The weight of words: textual data versus sentiment analysis in stock returns prediction

Il peso delle parole: dati testuali versus sentiment analysis per predire i rendimenti azionari

Riccardo Ferretti and Andrea Sciandra

Abstract The focus of this paper is to understand whether the words contained in a text corpus improves the explained variance of stock returns better than the use of the polarity of the same texts, obtained through a sentiment analysis using a generic ontological dictionary. The empirical analysis is based on the content of a weekly column in the most important Italian financial newspaper, which published past information and analysts' recommendations on listed companies. The use of textual data clearly increases the explained variance of stock returns but, through comparisons between data mining techniques, we observed minor differences in terms of MSE, by adding a selection of specific terms as features. In this context, the text mining approach proved to be very useful to improve the explanatory power of forecasting models, while it emerged the limited explanatory power of an automatic sentiment analysis based on a generic lexicon.

Abstract *Il focus di questo contributo è capire se le parole contenute in un corpus di testi migliorano la varianza spiegata dei rendimenti azionari rispetto all'uso della polarità degli stessi testi, ottenuta mediante una sentiment analysis utilizzando un dizionario ontologico generico. L'analisi empirica si basa sul contenuto di una rubrica settimanale del più importante quotidiano finanziario italiano, che ha pubblicato informazioni già note e raccomandazioni di analisti per una selezione di aziende. L'utilizzo di dati testuali chiaramente migliora la varianza spiegata dei rendimenti azionari ma, mediante il confronto tra diverse tecniche di data mining, abbiamo osservato modeste differenze in termini di MSE, quando abbiamo aggiunto alle variabili esplicative una selezione di parole. In questo contesto, l'approccio del text mining ha mostrato la sua utilità per migliorare la potenza esplicativa in modelli di previsione, mentre è emersa la scarsa capacità esplicativa di una sentiment analysis automatica basata su un dizionario ontologico generico.*

Key words: Text Mining, Sentiment Analysis, Stock Returns, Data Mining

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

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

1 Introduction

This paper is part of a wider project whose main goal is to analyse the market reaction to the dissemination of analysts' recommendations published in print media, in order to explain the market reaction to 'buy' advice. Specifically, the hypothesis under consideration is that past analysts' recommendations induces abnormal movements in stock prices and returns.

Previous research found a positive market reaction to the publication of the past information when analysts grade the stock as a good opportunity. In particular, Cervellati et al. (2014) showed that the asymmetric market reaction supports the Barber and Odean (2008) Attention-Grabbing Hypothesis (AGH), assuming that naïve investors' behaviour affects the market. Therefore, AGH predicts positive and significant abnormal returns for positively recommended stocks and no reaction for negative ratings. In other words, the market reaction is motivated by an attention-grabbing phenomenon, because only the publication of positive recommendations induces a significant (positive) price movement.

Moreover, previous literature showed that investor mood varies systematically across calendar months and weekdays, with possible shifts in investor attention (Hirshleifer et al., 2020), affecting financial decision-making and asset prices. Hirshleifer et al. (2020) have also documented the 'day of the week' effect, which would explain how aggregate stock markets tend to do better at the end of the week than at the beginning of the week. In this regard, they found consistent results with the mood-based theory, by documenting several mood recurrence and reversal effects across calendar months and weekdays.

Our empirical analysis is based on the content of two similar weekly columns in the most important Italian financial newspaper, which published past information and analysts' recommendations on listed companies. One, named 'The Stock of the Week' appears on Saturday and the other, named 'Letter to the investor" appears on Sunday. It's important to stress that the these columns have the same author, the same content (past balance sheet and P&L data; single analyst recommendations, consensus forecasts, company's profile), and their characteristics have remained unchanged during our observation period.

The focus of this paper is to understand whether the words contained in this particular text corpus improves the explained variance of stock returns better than the use of the polarity of the same texts, obtained through a sentiment analysis using a generic ontological dictionary. Although with very different methods, this approach has long existed, as Li et al. (2014) have documented the use of textual analysis for studies on the influence of news on stock markets.

2 Data collection and processing

We collected all the 'Stock of the Week' and 'Letter to the investor' columns published from January 2005 to March 2010 that were devoted each week to a domestic company listed on the Italian Stock Exchange. The final dataset consists of 214 records.

In order to explain the variance of the average returns (AR) on the stock exchange opening day following the publication of the column, we added some explanatory variables, including: the number of quoted analysts, the natural logarithm of the order size, the natural logarithm of market capitalization, a dummy variable indicating the presence of any confounding effect, a dummy variable indicating the day of the week of the column (Saturday or Sunday), the turnover ratio, the price-to-book and the past performance (applying the absolute value or not).

2.1 Pre-processing

The pre-processing phase of the column texts involves the following steps:

- text cleaning, in order to normalize text encoding, remove punctuation, handling capitalized words, etc.
- Stop words removing in Italian language (articles, prepositions, pronouns, etc.).

In this phase we used `TextWiller` package (Solari et al., 2019), one of the few R libraries for the Italian language.

Figure 1 shows a word cloud of the most frequent words after pre-processing.

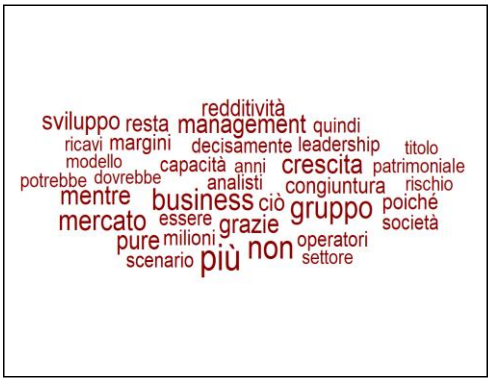

Figure 1: Most frequent words (word cloud)

2.2 Textual analysis

A textual analysis allowed us to select some terms to be used as features in the predictive models and to calculate the tf-idf weighting (Salton & Buckley, 1988). This index should show how important a word is to distinguish each weekly column in our corpus. In particular, some simple text mining procedures allowed us to create the document-term matrix we used for the regressions. In this phase we chose not to analyse the main multiword expressions (n-grams). We worked within the "bag of words" framework, as only tf-idf weighting was applied to the words and we don’t assume any Natural Language Processing rule.

Following, we used an ontological dictionary to obtain a measure of polarity for each text. Among the few resources available for the Italian language, we chose the NRC[1] lexicon (Liu, 2012), through which we extracted a polarity score for each column. The new variable containing the sentiment of each text had 41 different levels and, with reference to the sign (polarity), a positive sentiment had been attributed to 90% of the texts (a text is considered positive if the number of “positive” words is higher than the number of “negative” words).

Since we had a big sparse document-term matrix (7840 terms), we decided to allow maximal sparsity at 75%, or 25% in relation to document frequency. The resulting matrix contains only 113 terms, since we removed all the terms which have at least a 75% of empty elements (terms occurring 0 times in a document).

3 Data analysis

Concerning the aims presented above, the data analysis phase involved applying two stepwise regressions - with and without the use of words as features - to compare the two approaches in terms of explained variance. Subsequently, some data mining methods (Hastie et al., 2009) will be tested to compare the predictive power of the basic model with only the polarity feature and the one with words as features in addition (113 new features). In both cases we expect, through the use of words, an improvement in terms of R-squared (also adjusted) and Mean Squared Error (MSE).

We chose the following fitting techniques that can lead to better predictive accuracy and interpretability: Stepwise regression, Principal Components Regression (PCR), Partial Least Squares (PLS), Elastic-net, and Lasso. We estimated by cross-validation the number of components in the subset selection procedures (PCR and PLS) and the regularisation parameter in the shrinkage method (Elastic-net).

The dataset was randomly divided into two parts: a training set, used to fit the models, made up of a sample of 75% of the observations, and a validation set, used to estimate the prediction error for model selection, made up of the remaining 25% of the observations.

[1] The NRC lexicon is a list of words and their associations with eight emotions and two sentiments: negative and positive. This lexicon includes sentiment values for 13,901 words and has translations for just over 40 languages (see https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm).

4 Results

A summary of the performances of the two stepwise regressions is reported in Table 1. The R-squared indices showed a remarkable increase (about 2.5 times) of explained variance by introducing textual variables among the features (21 out of 113 terms were selected in the second model).

Table 1: Explained variance of the two models

Stepwise regression model	*R-squared*	*Adjusted R-squared*	*F-statistic*	*p-value*
M1: basic explanatory variables	0.1901	0.1667	8.099	7.178e-08
M2: adding text variables	0.4793	0.4005	6.082	7.489e-15

The variable that defines the polarity of the texts of the column was not significant in the basic model, while it was significant at 5% (p = 0.044) with a negative sign (estimated coefficient = -0.048) in the model that included the terms.

Instead, through comparisons between data mining models, we observed minor differences in terms of MSE, as shown in Table 2.

Table 2: Performances of predictive models

Technique	*MSE (basic features)*	*MSE (adding textual data)*
Stepwise	5.003575	9.434886
PCR	5.293135	5.293065
PLS	5.285215	5.284997
Elastic-net	4.788059	4.344863
Lasso	4.477144	4.702495

Lasso showed the lowest MSE for the basic model, while Elastic-net had the best performance for the model including textual variables, although the difference in terms of MSE is limited in favour of the second (4.48 vs 4.34).

5 Conclusion

The text mining approach has proven to be very useful to improve R-squared with respect to our dependent variable, the average returns. In addition, the poor ability to be a good predictor of a sentiment feature, automatically obtained from a generic lexicon, has emerged. In these cases, it would be preferable to use a supervised learning method with human tagging (Ceron et al., 2017) or, at least, a thematic lexicon.

If we observed a clear improvement in terms of R-squared, data mining techniques did not show substantial improvements in terms of MSE, by introducing some terms as features.

The use of a generic lexicon is surely a limitation for this work, as well as the presence of a quite small sample (214 records), which may affect the stability of the estimates for the data mining methods, as we were sampling just over 50 cases (training set) to compare the accuracy of various regression techniques.

References

1. Barber, B.M. and 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) doi: 10.1093/rfs/hhm079
2. Ceron, A., Curini, L., Iacus, S.M.: Politics and Big Data: Nowcasting and Forecasting Elections with Social Media. Routledge, New York (2017) doi: 10.1080/23248823.2019.1619298
3. 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) doi: 10.1080/00036846.2013.866206
4. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction. Springer, New York (2009) doi: 10.1111/j.1751-5823.2009.00095_18.x
5. Hirshleifer, D., Jiang, D., Meng, Y.: Mood beta and seasonalities in stock returns. J. of Financial Econ. (2020) doi; 10.1016/j.jfineco.2020.02.003
6. Li, Q., Wang, T., Li, P., Liu, L., Gong, Q., Chen, Y.: The effect of news and public mood on stock movements. Information Sci, 278, 826-840 (2014) doi: 10.1016/j.ins.2014.03.096
7. Liu, B.: Sentiment analysis and opinion mining. Synth. Lect. on Hum. Lang. Technol., 5(1), 1-167 (2012) doi: 10.2200/S00416ED1V01Y201204HLT016
8. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Process. & Manag. 24(5), 513-523 (1988) doi: 10.1016/0306-4573(88)90021-0
9. Solari, D., Sciandra, A., Finos, L.: TextWiller: Collection of functions for text mining, specially devoted to the Italian language. J. of Open Source Softw., 4(41), 1256, (2009) doi: 10.21105/joss.01256