wu, Ping Li, and Qing Li. 2015. A social-media-based approach to predicting stock comovement. *Expert Systems with Applications*, 42(8):3893–3901.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Hsin Min Lu, Hsinchun Chen, Tsai Jyh Chen, Mao Wei Hung, and Shu Hsing Li. 2010. Financial text mining: Supporting decision making using Web 2.0 content. *IEEE Intelligent Systems*, 25(2):78–82.

Ronny Luss and Alexandre d'Aspremont. 2015. Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15(6):999–1012.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Renate E Meyer and Markus A Höllerer. 2010. Meaning structures in a contested issue field: A topographic map of shareholder value in Austria. *Academy of Management Journal*, 53(6):1241–1262.

Andrew Moore, Paul Rayson, and Steven Young. 2016. Domain adaptation using stock market prices to refine sentiment dictionaries. In *ESA 2016 — Proceedings of the [6th] Workshop on Emotion and Sentiment Analysis @ LREC 2016. Portorož, Slovenia, 23 May 2016*, pages 63–66.

Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *WWW '12 — Proceedings of the 21st Annual Conference on World Wide Web. Lyon, France, April 16-20, 2012*, pages 191–200.

Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh-Ying Wah, and David Chek Ling Ngo. 2014. Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Dean Neu, Hussein Warsame, and Kathryn Pedwell. 1998. Managing public impressions: Environmental disclosures in annual reports. *Accounting, Organizations and Society*, 23(3):265–282.

Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *ACL-IJCNLP 2015 — Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics & 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China, July 26-31, 2015*, pages 1354–1364.

Nuno Oliveira, Paulo Cortez, and Nelson Areal. 2016. Stock market sentiment lexicon acquisition using microblogging data and statistical measures. *Decision Support Systems*, 85:62–73.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Viktor Pekar and Jane Binner. 2017. Forecasting consumer spending from purchase intentions expressed on social media. In *WASSA 2017 — Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP 2017. Copenhagen, Denmark, September 8, 2017*, pages 92–101.

Reid Pryzant, Young-joo Chung, and Daniel Jurafsky. 2017. Predicting sales from the language of product descriptions. In *SIGIR eCom 2017 — Proceedings of the ACM SIGIR Workshop on eCommerce. Tokyo, Japan, August 11, 2017*.

Xin Ying Qiu and Padmini Srinivasan. 2014. Supervised learning models to predict firm performance with annual reports: An empirical study. *Journal of the Association for Information Science and Technology*, 65(2):400–413.

Michael Rechenthin, W. Nick Street, and Padmini Srinivasan. 2013. Stock chatter: Using stock sentiment to predict price direction. *Algorithmic Finance*, 2(3-4):169–196.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Allan Hanbury, Alexander Dür, and Linda Anderson. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based ir models. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, British Columbia, Canada, July 30 - August 4, 2017*, volume 1: Long Papers, pages 1712–1721.

Brian A Rutherford. 2005. Genre analysis of corporate annual report narratives: A corpus linguistics-based approach. *The Journal of Business Communication*, 42(4):349–378.

Shigeyuki Sakaki, Francine Chen, Mandy Korpusik, and Yan-Ying Chen. 2016. Corpus for customer purchase behavior prediction in social media. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 2976–2980.

Saskia Schön, Veselina Mironova, Aleksandra Gabryszak, and Leonhard Hennig. 2018. A corpus study and annotation schema for named entity recognition and relation extraction of business products. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 4445–4451.

Robert P. Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The AZFINTEXT system. *ACM Transactions on Information Systems*, 27(2):#12.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Sinno Jialin Pan, Qing Li, and Huayi Li. 2014. Exploiting social relations and sentiment for stock prediction. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, October 25-29, 2014*, pages 1139–1145.

Herbert A. Simon. 1991. Organizations and markets. *Journal of Economic Perspectives*, 5(2):25–44.

Anna Stavrianou and Caroline Brun. 2015. Expert recommendations based on opinion mining of user-generated product reviews. *Computational Intelligence*, 31(1):165–183.

Pyry Takala, Pekka Malo, Ankur Sinha, and Oskar Ahlgren. 2014. Gold-standard for topic-specific sentiment analysis of economic texts. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 2152–2157.

Yi-jie Tang and Hsin-Hsi Chen. 2014. FADR: A system for recognizing false online advertisements. In *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland, USA, June 23-24, 2014*, pages 103–108.

Fangbo Tao, Bo Zhao, Ariel Fuxman, Yang Li, and Jiawei Han. 2015. Leveraging pattern semantics for extracting entities in enterprises. In *WWW 2015 —*

Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, May 18–22, 2015, pages 1078–1088.

Katrin Tomanek, Joachim Wermter, and Udo Hahn. 2007. A reappraisal of sentence and token splitting for life sciences documents. In *MedInfo 2007 — Proceedings of the 12th World Congress on Health (Medical) Informatics. Building Sustainable Health Systems. Brisbane, Australia, August 20-24, 2007*, number 129 in Studies in Health Technology and Informatics, pages 524–528, Amsterdam. IOS Press.

Andrew Trotman, Jon Degenhardt, and Surya Kallumadi. 2017. The architecture of EBAY SEARCH. In *SIGIR eCom 2017 — Proceedings of the ACM SIGIR Workshop on eCommerce. Tokyo, Japan, August 11, 2017*.

Ming-Feng Tsai and Chuan-Ju Wang. 2017. On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, 257(1):243–250.

Takaaki Tsunoda, Takashi Inui, and Satoshi Sekine. 2015. Utilizing review analysis to suggest product advertisement improvements. In *WASSA 2015 — Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis @ EMNLP 2015. Lisbon, Portugal, 17 September 2015*, pages 41–50.

Marjan Van De Kauter, Diane Breesch, and Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 42(11):4999–5010.

Damir Vandic, Steven S. Aanen, Flavius Frasincar, and Uzay Kaymak. 2017. Dynamic facet ordering for faceted product search engines. *IEEE Transactions on Knowledge and Data Engineering*, 29(5):1004–1016.

Martin Žnidaršič, Jasmina Smailović, Jan Gorše, Miha Grčar, Igor Mozetič, and Senja Pollak. 2018. Trust and doubt terms in financial tweets and periodic reports. In *FNP 2018 — Proceedings of the 1st Financial Narrative Processing Workshop @ LREC 2018. Miyazaki, Japan, 7 May 2018*, pages 59–65.

Hao Wang and Martin Ester. 2014. A sentiment-aligned topic model for product aspect rating prediction. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, October 25-29, 2014*, pages 1192–1202.

William Yang Wang and Zhenhao Hua. 2014. A semiparametric Gaussian copula regression model for predicting financial risks from earnings calls. In *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, USA, June 22-27, 2014*, pages 1155–1165.

Albert Weichselbraun, Daniel Streiff, and Arno Scharl. 2015. Consolidating heterogeneous enterprise data for named entity linking and Web intelligence. *International Journal on Artificial Intelligence Tools*, 24(2):#1540008.

Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer. 2013. Semantic frames to predict stock price movement. In *ACL 2013 — Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, August 4-9, 2013*, pages 873–883.

Frank Z. Xing, Erik Cambria, and Roy E. Welsch. 2018. Natural language based financial forecasting: A survey. *Artificial Intelligence Review*, 50(1):49–73.

Chao Yang, Shimei Pan, Jalal U. Mahmud, Huahai Yang, and Padmini Srinivasan. 2015. Using personal traits for brand preference prediction. In *EMNLP 2015 — Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 17-21 September 2015*, pages 86–96.

Kristi Yuthas, Rodney Rogers, and Jesse F Dillard. 2002. Communicative action and corporate annual reports. *Journal of Business Ethics*, 41(1-2):141–157.

Yongzheng Zhang and Marco Pennacchiotti. 2013. Predicting purchase behaviors from social media. In *WWW '13 — Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, May 13-17, 2013*, pages 1521–1532.

A Supplemental Material

In general, the estimating technique must address the main characteristics of the data at hand. Due to the repeated observations over the eight years (from 2008 to 2015), the investigated cases are not independent from each other which increases the likelihood of autocorrelation in the data. In order to appropriately deal with this issue, we employ the generalized estimating equations (GEE) method (Liang and Zeger, 1986). We report population-average estimators with fixed effects that allow us to control for organizational differences we cannot account for directly. Also, this model allows for omitting observable but stable organizational characteristics. We use a normal distribution for modeling the dependent variable.

	ROE	Valence	Arousal	Dom.	lnEmpls	lnAssets	Sales/Empl.	Sales/Assets	Country
ROE	1								
Valence	.03	1							
Arousal	-.02	-.68	1						
Dominance	.56	.90	-.70	1					
ln(Employees)	.10	-.05	.22	-.12	1				
ln(Assets)	.50	.02	-.19	-.10	.77	1			
Sales/Employee	-.02	.06	-.06	.22	-.29	.07	1		
Sales/Assets	.01	.00	-.03	.54	-.97	-.37	-.06	1	
Country	-.58	-.12	.07	-.37	.13	.83	.08	.50	1

Table 3: Correlation matrix of independent, dependent and control variables in the GEE model. 'Country' is coded as Germany(GER)= 1, UK= 0.

	Model I: Controls			Model II: Explanatory			Model III: Full		
	Beta	S.E.	Sig.	Beta	S.E.	Sig.	Beta	S.E.	Sig.
Arousal				-.067	.055	.228	-.158	.040	.000
Dominance				-.019	.040	.636	.016	.061	.795
Arousal*Country							.189	.047	.000
Dominance*Country							.022	.066	.737
lnEmployees	.080	.030	.007	.085	.312	.007	.082	.031	.008
lnAssets	-.058	.030	.050	-.057	.030	.056	-.056	.030	.059
Sales/Employee	.049	.023	.037	.048	.024	.041	.050	.024	.035
Sales/Assets	.004	.038	.915	.004	.038	.914	-.004	.038	.915
Country	-.191	.088	.030	-.196	.098	.046	-.159	.098	.103
Constant	.253	.357	.480	.196	.366	.591	.191	.368	.603

Table 4: Results of GEE panel regression with dependent variable ROE lagged by one year and interaction effects of arousal and dominance with the country dummy (GER=1). Columns give the respective slope coefficient (Beta), standard error (S.E.) and p-value (Sig.). The three models differ in the set of variables taken into account. The number of cases is 1,127 for each model (one AR per corporation per year in the application's subsample of the corpus).

.