ISSN 2007-9737

Automatic Analysis of Annual Financial Reports: A Case Study

Jasmina Smailović[1], Martin Žnidaršič[1], Aljoša Valentinčič[2], Igor Lončarski[2], Marko Pahor[2], Pedro Tiago Martins[1], Senja Pollak[1]

[1] Jožef Stefan Institute, Ljubljana, Slovenia

[2] University of Ljubljana, Faculty of Economics, Ljubljana, Slovenia

senja.pollak@ijs.si

Abstract. The main goal of reporting in the financial system is to ensure high quality and useful information about the financial position of firms, and to make it available to a wide range of users, including existing and potential investors, financial institutions, employees, the government, etc. Formal reports contain both strictly regulated, financial sections, and unregulated, narrative parts. Our research starts from the hypothesis that there is a relation between business performance and not only content, but also the linguistic properties of unregulated parts of annual reports. In the paper we first present our dataset of financial reports and the techniques we used to extract the unregulated textual parts. Next, we introduce our approaches of differential content analysis and analysis of correlation with financial aspects. The differential content analysis is based on TF-IDF weighting and is aimed at finding the characteristic terms for each year (i.e. the terms which were not prevailing in the previous reports by the same firm). For correlation of linguistic characteristics of reports with financial aspects, an array of linguistic features was considered and selected financial indicators were used. Linguistic features range from measurements, such as personal/impersonal pronouns ratio, to assessments of characteristics like financial sentiment, trust, doubt, and discursive features expressing certainty, modality, etc. While some features show strong correlation with industry (e.g., shorter and more personal reports by IT industry compared to automotive industry), doubt, communication – as well as necessity and cognition words to some extent – are positively correlated with failure.

Keywords. Financial reports, 10-K, differential content analysis, linguistic characteristics, financial indicators.

1 Introduction

The main goal of the financial system is to facilitate the transfer of funds from savers with excess funds to entities that require funds for capital investment. To ensure that this transfer is efficient, savers require high quality information. The objective of financial reporting is to ensure high quality and useful information about the financial position of firms, and to make it available to a wide range of users, including investors, employees, the government, etc.

Financial reports are required from publicly traded companies [44] and contain both strictly regulated financial sections, providing the information on the status, properties and practices of companies, and unregulated parts, where flexibility is allowed and required. Even though financial disclosures follow a reasonably well-established set of guidelines, there is variation in terms of how the content of these disclosures is conveyed, especially in sections that allow for a more narrative style. The choice of specific words and tone when framing a disclosure could be indicative of the underlying facts about a company's financial situation that cannot be conveyed through financial indicators alone.

The Securities and Exchange Commission (SEC), in the United States, stipulates filing of financial disclosures by publicly traded companies: an annual report (Form 10-K), quarterly reports (Form 10-Q) and other reports to communicate

major events within the company (Form 8-K). In our paper we focus on 10-K reports. They conform to a set number of sections containing information on the business practices of a company, their risks, and results for the fiscal year for which the report is filed. Many of these sections allow for relatively little variation, as their main function is to declare facts about the company, such as enumeration of their properties, shareholders and similar information.

However, there are also sections with more flexible contents. In particular, one section is of special interest: Management's Discussion and Analysis (MD&A), which discloses company operations and management in a way that is easy for investors and other interested parties to understand. It also includes information on what the company does in the face of risks, legislation, competition, etc. This section allows more freedom and variation in the framing of the information contained, and as such provides an insight into the nature of a company's management and its prospects to a degree that goes beyond the quantitative facts stated in other sections. They contain opinions and attitudes, and are subject to style. For this reason, the MD&A section is an important component in many works in qualitative analysis of financial reports, allowing for the exploitation of linguistic features.

There is now a growing body of literature dedicated to the analysis of non-financial information in financial reports. The motivation for such research is to assess whether there is a relationship between the way a report is framed and the past and future performances of the companies. The hypothesis is that e.g., negative performance is reported with different stylistic devices than positive performance, while textual analysis can be also an indicator for future performance. Even though some of the linguistic choices in financial reporting are subjective, research shows that a relationship exists between different textual aspects of financial reports and real-world financial indicators.

In our study, we contribute to this line of research by investigating the narrative parts of 10-K annual reports, where on the one hand we are interested in how through content analysis we can identify events, characteristic for the reported year, and on the other hand, how linguistic characteristics of the reports reflect a firm's business performance. We showcase our methodology on the analysis of annual reports by four companies.

2 Related Work

Given the financial and cultural prominence of the American Stock markets, as well as the availability of their gray-literature, 10-Q and 10-K are the kinds of reports most often used in research on financial reports, the latter being the focus of our study. However, other types of reports are occasionally used in related work (e.g., the filings of the firms in the London Stock Exchange [3], the Hong Kong Stock Exhange [30], or Euronext Paris [1]). Non-financial information from reports has been used for prediction of financially relevant events [40], such as next year performance (through indicators such as return on equity) [40, 9, 26, 4, 20, 30], contemporaneous returns around filing dates [15], stock return volatility [32, 33], earnings forecast dispersion [28, 33], costs of capital [28], financial distress [21], credibility of reports [3] or fraud detection [18].

The role of linguistic features has been explored by several authors. A lot of attention has been paid to linguistic sentiment available in dictionaries. Popular general dictionaries and dictionary-based programs are the General Inquirer/Harvard [45], used in works such as [15, 13, 34, 16, 39, 28] and DICTION [22], used in works such as [12, 11, 42, 16, 20]. Given their generic character, these dictionaries present some limitations, since words that are generally considered positive or negative might convey a different, and sometimes opposite, meaning in a financial context. However, attempts have been made to counter this by devising domain-specific dictionaries [23, 34], which were used in various studies [13, 42, 16, 25, 39, 19, 21, 35, 18]. Also in our study we use the LoughranMcDonald Master Dictionary [34]. Other linguistic features addressed in related work include forward looking performance keywords [3] and statements [32], risk sentiment [4], polarity/tone [12, 15, 18, 19, 20, 25], subjectivity [18], intensity [18], modality [20, 34] and uncertainty [20, 34].

The most related to our approach is the work by Pimenta et al. [38]. While they propose a system for extraction and cleaning of content from UK annual reports, we focus on the annual reports of US firms. They use the extracted data to assess the predictive value of various sections for future earnings, based on tone. We, on the other hand, investigate the relation between the linguistic characteristics of non-regulated textual parts of reports and firm's performance in the reported period.

We analyze the correlation between financial performance and document length (cf. also the findings by Li [31] and Loughran and McDonald [35]), the observed sentiment (we use a sentiment dictionary [34]), and several other features: the newly introduced trust and doubt keywords, discursive features by Biber [6] and the ratio between personal and impersonal pronouns, which can be related to the more general observation by Pannebaker [37] on the important role of pronouns, as well as to the passive voice by Merkl-Davies and Koller [36]. We showcase the methodology on a case study of four firms.

3 Dataset of Financial Reports

Our case study focuses on 10-K annual reports of four firms from two industries in a period of ten years (2005-2014). From the automotive industry we chose Ford Motor Company (F) and General Motors Company (GM), and from the IT industry we selected Google - Alphabet Inc. (GOOAV) and Yahoo! Inc. (YHOO).

From the original 10-K reports we extract Part I, and Items 7 and 7A from Part II, since these are the non-financial, non-regulated parts of annual reports, in which management can express their opinion on the past and future performance, etc. However, they are mandated in the sense that they must appear in every annual report and can thus be followed through time as well as in a cross-section. Automatic filtering of report files is challenging due to their inconsistent and error-prone format, so here we provide some useful details of our approach. The most important components of the cleaning/extracting process include:

— Read an original report line by line.

— Perform decoding, removing leading and trailing white-space characters, replacing multiple white-space characters with one space character.

— Skip potential .pdf, .xls, .jpg, .zip, .gif objects.

— If the line contains more than 10,000 characters, split it into a list of elements and process them individually.

— If the line contains more '<' signs than '>' signs (possible unfinished tag), merge it with the next line. After 100 iterations break the loop.

— Detect relevant parts by searching for the titles of the section (e.g. Part I) (which should not be mistaken for the references to that parts in text).

— Skip tables and remove html/xml tags.

4 Financial Indicators

For our analysis of the correlation between linguistic characteristics and financial aspects, we investigate the role of several measures of failure based on existing literature. The first group of indicators is based on financial statements. Firms avoid reporting losses and/or decreases in earnings (*LOSS, DECR_P*) [8, 17] for several reasons such as access to capital markets, loan contracting, management compensation, etc.

Because both measures depend in part on capital structure (i.e. the amount of debt used to finance the firm), a cleaner version is to define failure if operating income (i.e. earnings from business activities) is negative (*OP-LOSS*). Managers exert discretion over the financial reporting purposes and they would normally do so only through the accruals component of earnings [43, 27]. Hence, to avoid this subjective element, another measure of failure indicates when the cash flows from operations are negative (*N-CFO*).

The second group of indicators are market-based indicators. Failure here occurs when total returns are negative (*Neg-RET1*), i.e. when the

ISSN 2007-9737

share price and dividends fall below previous year's level. As not all decreases of share prices can be attributed to firms only, we also subtract the market-wide return (the return on a broad stock index) from individual company returns and define failure only based the residual part of the return that is attributable to the firm (*Neg-RET2*). Finally, another measure that defines failure is if firms fail to secure a return on equity of 8% (LowROE). This is a conservative (i.e. low) cost of equity that firms are expected to cover over longer periods of time.

The third group is based on dividends. Firms that pay no dividends (*NO_DIV*) are different from firms that do pay out dividends (see e.g. [41, 24]). Moreover, dividend decreases (*DIV_DECR*) are considered to be a very negative signal to the market and are typically accompanied by a negative market response.

We also look at investments expenditure, more importantly at R&D expenditures that also play an important role in firm valuation [2]. It is relatively easy for firms to reduce R&D expenditures (and expenses), by which they temporarily boost the net income in the short run, but destroy value over longer periods. We therefore add the indicator *R&D-RED*.

All the above-presented indicators were transformed to binary values, 1 indicating failure and 0 indicating positive performance. There are also missing values due to the following reasons: missing previous year data item in increases and decreases, negative book value of equity and a loss at the same time, companies not reporting R&D expenditure.

Finally, we also correlate text properties with the relative amount of financial debt (*D/A*). More indebted companies are more financially constrained and, hence, more sensitive to any deterioration in their business performance. As finance theory generally does not unequivocally prescribe the optimal level of debt financing, we use this variable in continuous form rather than as binary (indicator) variable.

5 Analysis

We first present a differential content analysis of 10-K annual reports (Section 5.1), followed by the analysis of correlation between various linguistic features in the reports and financial indicators of corresponding companies and time periods (Section 5.2).

5.1 Differential Content Analysis

The differential content analysis is based on TF-IDF weighting and is aimed at finding the characteristic terms that start appearing in annual reports, i.e. the terms which were not prevailing in the previous reports by the same firm. We employed the LATINO library[1] for preparing the features, creating the Bag-of-words model and calculating the TF-IDF weights. Table 1 presents the top five characteristic terms (unigrams and bigrams) per company and year.

The 2006-2014 period under study has been characterized by two major factors. First, the global financial and economic crisis, arguably one of the most severe crises ever, and certainly the most severe in the post-WW2 period. Virtually no company could escape its consequences one way or another. The second significant change in this period is the development of communication technology. As an example relevant to our research, Twitter, Inc. was founded only in 2006, yet it has rapidly become one of the main channels through which firms communicate in an informal way with customers, investors, etc. Today it has over 310 million active users. Academic research has shown that these communications can affect significantly the capital market (e.g., [29]). Both of these major factors determine significantly what we see in less regulated parts of annual reports of Google, Yahoo, Ford and GM.

As an example, Google, Yahoo and Ford all mention the following expressions: risk management, global economic crisis, restructuring plan. We note that GM in all likelihood would, too, had it not been rescued from Chapter 11 bankruptcy by the government in 2008. Ford has entered this period by recording a record loss for

[1] https://github.com/LatinoLib/LATINO

Table 1. Top 5 characteristic terms (unigrams and bigrams) per company and year

Company	Year	Top 5 characteristic terms
F	2006	contents item, table contents, cost changes, indebtedness, wholesale unit,
	2007	deferred tax, tax assets, valuation allowance, goodwill, uaw represented,
	2008	crisis, ghg, deferred tax, global economic, tax assets,
	2009	ghg, new note, ghg emissions, warrants, talf,
	2010	ghg, earnings per, ghg emissions, valuation allowance, costs expenses,
	2011	charts detail, causal, causal factor, tax operating, compared causal,
	2012	charts detail, causal factor, causal, compared causal, mid decade,
	2013	results compared, causal, causal factor, charts detail, compared causal,
	2014	field service, following two, two charts, east africa, causal factor,
GM	2010	confidential, contents confidential, confidential general, ally, ally financial,
	2011	ebit adjusted, due primarily, ally, ally financial, primarily increased,
	2012	retiree plan, due primarily, psa, annuity, ebit adjusted,
	2013	retail vehicle, due primarily, wholesale vehicle, acquisition ally, mix due,
	2014	ignition switch, switch recall, suvs, ignition, courtesy,
GOOAV	2006	checkout, google checkout, youtube, audio, audio ads,
	2007	including increase, headcount including, tv, checkout, december year,
	2008	exchange offer, new options, display advertising, risk management, management program,
	2009	average cost, visualize, management program, risk management, display advertising,
	2010	members websites, websites, web spam, google websites, average cost,
	2011	members websites, websites, motorola, google websites, web spam,
	2012	motorola, arris, class capital, motorola mobile, websites,
	2013	segment, motorola, google segment, motorola mobile, mobile segment,
	2014	motorola, class capital, motorola mobile, divestiture, certain currencies,
YHOO	2006	table contents, audience users, reorganization, global audience, digital home,
	2007	affiliate sites, table contents, operated sites, primarily display, publishers developers,
	2008	affiliate sites, reduction initiatives, cost reduction, restructuring charges, table contents,
	2009	search agreement, restructuring plan, restructuring charges, algorithmic paid, microsoft will,
	2010	ex, revenue ex, ex tac, asia pacific, pacific,
	2011	ex, ex tac, revenue ex, asia pacific, pacific,
	2012	ex, revenue ex, ex tac, ebitda, adjusted ebitda,
	2013	hedges, tumblr, marketable securities, ex tac, ex,
	2014	revenue ex, ex, ex tac, marketable securities, adss,

2006 ($>$12bn USD) and was massively in debt (hence the phrase "indebtedness"). In contrast, after 2012 Ford was emerging from crisis and expressions such as casual, results, etc., appear several times in the period. This is consistent with managers self-attributing success, but not failure (e.g., [10, 14]).

Yahoo had been in a state of permanent crisis up to mid-2012 when Marissa Mayer—a former Google executive—was appointed as CEO. Seen at the time as an important step for future development, that, too, ended in disarray with her departure announced in January 2017. An important asset of Yahoo is its stake in Alibaba, the largest e-commerce firm in the world that originates in China. The stake is worth about 35 bn USD. Hence, the term "pacific" appears often in its reports. For Google, an important milestone has been the acquisition of YouTube, which they announced in 2006. An important decision for Google was also the purchase of Motorola, which was announced in 2011. This was not only related to mobile handsets, but also to Motorola's patent library. Hence terms such as "Motorola" and "mobile segment" appear, but note also the term "divestiture" as a result of Motorola's sale to Lenovo in 2014.

5.2 Correlation of Linguistic Characteristics with Financial Aspects

In this section we present the process of computing the linguistic features from the 10-K reports and calculating the correlation between the features and the corresponding financial aspects of a company (see Figure 2).

ISSN 2007-9737

The simplest feature is the length of the documents (*Number of words*), followed by *pers_it* – the proportion of first person personal pronouns (I, we) compared to the impersonal pronoun "it".

Next, we were interested in linguistic features of three types, based on sentiment, doubt and discursive word lists. The first category was calculated on raw text, while for the second two categories we first lemmatized the documents by WordNet lemmatizer [7].

Positive and negative sentiment was calculated by applying a simple lexicon-based approach. For each report we count the occurrence of 354 *positive* and 2,355 *negative words* from the LoughranMcDonald Master Dictionary [34], as well as the relative frequencies of the positive and negative words with respect to the number of all the words in a document – *Positive (%)* and *Negative (%)*. Note that, as explained in Section 3, only Part I and Items 7 and 7A from Part II were extracted. In Figure 1, we also visualize the relative sentiment results and the word count per company. The results suggest that there is a difference in the length of reports between the companies in the two showcased industries, as well as a slight tendency for longer texts in difficult periods (2008 and 2009 for Ford and GM). While the sentiment indicators show some variation over years for certain companies (e.g., Google), they do not seem to follow a particular common rule or pattern over all the companies.

Next, we created the dictionary of *trust* and *doubt*-related words (available at: `http://kt.ijs.si/data/trust_doubt_wl.zip`). For that purpose, we collected (near) synonyms of words related to "trust" and "doubt" from WordNet[2] and online dictionaries[3]. The word lists contain 20 words for trust (e.g., trustful, confidence) and 72 for doubt (e.g., uncertainty, untrusting, suspicion). For each feature (word list), we represent the trust/doubt values as relative frequencies of words from a word list with respect to the total length of a report (only the extracted parts). To the best of our knowledge, these are the first publicly available trust/doubt word lists.

[2]http://wordnet.princeton.edu, last accessed: Nov. 5, 2016.

[3]For example: http://thesaurus.yourdictionary.com/doubt, last acc.: Nov. 5, 2016.

The last group of features is based on discourse markers by Biber et al. [6] (listed in [5, pp.69–72]), listing words and grammatical devices used to express stance. The relative frequencies of words from different word lists are used: *causation/modality/effort* (e.g., afford, allow), *premodifying adverbs* (e.g., completely, extremely), *communication* (e.g., add, announce), *modal_possibility* (e.g., can, may), *ability* (e.g., able), *evaluation* (e.g., acceptable, advisable), *modal_prediction* (e.g., will, would), *ease/difficulty* (e.g., difficult, ease), *cognition* (e.g., assume, believe), *modal_necessity* (e.g., must, should), *nouns_various*, *attitude/emotion*, *likelihood*, *desire/decision* (e.g., agreement, commitment), *certainty* (undoubtedly, certainly), *style* (e.g., accordingly, definitely).

Correlations among linguistic features and financial indicators are presented in Figure 2, where also an indicator of the industry (IT or automotive) is shown in order to provide an insight into linguistic differences of reports in different industries and to allow for assessing the impact of industry-specific reporting. Namely, strong correlations of some linguistic features and particular financial indicators (e.g., low debt/assets ratio (*D/A*) and no dividends (*NO_DIV*)) might be more related to specific industries than to financial performance of the companies. This issue is particularly relevant in our case, since we only investigate four companies in two industries. Therefore, we are particularly interested in linguistic features that have a high correlation with one or more financial indicators and a low correlation with the industry identifier, as those are presumably the features that are not industry specific, but rather imply a more general relationship with financial indicators and, hence, allow for a more generalized conjectures.

The correlation analysis on four showcased companies allows three general observations:

1. **Industry-specific writing characteristics.** Several linguistic characteristics seem to be strongly related to the industry classification of the reporting company. The IT industry appears to produce shorter annual reports reflected also in the lower absolute counts of both positive and negative words. However,

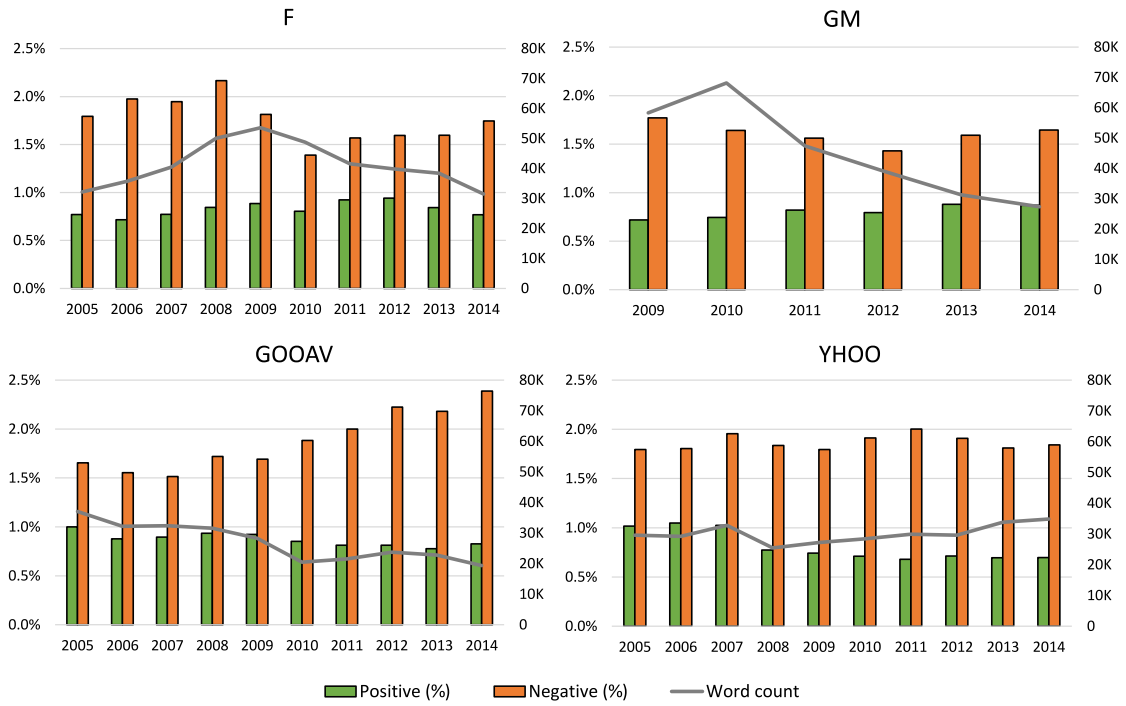

Fig. 1. The relative positive and negative sentiment results (left axis) and word count (right axis) per company in time

the relative sentiment does not differ widely between the two industries in our study, there is only a marginally significant difference in the share of negative words. Also some other linguistic characteristics display strong differences between the two industries. The IT sector seems to use a more personal way of writing (*pers_it*) with more first person personal pronouns used (I, we) compared to impersonal pronoun (it). Similarly *ModalPossibility* and *AttitudeEmotion* are more frequent in reports of IT companies, while *DesireDecision* is lower in these reports.

2. **Weak effect of positive/negative words.** Percentages of positive or negative words in reports are not strongly correlated with financial indicators. The most likely explanation for this is that writers of the reports tend to keep reports as neutral as possible, thus keeping the sentiment fairly stable over the years, success states and also industry. The expressions of sentiment can be controlled by carefully selected wording, while on the other hand more unconscious stylistic features might be more interesting to observe.

3. **Some linguistic features have strong correlations with financial indicators.** As explained earlier, our interest is focused on the features that exhibit a strong correlation with a financial indicator and are at the same time not attributable to a specific industry. As expected, the *Doubt* score is positively correlated with failure (weak or bad performance). We also find *Communication* to be of this kind, since it is positively correlated with financial indicators of decrease in profits and the reduction in R&D. This indicates that writers of reports are more likely to use words, such as “assure”, “announce”, “promise”. To some extent, we can also observe the correlation of the words of *Cognition* (e.g., assume, consider) and *Modality_necessity* (e.g. must, should).

	IT ind.	LOSS	DECR_P	OP-LOSS	N-CFO	Neg-RET1	Neg-RET2	LowROE	NO_DIV	DIV_DECR	R&D-RED	D/A
Positive words	**-0.605**	0.144	0.164	0.137	0.214	0.117	0.066	0.084	-0.038	0.005	0.010	**0.431**
Negative words	**-0.564**	0.413	0.203	0.254	**0.471**	0.264	0.194	**0.547**	0.052	0.209	0.220	0.421
Number of words	**-0.664**	0.205	0.175	0.270	0.315	0.153	0.061	0.270	-0.018	0.094	0.109	0.415
Positive (%)	0.111	-0.161	0.036	-0.268	-0.220	-0.020	0.070	-0.233	-0.013	-0.225	-0.349	-0.001
Negative (%)	0.385	0.315	-0.065	-0.098	0.205	0.213	0.301	0.191	0.209	0.174	0.109	-0.133
pers_it	**0.920**	-0.405	-0.220	-0.303	-0.152	-0.254	-0.124	-0.422	**0.489**	-0.350	-0.027	**-0.838**
trust	**-0.506**	-0.180	0.222	0.324	0.027	0.114	-0.023	-0.117	0.125	-0.133	0.138	-0.011
doubt	-0.081	**0.554**	0.298	-0.005	0.167	**0.545**	**0.433**	0.395	0.092	0.333	0.108	0.132
CausModEffort	**0.465**	-0.098	0.132	-0.313	-0.167	-0.154	-0.023	-0.103	0.152	-0.090	-0.092	-0.174
PremodAdv	**-0.586**	**0.459**	-0.141	-0.030	-0.010	0.038	0.006	**0.459**	**-0.507**	0.363	-0.045	**0.878**
Communication	-0.262	0.291	0.402	0.058	0.148	0.151	0.183	0.268	-0.242	0.271	**0.457**	0.382
ModalPossiblity	**0.943**	-0.294	-0.336	-0.345	-0.175	-0.214	-0.133	-0.317	**0.459**	-0.248	-0.212	**-0.733**
Ability	**0.453**	0.071	-0.206	-0.210	-0.087	-0.028	0.012	-0.031	0.421	-0.036	-0.109	-0.172
Evaluation	**0.853**	-0.370	-0.264	-0.388	-0.236	-0.178	-0.042	-0.414	0.430	-0.386	-0.341	**-0.665**
ModalPrediction	-0.085	0.167	-0.350	0.028	0.191	-0.305	-0.340	0.069	0.181	0.052	0.091	0.238
EaseDifficulty	**0.588**	-0.141	-0.295	-0.206	-0.063	0.005	0.010	-0.217	0.380	-0.198	-0.153	**-0.529**
Cognition	-0.379	**0.436**	0.002	-0.126	-0.075	0.108	0.133	0.404	**-0.480**	0.330	0.169	**0.774**
ModalNecessity	**-0.433**	**0.716**	0.300	0.294	0.179	0.356	0.265	**0.684**	-0.077	**0.567**	0.379	**0.528**
Nouns_various	**-0.647**	0.054	0.339	0.426	0.337	0.170	0.146	0.091	-0.087	0.036	0.415	0.285
AttituteEmotion	**0.862**	-0.226	-0.104	-0.110	-0.162	-0.170	-0.179	-0.166	**0.461**	-0.106	-0.030	**-0.754**
Likelihood	0.086	0.338	0.009	-0.163	0.196	0.224	0.252	0.201	-0.086	0.182	-0.026	0.294
DesireDecision	**-0.950**	0.266	0.272	0.427	0.189	0.248	0.202	0.299	-0.429	0.247	0.278	**0.667**
Certainty	**0.613**	-0.248	-0.207	-0.179	-0.135	-0.143	-0.208	-0.249	0.293	-0.196	-0.102	**-0.590**
Style	-0.240	0.072	-0.123	0.079	-0.140	0.003	-0.124	0.121	-0.130	0.124	-0.128	0.072

* Correlations above 0.320 are significant at 0.05, above 0.430 significant at 0.01 and above 0.540 significant at 0.001

Fig. 2. Correlations between financial indicators and sentiment/linguistic characteristics of reports. Correlations significant at 0.05 are colored green/red, while the ones significant at 0.01 are also shown in bold text

6 Conclusions

In this paper we perform a case study of the linguistic and sentiment properties of annual reports (10-K) filed by public companies. We do so by investigating cases of four companies from two diametrically different industries - a traditional, capital intensive automotive industry and a relatively fresh, high intellectual capital intensive IT industry. To further achieve diversity, we select companies with diverse situations and financial indicators in each of the two industries.

The main finding of this preliminary research is that even though there are important differences between the characteristics of the reports, our assessment of their general sentiment remains largely the same. Neither industry nor financial state seems to have an impact on the general sentiment of the reports. This seems to indicate that the writers of annual reports attempt to preserve their informational content neutral.

On the other hand, some other linguistic characteristics relatively strongly relate to the financial state of the company. Even after accounting for the inter-industrial differences, companies in financial distress use wording that expresses more doubt and necessity. Our explanation for this is that while the writers of the reports may choose the main wording, expressing the general positive/negative sentiment carefully enough in order to avoid a bias and keep the report neutral in terms of sentiment, they may have less conscious control over the modal words that are captured by these other linguistic characteristics.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency for research

core funding (No. P2-0103 and No. P5-0161), as well as for funding of the research project *Influence of formal and informal corporate communications on capital markets* (No. J5-7387).

References

1. **Ajina, A., Laouiti, M., & Msolli, B. (2016).** Guiding through the fog: Does annual report readability reveal earnings management? *Research in International Business and Finance*, Vol. 38, pp. 509–516.
2. **Akbar, S. & Stark, A. W. (2003).** Deflators, net shareholder cash flows, dividends, capital contributions and estimated models of corporate valuation. *Journal of Business Finance & Accounting*, Vol. 30, No. 9-10, pp. 1211–1233.
3. **Athanasakou, V. & Hussainey, K. (2014).** The perceived credibility of forward-looking performance disclosures. *Accounting and Business Research*, Vol. 44, No. 3, pp. 227–259.
4. **Balakrishnan, R., Qiu, X. Y., & Srinivasan, P. (2010).** On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, Vol. 202, No. 3, pp. 789–801.
5. **Biber, D. (2007).** *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*, volume 28. John Benjamins Publishing.
6. **Biber, D., Finegan, E., Johansson, S., Conrad, S., & Leech, G. (1999).** *Longman Grammar of Spoken and Written English*. Longman.
7. **Bird, S., Klein, E., & Loper, E. (2009).** *Natural Language Processing with Python*. O'Reilly Media.
8. **Burgstahler, D. & Dichev, I. (1997).** Earnings management to avoid earnings decreases and losses. *Journal of accounting and economics*, Vol. 24, No. 1, pp. 99–126.
9. **Butler, M. & Kešelj, V. (2009).** Financial forecasting using character n-gram analysis and readability scores of annual reports. *Canadian Conference on Artificial Intelligence*, Springer, pp. 39–51.
10. **Chen, W., Han, J., & Tan, H.-T. (2016).** Investor reactions to management earnings guidance attributions: The effects of news valence, attribution locus, and outcome controllability. *Accounting, Organizations and Society*, Vol. 55, pp. 83–95.
11. **Davis, A. K., Piger, J. M., & Sedor, L. M. (2012).** Beyond the numbers: Measuring the information content of earnings press release language. *Contemporary Accounting Research*, Vol. 29, No. 3, pp. 845–868.
12. **Davis, A. K. & Tama-Sweet, I. (2012).** Managers' use of language across alternative disclosure outlets: Earnings press releases versus md&a. *Contemporary Accounting Research*, Vol. 29, No. 3, pp. 804–837.
13. **Doran, J. S., Peterson, D. R., & Price, S. M. (2012).** Earnings conference call content and stock price: the case of reits. *The Journal of Real Estate Finance and Economics*, Vol. 45, No. 2, pp. 402–434.
14. **Doukas, J. A. & Petmezas, D. (2007).** Acquisitions, overconfident managers and self-attribution bias. *European Financial Management*, Vol. 13, No. 3, pp. 531–577.
15. **Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2008).** The incremental information content of tone change in management discussion and analysis. *Available at SSRN: http://ssrn.com/abstract=1126962*.
16. **Ferris, S. P. et al. (2012).** The effect of issuer conservatism on ipo pricing and performance. *Review of Finance*, Vol. 17, No. 3, pp. 993–1027.
17. **Garrod, N., Pirkovic, S. R., & Valentincic, A. (2006).** Testing for discontinuity or type of distribution. *Mathematics and Computers in Simulation*, Vol. 71, No. 1, pp. 9–15.
18. **Goel, S. & Uzuner, O. (2016).** Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, Vol. 23, No. 3, pp. 215–239.
19. **Gupta, A., Simaan, M., & Zaki, M. J.**, **.** When positive sentiment is not so positive: Textual analytics and bank failures. *Available at SSRN: https://ssrn.com/abstract=2773939*.
20. **Hájek, P. & Olej, V. (2013).** Evaluating sentiment in annual reports for financial distress prediction using neural networks and support vector machines. *International Conference on Engineering Applications of Neural Networks*, Springer, pp. 1–10.
21. **Hajek, P., Olej, V., & Myskova, R. (2014).** Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, Vol. 20, No. 4, pp. 721–738.

22. **Hart, R. P. (2000).** Diction 5.0. *Austin, TX: Digitex*.
23. **Henry, E. (2008).** Are investors influenced by how earnings press releases are written? *Journal of Business Communication*, Vol. 45, No. 4, pp. 363–407.
24. **Hutagaol-Martowidjojo, Y. & Valentincic, A. (2015).** Valuation and forecasting roles of dividends of indonesian listed firms. *Journal of International Financial Management & Accounting*, Vol. 27, No. 2.
25. **Jegadeesh, N. & Wu, D. (2013).** Word power: A new approach for content analysis. *Journal of Financial Economics*, Vol. 110, No. 3, pp. 712–729.
26. **Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009).** Predicting risk from financial reports with regression. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pp. 272–280.
27. **Kosi, U. & Valentincic, A. (2013).** Write-offs and profitability in private firms: Disentangling the impact of tax-minimisation incentives. *European accounting review*, Vol. 22, No. 1, pp. 117–150.
28. **Kothari, S., Li, X., & Short, J. E. (2009).** The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review*, Vol. 84, No. 5, pp. 1639–1670.
29. **Lee, L. F., Hutton, A. P., & Shu, S. (2015).** The role of social media in the capital market: evidence from consumer product recalls. *Journal of Accounting Research*, Vol. 53, No. 2, pp. 367–404.
30. **Leung, S., Parker, L., & Courtis, J. (2015).** Impression management through minimal narrative disclosure in annual reports. *The British Accounting Review*, Vol. 47, No. 3, pp. 275–289.
31. **Li, F. (2008).** Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and economics*, Vol. 45, No. 2, pp. 221–247.
32. **Li, F. (2010).** The information content of forward-looking statements in corporate filings - a naïve bayesian machine learning approach. *Journal of Accounting Research*, Vol. 48, No. 5, pp. 1049–1102.
33. **Loughran, T. & McDonald, B. (2011).** Barron's red flags: Do they actually work? *Journal of Behavioral Finance*, Vol. 12, No. 2, pp. 90–97.
34. **Loughran, T. & McDonald, B. (2011).** When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, Vol. 66, No. 1, pp. 35–65.
35. **Loughran, T. & McDonald, B. (2014).** Measuring readability in financial disclosures. *The Journal of Finance*, Vol. 69, No. 4, pp. 1643–1671.
36. **Merkl-Davies, D. & Koller, V. (2012).** 'Metaphoring' people out of this world: A critical discourse analysis of a chairman's statement of a uk defence firm. *Accounting Forum*, Vol. 36, No. 3, pp. 178–193.
37. **Pennebaker, J. W. (2011).** *The secret life of pronouns: W hat our words say about us*. Bloomsbury Press, New York.
38. **Pimenta, A., Alexandre, P., El-Haj, M., Rayson, P., Walker, M., & Young, S., .** Heterogeneous narrative content in annual reports published as PDF files: Extraction, classification and incremental predictive ability. *Available at SSRN: https://ssrn.com/abstract=2803275*.
39. **Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012).** Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, Vol. 36, No. 4, pp. 992–1011.
40. **Qiu, X. Y., Srinivasan, P., & Street, N. (2006).** Exploring the forecasting potential of company annual reports. *Proceedings of the American Society for Information Science and Technology*, Vol. 43, No. 1, pp. 1–15.
41. **Rees, W. & Valentincic, A. (2013).** Dividend irrelevance and accounting models of value. *Journal of Business Finance & Accounting*, Vol. 40, No. 5-6, pp. 646–672.
42. **Rogers, J. L., Van Buskirk, A., & Zechman, S. L. (2011).** Disclosure tone and shareholder litigation. *The Accounting Review*, Vol. 86, No. 6, pp. 2155–2183.
43. **Roychowdhury, S. (2006).** Earnings management through real activities manipulation. *Journal of Accounting and Economics*, Vol. 42, pp. 335–370.
44. **SEC (2012).** Securities exchange act of 1934. *Securities Exchange Act of 1934*.
45. **Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966).** *The General Inquirer: A Computer Approach to Content Analysis.* MIT press.

Article received on 27/12/2016; accepted on 25/02/2017.
Corresponding author is Senja Pollak.