A Study of Importance of Textual Features for Predictive Models of Financial Indicators

Kralj Novak, Petra; Pollak, Senja; Valentinčič, Aljoša; Lončarski, Igor; Pahor, Marko and Žnidaršič, Martin

Abstract: ***In this study, we experimentally assess the potential of informal and unregulated communication to contribute to predictive models of financial markets indicators. The data sources that were analyzed are unregulated parts of yearly reports of the companies of the DOW30 index, text of tweets that mention these companies, data from financial statements, and stock market data about stock prices and volume. We conducted correlation analysis of descriptive and target features and an analysis of impacts of descriptive features to predictive power of models for regression and classification. The results indicate that overall the studied features only weakly describe the complex and noisy target phenomena and that also the linguistic features can contribute to phenomena models, particularly the features that represent expressions of sentiment, both in tweets and annual reports.***

Index Terms: ***yearly reports, sentiment analysis, informal communication, unregulated communication, linguistic features, correlation, impact on financial markets***

Manuscript received May 2019.

The authors acknowledge the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103) and the project Influence of formal and informal corporate communications on capital markets (No. J5-7387). The authors would like to thank Sowa Labs who enabled us to collect the information from tweets and kindly provided the pre-trained tweet sentiment models.

P. Kralj Novak (petra.kralj.novak@ijs.si), S. Pollak and M. Žnidaršič (martin.znidarsic@ijs.si) are with the Jožef Stefan Institute, Ljubljana, Slovenia. S. Pollak is also affiliated to the University of Edinburgh, Edinburgh, United Kingdom. A. Valentinčič, I. Lončarski and M. Pahor are with the University of Ljubljana, School of Economics and Business, Ljubljana, Slovenia.

1. Introduction

The main goal of financial reporting in the financial system is to ensure that high-quality, useful information about the financial position of firms, their performance and changes in their financial position is available to a wide range of users, including existing and potential investors, financial institutions, employees, the government. Formal reports contain both strictly regulated financial sections, as well as unregulated, narrative parts. While formal communications are commonly the subject of academic research, studies of unregulated narratives and informal communication are scarcer, but are becoming more common in recent years, as reflected also in surveys [4, 12, 8].

Our research starts from the hypothesis that informal communications contain useful information to capital markets, and that there is a relation between business performance and linguistic properties of unregulated parts of annual reports and informal communications, such as microblogging posts. Similar hypothesis have been studied in related work. For example, relations between microblogging posts (tweets) and financial indicators have been addressed in e.g. [3, 17, 15, 5, 19], while non-financial information from reports has been used for prediction of financially relevant events, such as next year performance [14], stock return volatility [10] or fraud detection [6]. This study is based on work [18] that served as a pilot study, in which we have analysed the correlation between linguistic and financial indicators in a limited case of four companies.

The goal of our research is to study the importance of informal communications and unregulated parts of annual reports for cap-

ital markets. We aim to detect correlations among the financial indicators and the financial and linguistic features of unregulated parts of annual reports and microblogging posts (that mention the studied companies). Furthermore, the study is aimed at the assessment of potential predictive capabilities of the three groups of descriptive features (financial and linguistic features from reports, and linguistic from microblogging posts) and the financial indicators as prediction targets. In this paper, we document the experiments with statistical and machine learning methods that were conducted in order to detect any such impact and the features that contribute to that. Besides the novelty of some of the studied features, the paper's main contribution is the combined analysis of very diverse feature sets and of their individual and combined effects.

For the purpose of our study, we retrieved yearly financial (10-K) reports of the thirty DOW30 companies from last 20 years. Additionally, we used a collection of 11,309,609 tweets containing DOW30 stock tickers over the period of four years. We also used financial markets data which consists of stock price, stock market volume and return index data for all the analysed companies in the time period of the last twenty years.

Our results indicate that overall the studied features only weakly describe the (complex and noisy) target phenomena and that also the linguistic features can contribute to phenomena models, particularly the features that represent expressions of sentiment, both in tweets and annual reports.

This paper is organised as follows: Section 2 describes the data acquisition, cleaning and feature generation process. Section 3 is dedicated to the experimental setting and experimental results. Section 4 provides the concluding remarks and directions for further work.

2. Data Description

The data used in the experiments corresponds to the companies of the DOW30 index. We used three sources of data:

- Text from the unregulated parts of annual reports
- Numerical yearly financial data from financial statements
- Stock market data (stock prices and volumes)
- Text from Twitter messages that mention these companies

Since we do not have a complete overlap of all the data sources, we derived two datasets: one that is based on data from annual reports which covers the period of 20 years and contains 552 examples. We denote it with $Data_{20}$. A smaller one, denoted $Data_4$, contains also the data from tweets, but is limited on the period of four years and contains 113 examples.

2.1 Annual Reports

We retrieved yearly financial (10-K) reports of the thirty DOW30 companies from last 20 years. The financial reports of the companies were collected with a script based on the SECEdgar tool[1].

The reports are usually provided in raw format, which includes HTML tags, tables, images in binary formats and other non-textual or unreadable contents. In addition, we are only interested in the unregulated narrative parts of reports. It was therefore necessary that the reports were filtered and cleaned before the analysis of text. For this purpose, we used the approach that is described in the paper by Smailović et al. [18] and the corresponding tools that we have developed. We cleaned and processed the document to extract sentiment and other linguistic features.

Several types of linguistic features were considered (cf. [18]): simple ones, sentiment, trust-and-doubt features, and features based on discourse markers (by Biber et al. [2]).

The simple features include the *length* of the documents (i.e. number of words) and the proportion of first person personal pronouns "I" and "we" compared to the impersonal pronoun "it" (*pers/it*) or compared to personal and impersonal pronouns together (*pers_persit*). The corresponding feature names are:

[1] `https://pypi.org/project/SECEdgar/`

```
length
pers/it
pers_persit
```

The measures of sentiment that we used were of two kinds: (I) relative frequencies of *positive* and *negative* terms (using Loughran-McDonald Master Dictionary [11]) and the ratio of positive terms compared to all positive and negative terms (*pos2posneg*), (II) frequencies of positive, negative and neutral sentences, and an indicator computed as their aggregate, named the *Sentiment score* [9] (the mean of the discrete probability distribution of the sentiment), which were computed with a hybrid sentiment detection algorithm [20], an adapted reimplementation of the work by Malo et al. [13]. The features computed with this model that was learned only on the data from the original paper [13] have the suffix *_SM_PMdata_only*, while those by a model that was learned on the original dataset extended with additional manually labeled items have the suffix *_SM*.

The features related to sentiment in yearly reports as named in the dataset are:

```
positive
negative
pos2posneg
negative_SM
neutral_SM
positive_SM
SentimentScore_SM
negative_SM_PMdata_only
neutral_SM_PMdata_only
positive_SM_PMdata_only
SentimentScore_Pmdata_only
```

Next, we used a dictionary of *trust* and *doubt*-related words [21], containing (near) synonyms of words related to "trust" and "doubt" from WordNet[2] and online dictionaries.The word lists contain 25 words for trust (e.g., trustful, confidence) and 77 for doubt (e.g., uncertainty, untrusting, suspicion). For each feature, we represent the *trust* and *doubt* values as relative frequencies of words from a word list with respect to the total length of a report (only the extracted parts). In addition, we also compute the number of trust terms in relation to all trust/doubt terms. The corresponding names of the features are:

[2] http://wordnet.princeton.edu

```
trust
doubt
trust2trustdoubt
```

The last group of features is based on discourse markers by Biber et al. [2] (listed in [1, 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).

The complete list of discoursive features as named in the dataset is below:

```
Ability_biber
AttituteEmotion_biber
CausationModalityEffort_biber
Certainty_biber
Cognition_biber
Communication_biber
DesireDecision_biber
EaseDifficulty_biber
Evaluation_biber
Likelihood_biber
ModalNecessity_biber
ModalPossiblity_biber
ModalPrediction
Nouns_various
PremodAdv_biber
Style_biber
```

2.2 Tweets

We processed 11,309,609 tweets containing DOW30 stock tickers over the period of four years: 1.1.2014 - 31.12.2017. As features, we used the number of tweets and aggregated sentiment of the tweets. On average, there were 1,800 tweets per week (about 260 per day). In the tweets, the most represented stock is APPL. The number of tweets per stock is depicted in Figure 1.

For sentiment classification, a classifier consisting of two SVM models was build to distinguish between the three (ordered) classes: One

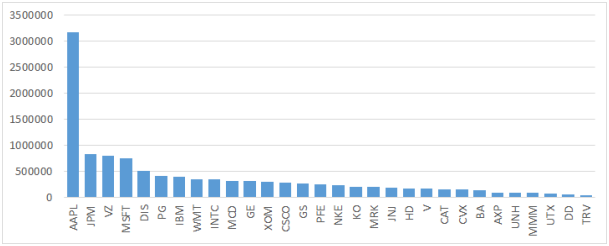


Figure 1: Total number of tweets per stock.

SVM model was trained to distinguish *Positive-or-Neutral* from the *Negative* tweets and another SVM model to distinguish *Positive* from *Neutral-or-Negative* tweets. For classification, both models are consulted and they vote for the final class assignment. The model structure is adapted from [7] and described in [16].

Once we have a sentiment label for each tweet, we aggregate the sentiment of the tweets for each stock for each selected period into a *Sentiment score*. We compute the Sentiment score as the mean of the discrete probability distribution of the sentiment (see [9] for more details). Sentiment score is in the range: $[-1, +1]$.

The features related to the tweets include both the aggregated sentiment in form of Sentiment score, and the count of tweets for the same periods. The periods duration range from short term (one day before date) to long term (one year before date). We have considered two relevant dates: financial year end (FYE) and filed as of date (FOD), which is the day of filing the annual report. These features are:

	Relative to FYE
count_y0m0d1	one day
sentiment_y0m0d1	one day
count_y0m0d7	one week
sentiment_y0m0d7	one week
count_y0m1d0	one month
sentiment_y0m1d0	one month
count_y0m3d0	three months
sentiment_y0m3d0	three months
count_y0m6d0	six months
sentiment_y0m6d0	six months
count_y1m0d0	one year
sentiment_y1m0d0	one year
	Relative to FOD
count_y0m0d1-fod	one day
sentiment_y0m0d1-fod	one day
count_y0m0d7-fod	one week
sentiment_y0m0d7-fod	one week
count_y0m1d0-fod	one month
sentiment_y0m1d0-fod	one month
count_y0m3d0-fod	three months
sentiment_y0m3d0-fod	three months
count_y0m6d0-fod	six months
sentiment_y0m6d0-fod	six months
count_y1m0d0-fod	one year
sentiment_y1m0d0-fod	one year

2.3 Financial Statements Data

The financial part of data contains numerical financial data from financial statements. With exception of *Total return*, most of this data is used as features, in particular these are:

```
BETA, MONTHLY, 5 YEARS
DS HISTORICAL BETA LOCAL INDEX
NET SALES OR REVENUES
COST OF GOODS SOLD (EXCL DEPRECIATION)
SALARIES AND BENEFITS EXPENSES
GROSS INCOME
SELLING, GENERAL & ADMINISTRATIVE EXPENSES
AMORTIZATION OF INTANGIBLES
AMORTIZATION OF DEFERRED CHARGES
DEPRECIATION, DEPLETION AND AMORTIZATION
RESEARCH & DEVELOPMENT
OPERATING INCOME
INTEREST EXPENSE ON DEBT
EXTRAORDINARY CREDIT - PRETAX
EXTRAORDINARY CHARGE - PRETAX
INTEREST CAPITALIZED
NON-OPERATING INTEREST INCOME
PRETAX INCOME
INCOME TAXES
MINORITY INTEREST
NET INCOME BEFORE EXTRA ITEMS/PREFERRED DIVIDENDS
EXTRA ITEMS & GAIN/LOSS SALE OF ASSETS
PREFERRED DIVIDEND REQUIREMENTS
NET INCOME USED TO CALCULATE DILUTED EARNINGS PER SHARE
NET INCOME USED TO CALCULATE BASIC EARNINGS PER SHARE
NET INCOME AVAILABLE TO COMMON
CASH & SHORT TERM INVESTMENTS
RECEIVABLES(NET)
INVENTORIES - TOTAL
CURRENT ASSETS - TOTAL
PROPERTY, PLANT AND EQUIPMENT - NET
TOTAL INTANGIBLE OTHER ASSETS - NET
TOTAL ASSETS
ACCOUNTS PAYABLE
SHORT TERM DEBT & CURRENT PORTION OF LONG TERM DEBT
CURRENT LIABILITIES - TOTAL
WORKING CAPITAL
LONG TERM DEBT
TOTAL DEBT
DEFERRED TAXES
MINORITY INTEREST
PREFERRED STOCK
COMMON STOCK
COMMON EQUITY
TOTAL CAPITAL
AMORTIZATION OF INTANGIBLE ASSETS
FUNDS FROM OPERATIONS
NET PROCEEDS FROM SALE/ISSUE OF COMMON & PREFERRED
CASH DIVIDENDS PAID - TOTAL
CAPITAL EXPENDITURES (ADDITIONS TO FIXED ASSETS)
INCREASE/DECREASE IN CASH AND SHORT TERM INVESTMENTS
NET CASH FLOW - OPERATING ACTIVITIES
NET CASH FLOW - INVESTING
NET CASH FLOW - FINANCING
COMMON DIVIDENDS (CASH)
EMPLOYEES
QUICK RATIO
```

```
CURRENT RATIO
INVENTORIES - DAYS HELD
ACCOUNTS RECEIVABLES DAYS
TOTAL DEBT \% TOTAL CAPITAL & SHORT TERM DEBT
TOTAL DEBT \% COMMON EQUITY
RETURN ON EQUITY - TOTAL (\%)
CASH FLOW/SALES
OPERATING PROFIT MARGIN
PRETAX MARGIN
NET MARGIN
RETURN ON INVESTED CAPITAL
CURRENT BETA
EARNINGS BEFORE INTEREST AND TAXES (EBIT)
DIVIDENDS PROVIDED FOR OR PAID - COMMON
EBIT & DEPRECIATION
PRICE EOY
```

The value not used as feature, but as a potential target attribute was TOTAL RETURN - MARKET (S&P500).

2.4 Financial Markets Data

This part of data consists of stock closing price, stock market volume and return index data for all the analysed companies in the time period of the last twenty years. This data was used for calculation of stock market indicators which we used as target classes in our correlations and predictions. For each of the three kinds of stock market data (price, volume and return index) we calculated six indicators for each date d of filling a yearly report in our database:

- relative change from day $d-1$ to d
- relative change from day d to $d+1$
- relative change from day $d-1$ to $d+1$
- relative change from day $d-3$ to d
- relative change from day d to $d+3$
- relative change from day $d-3$ to $d+3$

The indicator names start with a prefix of the type of data: CRDC, VRDC and RIDC for closing price, volume and return index respectively, followed by the span of the relative change. For example, VRDC_td_from_td-3 denotes the relative change of stock volume on day d from day $d-3$.

3. Experimental Setting and Experiments

The goal of our experiments is to empirically study the correlation and predictive power of different sets of features on the capital markets. In the following sections, we present the (I) correlations among descriptive and target variables, (II) assessment of importance of individual features, and (III) analysis of impacts of feature sets on the predictive power of machine-learned models.

3.1 Correlation Study

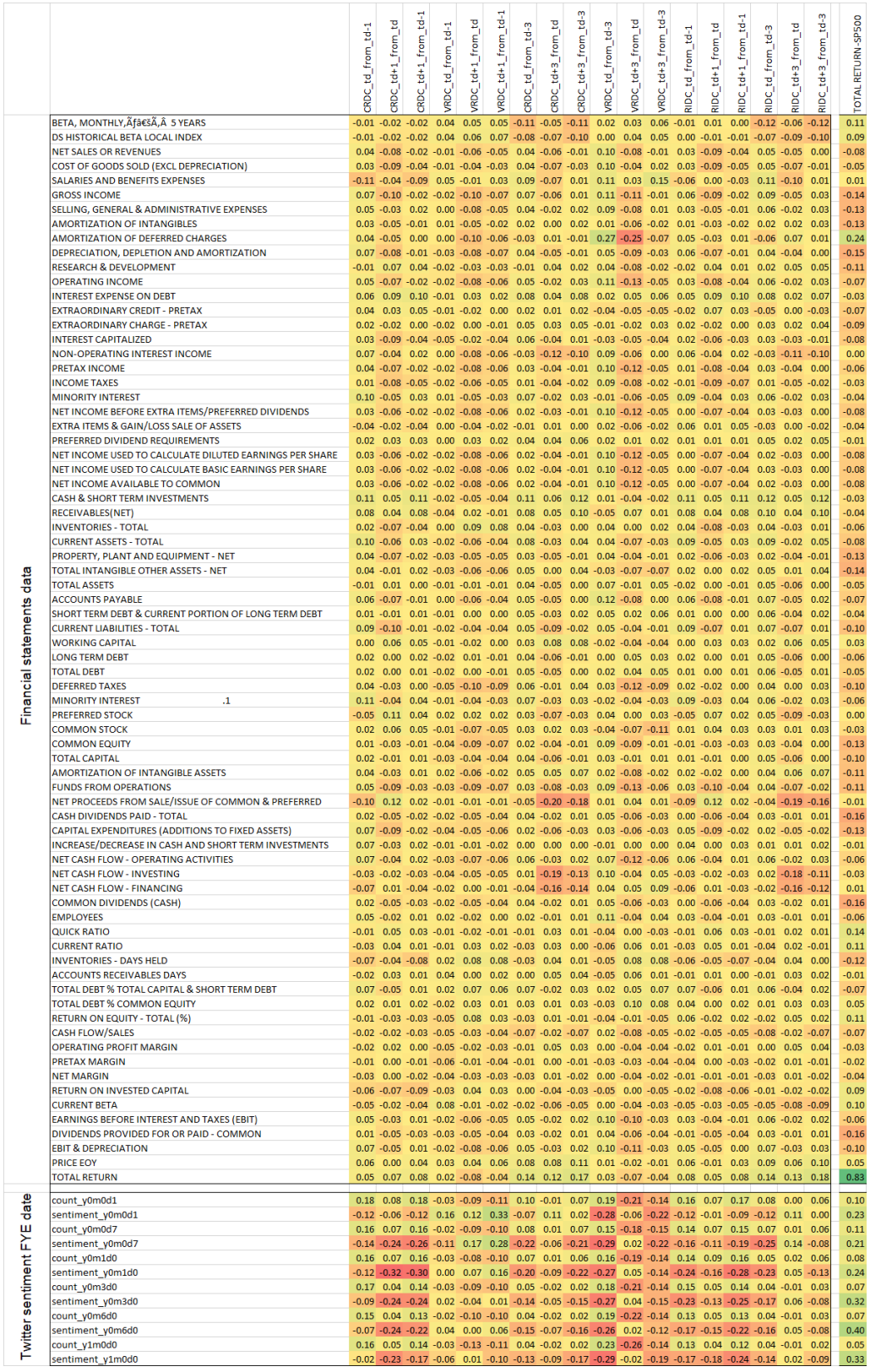
Correlation (Pearson) among all the used features and target variables was computed and is presented in Figures 2 and 3. The correlation coefficients were calculated on $Data_{20}$ for all features from annual reports and on $Data_4$ for the Twitter related features. The strength of correlations is color-coded: red for negative, green for positive, and yellow for no correlation. While most of the correlations are close to zero (Figures 2 and 3 are predominantly yellow), there are some deviations worth mentioning, especially among the Twitter features.

Figure 2 depicts the Pearson correlation coefficients between the linguistic and Twitter (FOD) features w.r.t. the FOD date targets and the market adjusted total return. Considering the complex phenomenon we are describing, the correlation coefficients are relatively high ($\geq$ 0.2) between the target (TOTAL RETURN - SP500) and longer term sentiment aggregates for one year (0.31), six months (0.33), three months (0.29) and one month (0.2). A pattern of positive correlation with the number of tweets and a negative correlation with the sentiment in them can be observed for the target VRDC_td_from_td-3. The tweet counts are also slightly ($\leq$ -0.2) negatively correlated with VRDC_td+3_from_td.

Similarly, Figure 3 depicts the Pearson correlation coefficients between the features from financial statements and Twitter (FYE) features. The most colorful is again the Twitter sentiment part, where the correlation coefficient between 6-months aggregated sentiment and adjusted total return is 0.4. Slight negative (or no) correlation of FYE sentiment features can be observed for most of the relative targets, except for VRDC_td+3_from_td which is negatively (cca -0.2) correlated with the number of tweets.

		CRDC_td_from_td-1	CRDC_td+1_from_td	CRDC_td+1_from_td-1	VRDC_td_from_td-1	VRDC_td+1_from_td	VRDC_td+1_from_td-1	CRDC_td_from_td-3	CRDC_td+3_from_td	CRDC_td+3_from_td-3	VRDC_td_from_td-3	VRDC_td+3_from_td	VRDC_td+3_from_td-3	RIDC_td_from_td-1	RIDC_td+1_from_td	RIDC_td+1_from_td-1	RIDC_td_from_td-3	RIDC_td+3_from_td	RIDC_td+3_from_td-3	TOTAL RETURN-SP500
All linguistic features	negative_SM	-0.05	0.06	0.01	-0.02	0.01	0.01	0.02	0.07	0.06	-0.02	0.01	-0.03	-0.05	0.06	0.01	0.03	0.05	0.06	0.04
	neutral_SM	-0.08	0.00	-0.05	-0.02	-0.01	-0.01	-0.01	0.04	0.02	0.02	-0.02	-0.03	-0.08	0.01	-0.05	-0.01	0.03	0.01	0.05
	positive_SM	-0.04	0.06	0.01	0.00	0.02	0.02	0.01	0.07	0.06	-0.03	0.03	-0.03	-0.04	0.05	0.01	0.02	0.07	0.05	0.07
	SentimentScore_SM	-0.03	0.10	0.05	0.00	0.04	0.04	0.06	0.04	0.07	-0.04	0.02	-0.04	-0.01	0.09	0.06	0.09	0.01	0.07	0.03
	negative_SM_PMdata_only	-0.06	0.06	0.00	-0.02	0.01	0.00	0.01	0.06	0.05	-0.02	0.00	-0.04	-0.06	0.06	0.00	0.02	0.05	0.04	0.04
	neutral_SM_PMdata_only	-0.07	0.03	-0.02	-0.02	0.00	0.00	0.00	0.05	0.04	0.01	-0.01	-0.03	-0.07	0.03	-0.03	0.01	0.04	0.03	0.05
	positive_SM_PMdata_only	-0.07	0.02	-0.03	-0.01	0.00	0.00	0.00	0.05	0.04	0.00	0.00	-0.02	-0.06	0.02	-0.03	0.01	0.04	0.04	0.04
	SentimentScore_Pmdata_only	0.04	0.05	0.06	-0.01	0.04	0.01	0.01	-0.01	0.00	-0.06	0.01	-0.06	0.00	0.05	0.04	-0.02	-0.02	-0.02	0.02
	length	-0.09	0.02	-0.04	-0.01	-0.01	0.00	-0.01	0.05	0.03	0.03	-0.03	-0.03	-0.09	0.03	-0.04	0.00	0.04	0.03	0.05
	pers_persit	0.05	0.01	0.04	-0.09	0.01	-0.05	0.06	0.04	0.07	-0.09	0.02	-0.07	0.04	0.01	0.04	0.09	0.05	0.09	-0.05
	pers/it	0.03	0.02	0.04	-0.09	0.04	-0.03	0.02	0.04	0.04	-0.08	0.01	-0.06	0.02	0.00	0.02	0.03	0.04	0.05	-0.03
	trust	-0.01	-0.01	-0.01	0.00	-0.02	0.00	-0.03	-0.06	-0.06	0.05	-0.08	0.01	0.00	-0.02	-0.01	-0.02	-0.06	-0.05	0.03
	doubt	-0.04	0.02	-0.01	0.05	0.01	0.06	0.04	0.07	0.08	-0.06	0.00	-0.05	-0.02	0.02	0.00	0.06	0.04	0.07	0.08
	trust2trustdoubt	0.04	-0.02	0.01	-0.01	-0.02	-0.02	-0.03	-0.11	-0.09	0.06	-0.06	0.04	0.03	-0.03	0.00	-0.04	-0.09	-0.09	-0.04
	positive	0.04	0.02	0.04	0.09	0.07	0.08	0.01	0.03	0.03	0.00	0.00	0.00	0.03	0.01	0.02	0.01	0.02	0.02	0.02
	negative	-0.05	0.06	0.01	0.02	-0.01	0.02	0.05	-0.01	0.03	0.04	-0.01	0.02	-0.05	0.04	-0.01	0.04	-0.03	0.01	0.05
	pos2posneg	0.05	-0.07	-0.02	0.02	0.05	0.03	-0.05	-0.03	-0.06	-0.03	0.02	0.00	0.04	-0.05	-0.01	-0.03	-0.01	-0.03	-0.03
	Ability_biber	0.01	0.09	0.07	-0.01	0.01	0.03	-0.01	0.06	0.03	-0.03	-0.06	-0.07	0.00	0.05	0.04	0.01	0.04	0.03	0.02
	AttituteEmotion_biber	0.01	0.05	0.04	-0.03	0.05	-0.01	0.05	0.05	0.07	-0.07	0.02	-0.03	0.00	0.05	0.03	0.04	0.02	0.04	0.06
	CausationModalityEffort_biber	0.01	0.03	0.03	-0.04	0.00	-0.03	0.05	0.01	0.04	-0.03	-0.02	-0.03	-0.02	0.04	0.01	0.05	0.00	0.04	0.04
	Certainty_biber	-0.04	0.05	0.00	0.10	0.03	0.08	-0.03	0.05	0.02	0.04	-0.02	0.00	-0.04	0.02	-0.01	-0.01	0.04	0.02	0.01
	Cognition_biber	0.01	0.01	0.02	0.03	0.11	0.07	0.01	0.06	0.05	-0.08	0.06	-0.02	0.03	0.01	0.03	0.03	0.05	0.05	0.03
	Communication_biber	0.04	-0.05	-0.01	0.03	-0.01	0.01	0.05	-0.02	0.02	0.00	0.02	0.06	0.04	-0.04	0.00	0.05	-0.01	0.03	-0.03
	DesireDecision_biber	0.01	0.07	0.05	0.00	0.01	0.00	0.07	0.07	0.10	-0.07	-0.04	-0.08	0.00	0.06	0.04	0.07	0.08	0.10	0.04
	EaseDifficulty_biber	-0.09	-0.04	-0.09	0.01	-0.07	-0.04	-0.03	-0.08	-0.07	0.04	-0.01	0.04	-0.09	-0.04	-0.08	-0.01	-0.08	-0.06	-0.03
	Evaluation_biber	0.01	-0.05	-0.03	0.02	-0.03	-0.02	0.00	0.00	0.00	0.08	-0.07	-0.01	-0.03	-0.09	-0.08	-0.06	-0.01	-0.05	-0.07
	Likelihood_biber	-0.03	0.06	0.02	0.01	0.08	0.04	0.02	0.04	0.05	-0.08	0.09	-0.01	-0.01	0.03	0.01	0.04	0.03	0.05	0.05
	ModalNecessity_biber	-0.05	0.03	-0.01	-0.01	0.01	0.01	-0.01	0.05	0.03	-0.04	-0.01	-0.04	-0.03	0.02	-0.01	0.03	0.05	0.05	0.03
	ModalPossiblity_biber	0.00	0.06	0.05	-0.01	0.00	0.00	0.04	0.06	0.07	0.01	-0.09	-0.07	-0.01	0.03	0.01	0.05	0.03	0.05	0.02
	ModalPrediction	-0.09	0.07	-0.01	0.02	0.01	0.01	-0.08	0.06	-0.01	0.02	-0.04	0.00	-0.08	0.04	-0.03	-0.08	0.03	-0.04	0.02
	Nouns_various	-0.04	0.11	0.05	0.08	0.02	0.07	0.05	0.08	0.09	0.03	-0.05	-0.02	-0.04	0.09	0.03	0.06	0.04	0.07	-0.01
	PremodAdv_biber	0.05	-0.01	0.03	0.07	0.01	0.04	0.07	0.03	0.07	0.00	0.04	0.06	0.02	-0.03	-0.01	0.04	0.02	0.04	0.01
	Style_biber	0.00	0.07	0.05	-0.02	0.08	0.05	0.02	0.04	0.04	-0.06	0.04	-0.06	0.03	0.06	0.05	0.06	0.03	0.06	-0.01
Twitter sentiment FOD date	count_y0m0d1-fod	0.17	0.07	0.17	0.04	-0.10	-0.06	0.02	0.03	0.04	0.22	-0.27	-0.12	0.18	0.08	0.18	0.02	0.05	0.05	0.08
	sentiment_y0m0d1-fod	0.13	-0.06	0.04	0.03	0.02	0.04	0.23	-0.04	0.15	0.06	0.09	0.14	0.12	-0.04	0.06	0.22	-0.01	0.15	0.00
	count_y0m0d7-fod	0.09	0.07	0.11	0.02	-0.09	-0.08	0.04	0.00	0.03	0.20	-0.23	-0.13	0.09	0.08	0.12	0.03	0.03	0.05	0.11
	sentiment_y0m0d7-fod	0.08	-0.10	-0.02	0.00	-0.04	-0.05	0.18	-0.14	0.04	-0.03	0.03	0.01	0.07	-0.10	-0.03	0.16	-0.12	0.03	0.04
	count_y0m1d0-fod	0.05	0.03	0.06	0.02	-0.09	-0.06	0.00	-0.02	-0.01	0.24	-0.24	-0.11	0.05	0.04	0.07	-0.01	0.01	0.00	0.11
	sentiment_y0m1d0-fod	-0.01	-0.03	-0.03	-0.05	-0.05	-0.09	-0.03	0.01	-0.02	-0.22	0.01	-0.10	-0.01	-0.02	-0.02	-0.04	0.03	-0.01	0.20
	count_y0m3d0-fod	0.09	0.05	0.10	0.02	-0.09	-0.07	0.02	-0.01	0.00	0.23	-0.23	-0.10	0.10	0.08	0.13	0.00	0.03	0.03	0.07
	sentiment_y0m3d0-fod	-0.05	-0.12	-0.12	-0.03	0.00	0.01	-0.09	-0.01	-0.07	-0.29	0.05	-0.12	-0.15	-0.08	-0.16	-0.12	0.07	-0.03	0.29
	count_y0m6d0-fod	0.10	0.04	0.10	0.01	-0.08	-0.07	0.02	-0.03	-0.01	0.23	-0.23	-0.10	0.10	0.05	0.10	0.01	-0.01	0.00	0.07
	sentiment_y0m6d0-fod	-0.07	-0.14	-0.15	-0.02	-0.01	0.00	-0.12	0.00	-0.10	-0.32	0.07	-0.11	-0.17	-0.11	-0.19	-0.14	0.07	-0.05	0.33
	count_y1m0d0-fod	0.12	0.04	0.11	0.02	-0.09	-0.07	0.02	-0.03	-0.01	0.25	-0.25	-0.10	0.11	0.04	0.10	0.02	-0.03	-0.01	0.05
	sentiment_y1m0d0-fod	-0.07	-0.18	-0.17	-0.04	0.01	0.03	-0.10	-0.07	-0.13	-0.33	0.09	-0.08	-0.12	-0.11	-0.16	-0.11	0.04	-0.05	0.31

Figure 2: Pearson correlation of linguistic and Twitter FOD (filed as of date) features w.r.t. the FOD date targets and the market adjusted total return. The strength of correlations is color-coded: red for negative, green for positive, and yellow for no correlation.



		CRDC_td_from_td-1	CRDC_td+1_from_td	CRDC_td+1_from_td-1	VRDC_td_from_td-1	VRDC_td+1_from_td	VRDC_td+1_from_td-1	CRDC_td_from_td-3	CRDC_td+3_from_td	CRDC_td+3_from_td-3	VRDC_td_from_td-3	VRDC_td+3_from_td	VRDC_td+3_from_td-3	RIDC_td_from_td-1	RIDC_td+1_from_td	RIDC_td+1_from_td-1	RIDC_td_from_td-3	RIDC_td+3_from_td	RIDC_td+3_from_td-3	TOTAL RETURN-SP500
Financial statements data	BETA, MONTHLY,Ãƒâ€šÃ‚Â 5 YEARS	-0.01	-0.02	-0.02	0.04	0.05	0.05	-0.11	-0.05	-0.11	0.02	0.03	0.06	-0.01	0.01	0.00	-0.12	-0.06	-0.12	0.11
	DS HISTORICAL BETA LOCAL INDEX	-0.01	-0.02	-0.02	0.04	0.06	0.07	-0.08	-0.07	-0.10	0.00	0.04	0.05	0.00	-0.01	-0.01	-0.07	-0.09	-0.10	0.09
	NET SALES OR REVENUES	0.04	-0.08	-0.02	-0.01	-0.06	-0.05	0.04	-0.06	-0.01	0.10	-0.08	-0.01	0.03	-0.09	-0.04	0.05	-0.05	0.00	-0.08
	COST OF GOODS SOLD (EXCL DEPRECIATION)	0.03	-0.09	-0.04	-0.01	-0.04	-0.03	0.04	-0.07	-0.03	0.10	-0.04	0.02	0.03	-0.09	-0.05	0.05	-0.07	-0.01	-0.05
	SALARIES AND BENEFITS EXPENSES	-0.11	-0.04	-0.09	0.05	-0.01	0.03	0.09	-0.07	0.01	0.11	0.03	0.15	-0.06	0.00	-0.03	0.11	-0.10	0.01	0.01
	GROSS INCOME	0.07	-0.10	-0.02	-0.02	-0.10	-0.07	0.07	-0.06	0.01	0.11	-0.11	-0.01	0.06	-0.09	-0.02	0.09	-0.05	0.03	-0.14
	SELLING, GENERAL & ADMINISTRATIVE EXPENSES	0.05	-0.03	0.02	0.00	-0.08	-0.05	0.04	-0.02	0.02	0.09	-0.08	0.01	0.03	-0.05	-0.01	0.06	-0.02	0.03	-0.13
	AMORTIZATION OF INTANGIBLES	0.03	-0.05	-0.01	0.01	-0.05	-0.02	0.02	0.00	0.02	0.01	-0.06	-0.02	0.01	-0.03	-0.02	0.02	0.02	0.03	-0.13
	AMORTIZATION OF DEFERRED CHARGES	0.04	-0.05	0.00	0.00	-0.10	-0.06	-0.03	0.01	-0.01	0.27	-0.25	-0.07	0.05	-0.03	0.01	-0.06	0.07	0.01	0.24
	DEPRECIATION, DEPLETION AND AMORTIZATION	0.07	-0.08	-0.01	-0.03	-0.08	-0.07	0.04	-0.05	-0.01	0.05	-0.09	-0.03	0.06	-0.07	-0.01	0.04	-0.04	0.00	-0.15
	RESEARCH & DEVELOPMENT	-0.01	0.07	0.04	-0.02	-0.03	-0.03	-0.01	0.04	0.02	0.04	-0.08	-0.02	-0.02	0.04	0.01	0.02	0.05	0.05	-0.11
	OPERATING INCOME	0.05	-0.07	-0.02	-0.02	-0.08	-0.06	0.05	-0.02	0.03	0.11	-0.13	-0.05	0.03	-0.08	-0.04	0.06	-0.02	0.03	-0.07
	INTEREST EXPENSE ON DEBT	0.06	0.09	0.10	-0.01	0.03	0.02	0.08	0.04	0.08	0.02	0.05	0.06	0.05	0.09	0.10	0.08	0.02	0.07	-0.03
	EXTRAORDINARY CREDIT - PRETAX	0.04	0.03	0.05	-0.01	-0.02	0.00	0.02	0.01	0.02	-0.04	-0.05	-0.05	-0.02	0.07	0.03	-0.05	0.00	-0.03	-0.07
	EXTRAORDINARY CHARGE - PRETAX	0.02	-0.02	0.00	-0.02	0.00	-0.01	0.05	0.03	0.05	-0.01	-0.02	0.03	0.02	-0.02	0.00	0.03	0.02	0.04	-0.09
	INTEREST CAPITALIZED	0.03	-0.09	-0.04	-0.05	-0.02	-0.04	0.06	-0.04	0.01	-0.03	-0.05	-0.04	0.02	-0.06	-0.03	0.03	-0.03	-0.01	-0.08
	NON-OPERATING INTEREST INCOME	0.07	-0.04	0.02	0.00	-0.08	-0.06	-0.03	-0.12	-0.10	0.09	-0.06	0.00	0.06	-0.04	0.02	-0.03	-0.11	-0.10	0.00
	PRETAX INCOME	0.04	-0.07	-0.02	-0.02	-0.08	-0.06	0.03	-0.04	-0.01	0.10	-0.12	-0.05	0.01	-0.08	-0.04	0.03	-0.04	0.00	-0.06
	INCOME TAXES	0.01	-0.08	-0.05	-0.02	-0.06	-0.05	0.01	-0.04	-0.02	0.09	-0.08	-0.02	-0.01	-0.09	-0.07	0.01	-0.05	-0.02	-0.03
	MINORITY INTEREST	0.10	-0.05	0.03	0.01	-0.05	-0.03	0.07	-0.02	0.03	-0.01	-0.06	-0.05	0.09	-0.04	0.03	0.06	-0.02	0.03	-0.04
	NET INCOME BEFORE EXTRA ITEMS/PREFERRED DIVIDENDS	0.03	-0.06	-0.02	-0.02	-0.08	-0.06	0.02	-0.03	-0.01	0.10	-0.12	-0.05	0.00	-0.07	-0.04	0.03	-0.03	0.00	-0.08
	EXTRA ITEMS & GAIN/LOSS SALE OF ASSETS	-0.04	-0.02	-0.04	0.00	-0.04	-0.02	-0.01	0.01	0.00	0.02	-0.06	-0.02	0.06	0.01	0.05	-0.03	0.00	-0.02	-0.04
	PREFERRED DIVIDEND REQUIREMENTS	0.02	0.03	0.03	0.00	0.03	0.02	0.04	0.04	0.06	0.02	0.01	0.02	0.01	0.01	0.01	0.05	0.02	0.05	-0.01
	NET INCOME USED TO CALCULATE DILUTED EARNINGS PER SHARE	0.03	-0.06	-0.02	-0.02	-0.08	-0.06	0.02	-0.04	-0.01	0.10	-0.12	-0.05	0.00	-0.07	-0.04	0.02	-0.03	0.00	-0.08
	NET INCOME USED TO CALCULATE BASIC EARNINGS PER SHARE	0.03	-0.06	-0.02	-0.02	-0.08	-0.06	0.02	-0.04	-0.01	0.10	-0.12	-0.05	0.00	-0.07	-0.04	0.03	-0.03	0.00	-0.08
	NET INCOME AVAILABLE TO COMMON	0.03	-0.06	-0.02	-0.02	-0.08	-0.06	0.02	-0.04	-0.01	0.10	-0.12	-0.05	0.00	-0.07	-0.04	0.02	-0.03	0.00	-0.08
	CASH & SHORT TERM INVESTMENTS	0.11	0.05	0.11	-0.02	-0.05	-0.04	0.11	0.06	0.12	0.01	-0.04	-0.02	0.11	0.05	0.11	0.12	0.05	0.12	-0.03
	RECEIVABLES(NET)	0.08	0.04	0.08	-0.04	0.02	-0.01	0.08	0.05	0.10	-0.05	0.07	0.01	0.08	0.04	0.08	0.10	0.04	0.10	-0.04
	INVENTORIES - TOTAL	0.02	-0.07	-0.04	0.00	0.09	0.08	0.04	-0.03	0.00	0.04	0.00	0.02	0.04	-0.08	-0.03	0.04	-0.03	0.01	-0.06
	CURRENT ASSETS - TOTAL	0.10	-0.06	0.03	-0.02	-0.06	-0.04	0.08	-0.03	0.04	0.04	-0.07	-0.03	0.09	-0.05	0.03	0.09	-0.02	0.05	-0.08
	PROPERTY, PLANT AND EQUIPMENT - NET	0.04	-0.07	-0.02	-0.03	-0.05	-0.05	0.03	-0.05	-0.01	0.04	-0.04	-0.01	0.02	-0.06	-0.03	0.02	-0.04	-0.01	-0.13
	TOTAL INTANGIBLE OTHER ASSETS - NET	0.04	-0.01	0.02	-0.03	-0.06	-0.06	0.05	0.00	0.04	-0.03	-0.07	-0.07	0.02	0.00	0.02	0.05	0.01	0.04	-0.14
	TOTAL ASSETS	-0.01	0.01	0.00	-0.01	-0.01	-0.01	0.04	-0.05	0.00	0.07	-0.01	0.05	-0.02	0.00	-0.01	0.05	-0.06	0.00	-0.05
	ACCOUNTS PAYABLE	0.06	-0.07	-0.01	0.00	-0.06	-0.04	0.05	-0.05	0.00	0.12	-0.08	0.00	0.06	-0.08	-0.01	0.07	-0.05	0.02	-0.07
	SHORT TERM DEBT & CURRENT PORTION OF LONG TERM DEBT	0.01	-0.01	0.01	-0.01	0.00	0.00	0.05	-0.03	0.02	0.05	0.02	0.06	0.01	0.00	0.00	0.06	-0.04	0.02	-0.04
	CURRENT LIABILITIES - TOTAL	0.09	-0.10	-0.01	-0.02	-0.04	-0.04	0.05	-0.09	-0.02	0.05	-0.04	-0.01	0.09	-0.07	0.01	0.07	-0.07	0.01	-0.10
	WORKING CAPITAL	0.00	0.06	0.05	-0.01	-0.02	0.00	0.03	0.08	0.08	-0.02	-0.04	-0.04	0.00	0.03	0.03	0.02	0.06	0.05	0.03
	LONG TERM DEBT	0.02	0.00	0.02	-0.02	0.01	-0.01	0.04	-0.06	-0.01	0.00	0.05	0.03	0.02	0.00	0.01	0.05	-0.06	0.00	-0.06
	TOTAL DEBT	0.02	0.00	0.01	-0.02	0.00	-0.01	0.05	-0.05	0.00	0.02	0.04	0.05	0.01	0.00	0.01	0.06	-0.05	0.01	-0.05
	DEFERRED TAXES	0.04	-0.03	0.00	-0.05	-0.10	-0.09	0.06	-0.01	0.04	0.03	-0.12	-0.09	0.02	-0.02	0.00	0.04	0.00	0.03	-0.10
	MINORITY INTEREST .1	0.11	-0.04	0.04	-0.01	-0.04	-0.03	0.07	-0.03	0.03	-0.02	-0.04	-0.03	0.09	-0.03	0.04	0.06	-0.02	0.03	-0.06
	PREFERRED STOCK	-0.05	0.11	0.04	0.02	0.02	0.02	0.03	-0.07	-0.03	0.04	0.00	0.03	-0.05	0.07	0.02	0.05	-0.09	-0.03	0.00
	COMMON STOCK	0.02	0.06	0.05	-0.01	-0.07	-0.05	0.03	0.02	0.03	-0.04	-0.07	-0.11	0.01	0.04	0.03	0.03	0.01	0.03	-0.03
	COMMON EQUITY	0.01	-0.03	-0.01	-0.04	-0.09	-0.07	0.02	-0.04	-0.01	0.09	-0.09	-0.01	-0.01	-0.03	-0.03	0.03	-0.04	0.00	-0.13
	TOTAL CAPITAL	0.02	-0.01	0.01	-0.03	-0.04	-0.04	0.04	-0.06	-0.01	0.03	-0.01	0.01	0.01	-0.01	0.00	0.05	-0.06	0.00	-0.10
	AMORTIZATION OF INTANGIBLE ASSETS	0.04	-0.03	0.01	0.02	-0.06	-0.02	0.05	0.05	0.07	0.02	-0.08	-0.02	0.02	-0.02	0.00	0.04	0.06	0.07	-0.11
	FUNDS FROM OPERATIONS	0.05	-0.09	-0.03	-0.03	-0.09	-0.07	0.03	-0.08	-0.03	0.09	-0.13	-0.06	0.03	-0.10	-0.04	0.04	-0.07	-0.02	-0.11
	NET PROCEEDS FROM SALE/ISSUE OF COMMON & PREFERRED	-0.10	0.12	0.02	-0.01	-0.01	-0.01	-0.05	-0.20	-0.18	0.01	0.04	0.01	-0.09	0.12	0.02	-0.04	-0.19	-0.16	-0.01
	CASH DIVIDENDS PAID - TOTAL	0.02	-0.05	-0.02	-0.02	-0.05	-0.04	0.04	-0.02	0.01	0.05	-0.06	-0.03	0.00	-0.06	-0.04	0.03	-0.01	0.01	-0.16
	CAPITAL EXPENDITURES (ADDITIONS TO FIXED ASSETS)	0.07	-0.09	-0.02	-0.04	-0.05	-0.06	0.02	-0.06	-0.03	0.03	-0.06	-0.03	0.05	-0.09	-0.02	0.02	-0.05	-0.02	-0.13
	INCREASE/DECREASE IN CASH AND SHORT TERM INVESTMENTS	0.07	-0.03	0.02	-0.01	-0.01	-0.02	0.00	0.00	0.00	-0.01	0.00	0.00	0.04	0.00	0.03	0.01	0.01	0.02	-0.01
	NET CASH FLOW - OPERATING ACTIVITIES	0.07	-0.04	0.02	-0.03	-0.07	-0.06	0.06	-0.03	0.02	0.07	-0.12	-0.06	0.06	-0.04	0.01	0.06	-0.02	0.03	-0.06
	NET CASH FLOW - INVESTING	-0.03	-0.02	-0.03	-0.04	-0.05	-0.05	0.01	-0.19	-0.13	0.10	-0.04	0.05	-0.03	-0.02	-0.03	0.02	-0.18	-0.11	-0.03
	NET CASH FLOW - FINANCING	-0.07	0.01	-0.04	-0.02	0.00	-0.01	-0.04	-0.16	-0.14	0.04	0.05	0.09	-0.06	0.01	-0.03	-0.02	-0.16	-0.12	0.01
	COMMON DIVIDENDS (CASH)	0.02	-0.05	-0.03	-0.02	-0.05	-0.04	0.04	-0.02	0.01	0.05	-0.06	-0.03	0.00	-0.06	-0.04	0.03	-0.02	0.01	-0.16
	EMPLOYEES	0.05	-0.02	0.01	0.02	-0.02	0.00	0.02	-0.01	0.01	0.11	-0.04	0.04	0.03	-0.04	-0.01	0.03	-0.01	0.01	-0.06
	QUICK RATIO	-0.01	0.05	0.03	-0.01	-0.02	-0.01	-0.01	0.03	0.01	-0.04	0.00	-0.03	-0.01	0.06	0.03	-0.01	0.02	0.01	0.14
	CURRENT RATIO	-0.03	0.04	0.01	-0.01	0.03	0.02	-0.03	0.03	0.00	-0.06	0.06	0.01	-0.03	0.05	0.01	-0.04	0.02	-0.01	0.11
	INVENTORIES - DAYS HELD	-0.07	-0.04	-0.08	0.02	0.08	0.08	-0.03	0.04	0.01	-0.05	0.08	0.08	-0.06	-0.05	-0.07	-0.04	0.04	0.00	-0.12
	ACCOUNTS RECEIVABLES DAYS	-0.02	0.03	0.01	0.04	0.00	0.02	0.00	0.05	0.04	-0.05	0.06	0.01	-0.01	0.01	0.00	-0.01	0.03	0.02	-0.01
	TOTAL DEBT % TOTAL CAPITAL & SHORT TERM DEBT	0.07	-0.05	0.01	0.02	0.07	0.06	0.07	-0.02	0.03	0.02	0.05	0.07	0.07	-0.06	0.01	0.06	-0.04	0.02	-0.07
	TOTAL DEBT % COMMON EQUITY	0.02	0.01	0.02	-0.02	0.03	0.01	0.03	0.01	0.03	-0.03	0.10	0.08	0.04	0.00	0.02	0.01	0.03	0.03	0.05
	RETURN ON EQUITY - TOTAL (%)	-0.01	-0.03	-0.03	-0.05	0.08	0.03	-0.03	0.01	-0.01	-0.04	-0.01	-0.05	0.06	-0.02	0.02	-0.02	0.05	0.02	0.11
	CASH FLOW/SALES	-0.02	-0.02	-0.03	-0.05	-0.03	-0.04	-0.07	-0.02	-0.07	0.02	-0.08	-0.05	-0.02	-0.05	-0.05	-0.08	-0.02	-0.07	-0.07
	OPERATING PROFIT MARGIN	-0.02	0.02	0.00	-0.05	-0.02	-0.03	-0.01	0.05	0.03	0.00	-0.04	-0.04	-0.02	0.01	-0.01	0.00	0.05	0.04	-0.03
	PRETAX MARGIN	-0.01	0.00	-0.01	-0.06	-0.01	-0.04	-0.01	0.00	-0.01	-0.03	-0.03	-0.04	-0.04	0.00	-0.03	-0.02	0.01	-0.01	-0.02
	NET MARGIN	-0.03	0.00	-0.02	-0.04	-0.03	-0.03	-0.03	0.01	-0.02	0.00	-0.04	-0.02	-0.01	-0.01	-0.01	-0.03	0.01	-0.02	-0.04
	RETURN ON INVESTED CAPITAL	-0.06	-0.07	-0.09	-0.03	0.04	0.03	0.00	-0.04	-0.03	-0.05	0.00	-0.05	-0.02	-0.08	-0.06	-0.01	-0.02	-0.02	0.09
	CURRENT BETA	-0.05	-0.02	-0.04	0.08	-0.01	-0.02	-0.02	-0.06	-0.05	0.00	-0.04	-0.03	-0.05	-0.03	-0.05	-0.05	-0.08	-0.09	0.10
	EARNINGS BEFORE INTEREST AND TAXES (EBIT)	0.05	-0.03	0.01	-0.02	-0.06	-0.05	0.05	-0.02	0.02	0.10	-0.10	-0.03	0.03	-0.04	-0.01	0.06	-0.02	0.02	-0.06
	DIVIDENDS PROVIDED FOR OR PAID - COMMON	0.01	-0.05	-0.03	-0.03	-0.05	-0.04	0.03	-0.02	0.01	0.04	-0.06	-0.04	-0.01	-0.05	-0.04	0.03	-0.01	0.01	-0.16
	EBIT & DEPRECIATION	0.07	-0.05	0.01	-0.02	-0.08	-0.06	0.05	-0.03	0.02	0.10	-0.11	-0.03	0.05	-0.05	0.00	0.07	-0.03	0.03	-0.10
	PRICE EOY	0.06	0.00	0.04	0.03	0.04	0.06	0.08	0.08	0.11	0.01	-0.02	-0.01	0.06	-0.01	0.03	0.09	0.06	0.10	0.05
	TOTAL RETURN	0.05	0.07	0.08	0.02	-0.08	-0.04	0.14	0.12	0.17	0.03	-0.07	-0.04	0.08	0.05	0.08	0.14	0.13	0.18	0.83
Twitter sentiment FYE date	count_y0m0d1	0.18	0.08	0.18	-0.03	-0.09	-0.11	0.10	-0.01	0.07	0.19	-0.21	-0.14	0.16	0.07	0.17	0.08	0.00	0.06	0.10
	sentiment_y0m0d1	-0.12	-0.06	-0.12	0.16	0.12	0.33	-0.07	0.11	0.02	-0.28	-0.06	-0.22	-0.12	-0.01	-0.09	-0.12	0.11	0.00	0.23
	count_y0m0d7	0.16	0.07	0.16	-0.02	-0.09	-0.10	0.08	0.01	0.07	0.15	-0.18	-0.15	0.14	0.07	0.15	0.07	0.01	0.06	0.11
	sentiment_y0m0d7	-0.14	-0.24	-0.26	-0.11	0.17	0.28	-0.22	-0.06	-0.21	-0.29	0.02	-0.22	-0.16	-0.11	-0.19	-0.25	0.14	-0.08	0.21
	count_y0m1d0	0.16	0.07	0.16	-0.03	-0.08	-0.10	0.07	0.01	0.06	0.16	-0.19	-0.14	0.14	0.09	0.16	0.05	0.02	0.06	0.08
	sentiment_y0m1d0	-0.12	-0.32	-0.30	0.00	0.07	0.16	-0.20	-0.09	-0.22	-0.27	0.05	-0.14	-0.24	-0.16	-0.28	-0.23	0.05	-0.13	0.24
	count_y0m3d0	0.17	0.04	0.14	-0.03	-0.09	-0.10	0.05	-0.02	0.02	0.18	-0.21	-0.14	0.15	0.05	0.14	0.04	-0.01	0.03	0.07
	sentiment_y0m3d0	-0.09	-0.24	-0.24	0.02	-0.04	0.01	-0.14	-0.05	-0.15	-0.27	0.04	-0.15	-0.23	-0.13	-0.25	-0.17	0.06	-0.08	0.32
	count_y0m6d0	0.15	0.04	0.13	-0.02	-0.10	-0.10	0.04	-0.02	0.02	0.19	-0.22	-0.14	0.13	0.05	0.13	0.04	-0.01	0.03	0.07
	sentiment_y0m6d0	-0.07	-0.24	-0.22	0.04	0.00	0.06	-0.15	-0.07	-0.16	-0.26	0.02	-0.12	-0.17	-0.15	-0.22	-0.16	0.05	-0.08	0.40
	count_y1m0d0	0.16	0.05	0.14	-0.03	-0.13	-0.11	0.04	-0.02	0.02	0.23	-0.26	-0.14	0.13	0.04	0.12	0.04	-0.01	0.02	0.05
	sentiment_y1m0d0	-0.02	-0.23	-0.17	-0.06	0.01	-0.10	-0.13	-0.09	-0.17	-0.29	-0.02	-0.19	-0.17	-0.18	-0.24	-0.14	0.02	-0.09	0.33

Figure 3: Pearson correlation of financial features and Twitter (FYE) features w.r.t. the FOD date targets and the market adjusted total return. The strength of correlations is color-coded: red for negative, green for positive, and yellow for no correlation.

3.2 Individual Feature Importance

Importance of individual features was calculated on the $Data_4$ dataset with the *Feature importance with forests of trees* approach in scikit-learn. To get an overall assessment of individual features, we have collected the feature importance rankings for each target variable and calculated the average rank of each feature. The ranking of features varies with each target class, but the linguistic features are commonly represented among the most important features, primarily the ones that are related to sentiment assessment, either in annual reports or tweets.

The twenty best ranked features according to this approach are:

```
16.22 : INCREASE/DECREASE IN CASH AND SHORT TERM INVESTMENTS
19.28 : sentiment_y0m0d7-fod
20.33 : sentiment_y0m0d1-fod
22.56 : DS HISTORICAL BETA LOCAL INDEX
22.89 : SentimentScore_Pmdata_only
26.06 : BETA, MONTHLY,Ã‚Â  5 YEARS
26.44 : sentiment_y1m0d0-fod
29.06 : NET CASH FLOW - FINANCING
29.28 : Cognition_biber
29.72 : TOTAL DEBT % TOTAL CAPITAL & SHORT TERM DEBT
29.83 : sentiment_y0m1d0-fod
30.06 : ModalNecessity_biber
30.28 : Evaluation_biber
30.56 : ModalPrediction
31.56 : Ability_biber
34.11 : sentiment_y0m6d0-fod
34.78 : SentimentScore_SM
34.94 : count_y1m0d0-fod
36.00 : sentiment_y0m3d0-fod
36.22 : negative_SM_PMdata_only
```

3.3 Predictive Modeling

We perform predictive modeling of financial indicators to assess the potential predicitve capabilities of our three groups of descriptive features. We considered these three groups of features:

- financial features from reports,
- linguistic features from reports,
- linguistic features and counts from tweets

as well as their union (combined feature set). More specifically, our targets were relative changes in closing price, volume and return index, described in Section 2.4. In one set of experiments, we have modeled the predictive task as a regression prediction problem. In the second set of experiments, we have discretized our target and tried to predict the binary classification problem: target variable increase vs. decrease. Both sets of experiments were performed on both datasets: $Data_{20}$ and $Data_4$.

The scikit-learn Python package was used to perform the predictive performance analysis. We used leave-one-out cross validation to estimate the predictive performance of the predictive models.

Regression

In the regression problem setting, we used mean absolute error, mean squared error and R^2 (coefficient of determination) to measure the performance. We use the following modelling algorithms:

- DummyRegressor: This model predicts the mean of the target variable of the training set. This regressor is useful as a simple baseline to compare with other regressors.
- LinearRegression: Ordinary least squares Linear Regression.
- DecisionTreeRegressor: A decision tree regressor with parameters (max_depth=5, min_samples_split=10, min_samples_leaf=10).
- SVR: A support vector machine approach that is adapted for regression. With RBF kernel and parameters C=1.0 and ϵ=0.1
- Random forest regressor: A random forest for regression, parameters: (max_depth=10 n_estimators=100).
- GradientBoostingRegressor: A gradient boosting approach for regression. Parameters: (n_estimators=10, learning_rate=0.1, max_depth=3)

We had 18 target variables and evaluated (leave-one-out) the seven machine learning regression models on each of the four sets of features on both datasets. The learned models rarely beat the simple overall average approach (DummyRegressor). With all of the targets, only the SVR model managed to beat the baseline with marginally better results, but rarely on

all the considered measures at once. Our conclusion from these experiments is that the regression problem formulation is too demanding for modeling of such noisy phenomena with the given data.

Classification

In case of classification, performance was measured by classification accuracy and the macroaveraged F1 measure on the models built by the following models:

- Dummy: The model that always predicts the majority class.
- Nearest Neighbors: The k-nearest neighbors approach (k=3).
- RBF_SVM: Support vector machine classifier with RBF kernel and parameters: C(gamma=2, C=1).
- Decision_Tree: A decision tree classifier (max_depth=5).
- Random_Forest: = A random forest classifier with parameters: (max_depth=5, n_estimators=10, max_features=1)
- AdaBoost: The AdaBoost classifier.
- Naive_Bayes = Naive Bayes classifier.

We discretized the targets from the regression setting into two classes: target variable increase and decrease. The targets had a slightly unbalanced class distribution (72% majority class in the worst case). We evaluated (leave-one-out) the seven machine learning classification models on each of the 18 targets on both datasets with different sets of features. We used the results of the Dummy classifier as a baseline.

On $Data_{20}$, 54 experiments (three sets of features on 18 target variables) with seven classifiers each were performed. On $Data_4$, 72 experiments (four sets of features on 18 target variables) with seven classifiers each were performed. In classification, the learned models improve on the F1 measure of the *majority class* baseline and in several cases improve on its classification accuracy as well. This indicates that the target concept might be learnable to some extent from the given data.

Table 1: Number of experiments with respect to the effect of addition of linguistic features to the financial ones. Only the cases when combined feature set beat the baseline are considered.

	$Data_{20}$	$Data_4$
better	15	7
equal	10	9
worse	8	2

Out of 126 experiments with seven classifiers each, we here focus on those that have learnable targets. We consider targets learnable whenever at least one of the classifiers learned on the combined feature set beats the baseline. This happened 33 times on the $Data_{20}$ dataset and 18 times on the $Data_4$ dataset. To get an insight into the question whether the textual information adds any new information to the financial one, we have counted the number of experiments in which a classifier that is trained on all features achieves a better classification accuracy than a classifier trained only on financial ones. This was done only for the cases considered learnable.

Overall, the linguistic features seem to contribute to predictive performance of classifiers in 22 cases, reduce the performance in 10 and do not have an impact in 19 cases. Details are provided in Table 1.

The results show that all the feature sets (linguistic, tweets-related and financial) potentially contain information to be learned about the target concepts. However, as there are no consistent relations among the influences of the feature sets, any interpretations of these results must be cautious.

4. Conclusion

The study described in this paper was aimed at the assessment of relations among various financial and linguistic data features and the indicators of capital markets, such as relative changes in stock closing prices and market adjusted total return. The linguistic features that were used originate from two sources of data:

unregulated textual parts of yearly reports and tweets that were mentioning a particular company. There were numerous features created from these data items, also some relatively new ones, such as sentiment scores of novel domain specific sentiment analysis approaches. All three feature sets (financial data based, annual report text based and tweet based) were used in experiments with an array of target indicators.

The results of our experiments show that overall the studied features only weakly describe the target phenomena. This was mostly expected, as these phenomena (e.g., stock price changes) are complex, noisy and depend also on a number of other factors, some of which are difficult or impossible to assess. However, the results of experiments with various feature sets indicate that all the feature sets have the potential to contribute to phenomena models, also the linguistic ones. According to results of the assessment of individual feature importance, among the linguistic ones, particularly the features that represent expressions of sentiment seem to be relevant in this aspect.

References

[1] Douglas Biber. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*, volume 28. John Benjamins Publishing, 2007.

[2] Douglas Biber, Edward Finegan, Stig Johansson, Susan Conrad, and Geoffrey Leech. *Longman Grammar of Spoken and Written English*. Longman, 1999.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[4] Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. Natural language processing in accounting, auditing and finance: a synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214, 2016.

[5] Peter Gabrovšek, Darko Aleksovski, Igor Mozetič, and Miha Grčar. Twitter sentiment around the earnings announcement events. *PLoS ONE*, 12(2):e0173151, 2017.

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

[7] Miha Grčar, Darko Cherepnalkoski, Igor Mozetič, and Petra Kralj Novak. Stance and influence of twitter users regarding the Brexit referendum. *Computational social networks*, 4(1):6, 2017.

[8] Colm Kearney and Sha Liu. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185, 2014.

[9] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.

[10] Feng Li. The information content of forward-looking statements in corporate filings - a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5):1049–1102, 2010.

[11] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

[12] Tim Loughran and Bill McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016.

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

[14] Xin Ying Qiu, Padmini Srinivasan, and Nick Street. Exploring the forecasting potential of company annual reports. *Proceedings of the American Society for Information Science and Technology*, 43(1):1–15, 2006.

[15] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič. The effects of Twitter sentiment on stock price returns. *PLoS ONE*, 10(9):e0138441, 2015.

[16] Sašo Rutar. Empirična evalvacija procesa avtomatske klasifikacije sentimenta na finančni domeni, 2016.

[17] Jasmina Smailović, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, 285:181–203, 2014.

[18] Jasmina Smailović, Martin Žnidaršič, Aljoša Valentinčič, Igor Lončarski, Marko Pahor, Pedro Tiago Martins, and Senja Pollak. Automatic analysis of annual financial reports: A case study. *Computación y Sistemas*, 21(4):809–818, 2017.

[19] Timm O Sprenger, Andranik Tumasjan, Philipp G Sandner, and Isabell M Welpe. Tweets and trades: the information content of stock microblogs. *European Financial Management*, 20(5):926–957, 2014.

[20] Jan Štihec, Martin Žnidaršič, and Senja Pollak. Simplified hybrid approach for detection of semantic orientations in economic texts. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, pages 692–698, 2018.

[21] Martin Žnidaršič, Jasmina Smailović, Jan Gorše, Miha Grčar, Igor Mozetič, and Senja Pollak. Trust and doubt terms in financial tweets and periodic reports. In Mahmoud El-Haj, Paul Rayson, and Andrew Moore, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA).

Petra Kralj Novak received her doctoral degree in Information Communication Technologies from the Jožef Stefan Inernational Postgraduate School in 2009. Currently, she is with the Department of Knowledge Technologies at Jožef Stefan Institute, Ljubljana, Slovenia. Her research focus is on the analysis of social and mainstream media focusing on the mediated sentiment and stance. She also researched the role of emojis in conveying sentiment.

Senja Pollak is a member of the Department of Knowledge Technologies, Jožef Stefan Institute, currently a research fellow at the Usher institute, University of Edinburgh. Her fields of research cover corpus linguistics and natural language processing, with the focus on author profiling and terminology extraction. She is the coordinator of the H2020 project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). She was the PI of the industrial project in automatization of terminology management for the largest Slovenian translation company, and collaborated in several national and EU research projects. She was invited speaker of Translation, Technology, Terminology conferences in 2014 and 2015, is organizing co-chair of SLSP 2019, and has been serving as a program and organizing committee member of several international conferences and workshops (e.g., ICCC, ECIR, 4REAL).

Aljoša Valentinčič is a professor of accounting and finance at the Faculty of Economics, University of Ljubljana. He received his PhD at the University of Glasgow, UK. He has been an active member of the European Accounting Association, including chairing the Standing Scientific Committee. His research is focused on financial reporting processes of private firms, payout policies of public firms and is currently working on a series of papers connecting accounting and finance to neuroscience, cognitive science and psychology. He is a member of editorial boards of and reviews for accounting journals and is an academic editor at PLOS One.

Igor Lončarski received his doctoral degree in Finance from Tilburg University in the Netherlands in 2007. He is currently an associate professor of Finance at the University of Ljubljana, School of Economics and Business, Ljubljana, Slovenia. His current research is focused on the use of multifactor models in corporate valuation, impact of market sentiment on sovereign credit ratings, and ethics in finance. In terms of professional service, Igor is the editor-in-chief of Risk Management journal (Springer), an associate editor of Emerging Markets Review and International Review of Financial Analysis, as well as a member of editorial boards of several other journals.

Marko Pahor received the PhD in Economics from the University of Ljubljana after spending an extended period at the University of Groningen, The Netherlands. He teaches courses in applied statistics and research methods at all three Bologna levels. His main research interests are in application of novel approaches in advanced research methods and data analysis, including social network analysis, agent based modeling and non-parametric statistics, to problems in business and economics. His interests involve also applications of data analytics, in particular text mining, as well as the economic and financial aspects of crypto revolution. He publishes with co-authors in a wide area of business and economics, including tourism economics, marketing and finance.

Martin Žnidaršič received his doctoral degree in Computer and Information Science from the University of Ljubljana in 2007. He is currently employed at the Department of Knowledge Technologies at Jožef Stefan Institute, Ljubljana, Slovenia. His research is focused on decision support and data mining, in particular on human assessment modelling and sentiment analysis. Applications of his work are conducted mostly in domains of ecology and finance.