Predictive performance comparisons of different feature extraction methods in a financial column corpus

Confronto della capacità predittiva di diversi metodi di estrazione delle variabili dal corpus di una rubrica finanziaria

Andrea Sciandra and Riccardo Ferretti

Abstract This work concerns the processing of a corpus made up of a financial weekly column. Specifically, we focused on document-level index extraction and textual feature extraction. Moreover, some feature extraction methods had been compared to evaluate their predictive capacity. Results confirm the hypothesis that vectors derived from word embedding do not improve the predictive power compared to other feature extraction methods but remain a fundamental resource for capturing semantics in texts.

Abstract *Questo contributo riguarda il trattamento di un corpus costituito da una rubrica finanziaria settimanale. In particolare, ci siamo concentrati sull'estrazione di indici a livello di documento e sull'estrazione di variabili testuali. Inoltre, abbiamo confrontato alcuni metodi di estrazione delle variabili per valutare la loro capacità predittiva. I risultati confermano l'ipotesi che i vettori derivati dal word embedding non migliorano la capacità predittiva rispetto ad altri metodi di estrazione delle variabili, ma restano una risorsa fondamentale per cogliere la semantica nei testi.*

Key words: behavioural finance, sentiment analysis, lexical complexity, feature extraction, word embedding, principal component regression

---

[1] Andrea Sciandra, Department of Communication and Economics, University of Modena and Reggio Emilia; email: andrea.sciandra@unimore.it

Riccardo Ferretti, Department of Communication and Economics, University of Modena and Reggio Emilia; email: riccardo.ferretti@unimore.it

This research has been partially supported by the University fund for Research (FAR 2020) of the University of Modena and Reggio Emilia.

1 Introduction

This work focuses on a financial column, named ‘Letter to investor’ (*Lettera all’investitore*), which has been published on the Sunday edition of the leading Italian financial newspaper (Il Sole 24 Ore). The column analyses every week an Italian stock through second-hand news, reporting balance sheet and income statement data, managers’ outlook, stock’s past performance, and, in some cases, analysts’ recommendations. In previous research, we showed how the mechanism of attention grabbing (AGH) is at work. According to AGH, stale information published in print media can lead retail investors to buy stocks that grab their attention [1] to the extent that past analysts’ recommendations may induce abnormal movements in stock prices and returns. Cervellati et al. [2] and Ferretti and Sciandra [3] showed that the publication of articles concerning single listed companies’ profiles and financial analysts’ recommendations are followed by an asymmetric reaction of stock prices. More precisely, they find a statistically significant stock price increase when the recommendation is positive (overweight or buy) and a substantial stationarity when the recommendation is not positive (hold or underweight or sell). In a more recent work, Ferretti and Sciandra [4] pointed out how the absence of explicit recommendations (approximately from 2015) in the same column, calls into question the role of the article sentiment, as they showed how investors transform articles content into implicit recommendations that, when highly positive, can direct their buying decisions. This result explained the importance of the textual analysis of this corpus, which will be deepened in this work in terms of text processing, textual feature extraction, and summary indexes especially related to the polarity and to the lexical complexity of the articles. Moreover, some feature extraction methods will be compared to evaluate their predictive capacity. The underlying hypothesis, based on previous research from other fields (especially social media [12]), is that vectors derived from word embedding do not improve the predictive ability compared to other feature extraction methods. The prediction targets are the abnormal returns calculated on the first day after the publication of the column.

2 Data processing

We collected all the ‘Letter to investor’ columns published, from January 2005 to December 2020, mentioning single Italian companies listed on the domestic Stock Exchange. In the time span 2005-2014 most of the columns contain explicit trading advice, that disappear since 2015 (overall: 350 stocks with explicit recommendation, and 366 without). Therefore, the 'Letter to investor' corpus consists of 716 articles, with an average length of about 1500 words, totalling 1104925 tokens and 482735 types. The type-token ratio is therefore very high (0.437), primarily due to the presence of: proper names (managers, companies, banks, rating agencies, countries), numbers and shares, dates, acronyms, anglicisms, etc.

The first task performed on the corpus was the lemmatisation using the R `udpipe` library [13]. The treebank on which this procedure was based, the Italian Stanford Dependency Treebank (ISDT), seems quite suitable for the purpose, as it was created using newspaper articles and Wikipedia pages. Following, we chose to select only nouns, adjectives, verbs, and adverbs for the next phase of feature extraction.

We then calculated, using a bag-of-words approach, some stylistic features pertaining to the lexical complexity and some lexica-based features pertaining the polarity of the texts. With regard to sentiment analysis, in previous works we exploited the NRC general lexicon, pointing out the need for resources in Italian comparable to the financial lexicon of Loughran and McDonald [6]. Loughran and McDonald lexicon contains lists of positive and negative terms, and other potentially interesting lists of words, e.g., related to uncertainty. Therefore, we decided to automatically translate the lexicon via 'eTranslation', an online machine translation service provided by the European Commission[2], qualitatively reviewing the result.

We also computed some lexical complexity measures, regarding readability and lexical diversity. The purpose of this task was to provide further dimensions that could potentially affect the reader and consequently the abnormal returns. In particular, these indices aim to discriminate the articles complexity and the authors' style of writing, as some journalists have taken turns as editors of the column over the years. For the predictive models, among several metrics we selected two indices of readability (mean sentence length, mean word syllables) and two indices of lexical diversity (Dugast's Uber Index U, Simpson's D) [14]. Since readability indices often contain specific weights for a given language, in this work we chose two unweighted indices[3]. Instead, lexical diversity is generally measured with respect to the type-token ratio. Considering the high level of correlation found between the several available indices, we chose U and D because we already tested them [5] and, even though they are dependent on the text length, weekly column's structure and layout did not vary in the observed period [4].

2.1 Feature extraction

The main goal of the feature extraction phase is to obtain a limited set of variables from the texts of the column, which will be used as predictors of abnormal returns. We chose to compare three different strategies: using the frequencies of a set of words determined by the value of the RAKE (Rapid Automatic Keyword Extraction) index, selecting the most important words based on the TF-IDF index, and creating a set of vectors derived from word embeddings.

RAKE [10] index derived from a keyword extraction algorithm based on the ratio of the degree to the frequency of each word. The algorithm creates a word

[2] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

[3] In future studies it would be useful to exploit Italian based indices, such as READ.IT and GulpEase.

degree matrix with each row displaying the number of times a given word co-occurs with another word in the sentences that make up a document. The degree of a word is calculated as the sum of the number of co-occurrences, then it is divided by the occurrence frequency. In this way, a ranking of the most relevant words in a text can be performed.

TF-IDF [11] is a widely used index that evaluates how relevant a word is to a document in a corpus and it is based on the ratio between the term frequency and the inverse document frequency of the given term. TF-IDF brings out the words that occur many times in a few documents and those words would be relevant to distinguish documents. We decide to compute the numerator as the normalized term frequency, i.e., the relative term frequency within a document. In order to obtain a selection of the most relevant terms (with a minimum frequency of 5), we then summed the TF-IDF normalised values within each document, thus obtaining a ranking for the terms. In this case, the selected predictors will be weighted precisely according to the TF-IDF index with respect to each document.

Word embedding is a popular text representation where words that have the same meaning will have a similar vector representation [7]. We chose to train our word embedding with the column corpus using the Global Vector (GloVe) model [9], usually able to identify synonyms or to suggest a word to complete a sentence. The GloVe model use an unsupervised learning algorithm to map the words into a N-dimensional space, where the semantic similarity among words is explored through the distance among the words. GloVe model builds a words co-occurrence matrix and then uses the matrix factorization technique for word embedding. Since word embedding techniques use context to create the word representations, after the corpus lemmatisation and the vocabulary creation (with a minimum frequency of 5), we defined a context window (a string of words before and after a focal word) of 3 words that was used to train our word embedding model. After obtaining 100 vectors for each word included in the corpus vocabulary, the word embedding features for each column were computed as the averages of the word vectors for all the vocabulary words appearing in the column [8]. A few examples of the semantic power of the word embedding trained in the corpus are provided in Table 1.

Table 1: Examples of similar words (cosine similarity) extracted from trained word embedding.

Input word	***Word 1 (similarity)***	***Word 2 (similarity)***	***Word 3 (similarity)***
strategy	development (0.751)	company (0.733)	growth (0.731)
plan	foresees (0.761)	industrial (0.723)	investment (0.671)
business	activity (0.905)	group (0.778)	industry (0.741)

3 Experiments and results

To compare the predictive power of 100 features obtained using RAKE, TF-IDF and word embedding respectively, we tested different statistical learning models (Linear Regression, Partial Least Squares Regression (PLS), Principal Component Regression (PCR), Random Forest, and Support Vector Machines with Radial Basis Function Kernel (SVM)) estimating the value of abnormal returns. Abnormal returns (ARs) were computed following the Market Adjusted Model:

$$AR_{jt} = R_{jt} - R_{mt}$$

where R_{jt} is the stock return of company j (mentioned in the column) on day t, R_{mt} is the stock market return (MILAN COMIT GLOBAL + R - PRICE INDEX) on day t, and AR_{jt} is the abnormal return of company j on day t (AR_{jt} are averaged across companies to get the mean Abnormal Return on day t, AR_t).

The experiments setting was the same for comparison purposes. The values of the ARs were then estimated through each of the five statistical models using 100 features selected through RAKE, TF-IDF and word embedding. To the 100 features of each model, we added five econometric control variables collected from DataStream and Borsa Italiana databases (company's size, price-to-book value (PBV), past performance, company's beta, and presence of concurrent news), four sentiment scores (NRC sentiment; Loughran-McDonald sentiment, uncertainty, and modal words scores) and four lexical complexity indices. Hence, a total of 114 predictors are included in each model. We performed a 5-fold cross-validation with 100 repetitions on the training set. The training set was made up of stocks with recommendations, while the test set was made up of stocks without recommendations. We obtained the best results through Principal Component Regression in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE)[4]. Table 2 shows the results of the models on the test set, while Figure 1 shows the most important features in the PCR-RAKE model based on weighted sums of the absolute regression coefficients[5]. It is important to stress that we found among the most important features: words extracted from RAKE, indices of sentiment and uncertainty, econometrics, and lexical complexity (Fig. 1). In contrast, in models with TF-IDF and word embedding few non-textual variables appeared among the most important ones. Furthermore, it should be mentioned that using terms in the models also allows for greater interpretability, which is simply not possible using word embeddings.

[4] Given a large set of variables, PCR probably overcomes the multicollinearity issue better than other techniques.

[5] Partial Least Squares regression also showed good results, especially through features extracted by word embedding and TF-IDF.

Table 2: Models results (MAE and RMSE) for each set of features – test set.

Model	***Features type***	***MAE***	***RMSE***
Linear Regression	RAKE	0.89281880	1,13017700
	TF-IDF	0.82722900	1.05565800
	Word Embedding	0.88001680	1.10798300
PLS	RAKE	0.11087380	0.12742180
	TF-IDF	0.03618955	0.04379134
	Word Embedding	0.03621648	0.04382179
PCR	RAKE	0.02606201	0.03253241
	TF-IDF	0.03017980	0.03764856
	Word Embedding	0.03017943	0.03764809
Random Forest	RAKE	0,07227743	0,09321358
	TF-IDF	0,04359055	0,05658084
	Word Embedding	0,06354684	0,07858456
SVM	RAKE	0,07826125	0,09849686
	TF-IDF	0,10417250	0,12097520
	Word Embedding	0,07064870	0,09008151

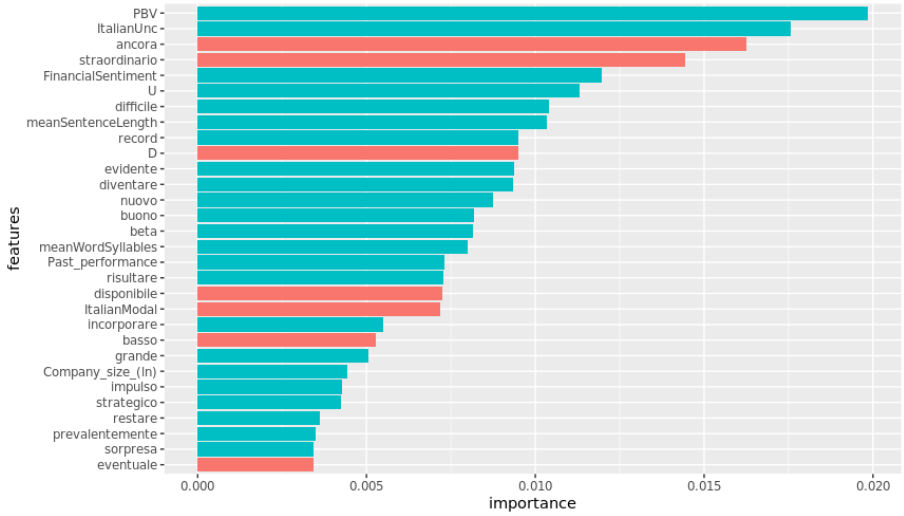

Figure 1: Most important features - RAKE PCR model (blue bars indicate a positive effect, red bars a negative effect).

4 Conclusions

Results confirmed our hypothesis, as RAKE features performed better in terms of both MAE and RMSE in the PCR model. The PCR model result for the word embedding features is similar to that achieved with TF-IDF. The main reason in our

opinion is that word embedding defines the multidimensional coordinates of each word, but to extract features for each text, we have to average each coordinate among the document words, resulting in fuzzy measures. We believe that the main usefulness of word embedding is in the recovery of semantics, while its use as features should be reviewed, for example through universal dependencies [5]. A further possibility to explore for exploiting word embeddings could be the use of measures like RAKE and TF-IDF to weight differently the numerical vectors. Future developments of this research should also consider n-grams and improve the translation of the financial lexicon for sentiment.

References

1. Barber, B.M., Odean, T.: All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors. The Rev. of Financial Stud., 21,785–818 (2008).
2. Cervellati, E.M., Ferretti, R., Pattitoni, P.: Market reaction to second-hand news: Inside the attention-grabbing hypothesis. Appl. Econ, 46(10), 1108-1121 (2014).
3. Ferretti, R., Sciandra, A.: Does the attention-grabbing mechanism work on Sundays? Influence of social and religious factors on investors' attention. Rev. of Behav. Fin. (2021)
4. Ferretti, R., Sciandra, A.: Media and Investors' Attention. Estimating analysts' ratings and sentiment of a financial column to predict abnormal returns. In: SIS 2021 Book of Short Papers, Pearson, 1543-1548 (2021).
5. Lai, M., Cignarella, A.T., Finos, L., Sciandra, A.: WordUp! at VaxxStance 2021: Combining Contextual Information with Textual and Dependency-Based Syntactic Features for Stance Detection. In: XXXVII Int. Conf. of the Spanish Society for NLP, 2943, 210-232. CEUR (2021).
6. Loughran, T. and McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The J. of Finance, 66(1), 35-65 (2011).
7. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient estimation of word representations in vector space. Proc. of International Conf. on Learning Representations (2013).
8. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Transactions on Internet Technology (TOIT), 17(3), 1-23. (2017).
9. Pennington J., Socher R., Manning C.D.: GloVe: Global Vectors for Word Representation, in Empirical Methods in Natural Language Processing (EMNLP), 1532-1543 (2014).
10. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text mining: applications and theory, 1, 1-20 (2010).
11. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523 (1988).
12. Sciandra, A.: COVID-19 Outbreak through Tweeters' Words: Monitoring Italian Social Media Communication about COVID-19 with Text Mining and Word Embeddings, 2020 IEEE Symposium on Computers and Communications (ISCC), 1-6 (2020).
13. Straka M., Straková J.: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, in: Proc. of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. ACL, Vancouver, Canada, 88-99 (2017).
14. Tweedie, F.J., Baayen, R.H.: How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities, 32(5), 323-352 (1998).