Paweł Oleksy
Marcin Czupryna

Market Commentaries and Stock Prices in Poland: A Text Mining Approach*

Abstract

From a theoretical point of view, the scope and quality of available information determines the market efficiency and, thus, investors' decisions. However, an excessive amount of information leads to information overload. In the case of textual data, advanced analytical methods must be applied to identify some regularities and trends within the analysed text corpora. Text mining may be useful in supporting the decision-making process.

The paper examines the interdependencies between market commentaries and stock prices. More specifically, it verifies the linguistic characteristics of opinions distributed by institutional investors (investment fund company) and their intertemporal links to the price movements on the Warsaw Stock Exchange.

The results indicate that: 1) there is no significant linguistic difference between market commentaries written after weeks of relatively low and relatively high rates of returns on the Warsaw Stock Exchange; 2) the linguistic content of selected market commentaries does not have a predictive value for the Polish stock market; 3) commentaries with

Paweł Oleksy, Cracow University of Economics, Faculty of Finance and Law, Department of Financial Markets, Rakowicka 27, 31-510 Kraków, e-mail: oleksyp@uek.krakow.pl

Marcin Czupryna, Cracow University of Economics, Faculty of Finance and Law, Department of Financial Markets, Rakowicka 27, 31-510 Kraków, e-mail: czuprynm@uek.krakow.pl

a one-week time difference linguistically differ less than the commentaries with two or more weeks' time difference.

1. Introduction

The growing quantity of information produced, the speed of its distribution and the ease of its accessibility in recent years, combined with ongoing progress in information and communication technology (ICT) are some of the forces shaping today's financial markets – and posing a real challenge for all market participants making investment decisions in a globalised and interconnected environment. From a theoretical point of view, the scope and quality of available information determines the market efficiency and, thus, investors' decisions. However, an excessive amount of information leads to information overload, which makes searching for relevant sets of information a complex and time--consuming process requiring one to simultaneously explore and process various textual and non-textual data sets. Text information ranges from short text tweet messages and market commentaries to multi-page analyses and recommendations disseminated by both, private and public entities.

The volume of information to be analysed by investors justifies the use of advanced analytical methods to identify regularities and trends within the text corpora. Text mining, a programmed multi-stage and multidirectional procedure, may be useful in supporting decision-making by providing high--quality information obtained with relatively little human effort. Regardless of the approach (e.g. semantic parsing, bag of words), it enables the automatic sorting of collected text data and extraction of its essential features, which are then subject to detailed analysis. The text reorganisation and transition from an unorganised to an organised state (through stemming, white space elimination, lower case conversion, synonym replacement, tagging, etc.) may therefore provide investors with a potentially useful information.

The paper examines the interdependencies between market commentaries and stock prices. More specifically, we verify the linguistic characteristics of selected text corpora distributed by institutional investor (investment fund company) and their intertemporal links to the price movements at the Warsaw Stock Exchange. Our main focus lies in diagnosing the linguistic differentiation of market commentaries under different market conditions and checking whether a lexical (language style-related) analysis of analyst opinions may create an informational

advantage in the stock market. Based on observation of the market, we postulate following hypotheses:

1. Market commentaries are written in a different way after weeks of relatively low or relatively high observed rates of return.

2. The content of market commentaries has a predictive value.

3. Commentaries with shorter time difference linguistically differ less than the commentaries with a longer time difference.

2. Literature Review

The question of how public or private information affects asset price behaviour has for decades been a mainstream issue in financial market research. The problem is multifaceted and different perspectives (legal, behavioural, microstructural, technological, etc.) have been taken to examine their mutual relations and dependencies, both theoretically and empirically.

The main emphasis lies in demonstrating the impact of endogenous or exogenous information on market efficiency (Fama 1970, Boya 2013) and the consequences of the informational asymmetry between market participants (Kyle 1985, Wang & Wang 2017). The market responses to new information are usually investigated in the context of stock splits (Arbel & Swanson 1993, Kalay & Kronlund 2013), dividends (Kaestner & Liu 1998, Dasilas & Leventis 2011) or earning announcements (Basu *et al*. 2013). Due to the advances in ICT, written communication is increasingly distributed online through information services, official websites (e.g. of stock exchanges) or social media platforms (Azar & Lo 2016). This makes it easily accessible and ready for further processing. However, high quality analytical data (e.g. real-time quotes and news) is costly; it is particularly the costs associated with acquiring information that pose a significant barrier for small investors. This hampers overall market efficiency and prompts the need for appropriate regulations, which should primarily be introduced in relation to data being offered for sale by stock exchanges (Easley, O'Hara & Yang 2016).

One of the popular sources of exogenous information for investors is the opinions and recommendations of analysts, which can induce noticeable stock price movements (Jagadesh & Kim 2006, Ramnath, Rock & Shane 2008, Mielcarz 2015). The scope of price reactions depends on the recommended trading direction and tends to lead to greater price drifts in case of "sell" than "buy" intimations (Womack 1996). In general, the informativeness of analysts' recommendations translates into abnormal returns mostly in the short term than in the long term (Wnuczak 2015, Murg 2016).

From a methodological point of view, event studies or analyses of informational flux play a predominant role in testing the information-price relationship in the case of exogenous information, and various predictability tests based on random walk, long memory, seasonal effects and trading volume for endogenous information (Boya 2013). Due to the problem of interdisciplinarity and the enormous production and output of news in recent years, approaches utilising linguistics, machine learning and behavioural economics-related methods have also found use in searching for interdependencies between textual information and price changes (Khadjeh Nassirtoussi *et al*. 2014). In this context, text mining tools have become increasingly popular, especially as they prove their usefulness in verifying market sentiment (Ishijima, Kazumi & Maeda 2015) and predicting stock returns (Tetlock, Saar-Tsechansky & Macskassy 2008, Sun, Lachanski & Fabozzi 2016), corporate earnings (Tetlock, Saar-Tsechansky & Macskassy 2008) and market volatility (Mamaysky & Glasserman 2016). The analytical potential of these tools is expected to gain in importance in future financial market research.

3. Data

108 weekly market commentaries disseminated by the analyst team of a Polish investment fund company have been examined. Each commentary is freely available at the company's web site and additionally distributed through information services and financial portals. It is written in English, saved as a PDF file, covers one page and is organised in two subsections: market summary and technical analysis. Our analysis covers the period 9.06.2014 to 13.03.2017. Additionally we used the WIG time series (www.stooq.pl) covering the same period.

4. Methodology and Results

We applied standard text mining tools to verify the hypotheses. First each commentary was read into analysis environment, a tm package of R (Feinerer, Hornik & Meyer 2008, Feinerer & Hornik 2017). Each verb was then tagged using the wordnet dictionary and relevant R interface (Fellbaum 1998, Feinerer & Hornik 2017). Tagging provides the part-of-speech (POS) information. For example, \JJR denotes adjectives in their superlative form, while \NNPS is for plural proper nouns. For a systematic description of all tags used, see (Marcus, Marcinkiewicz & Santorini 1993). We constructed the following metrics for further analysis: the percentage share of adjectives in the whole text,

the percentage share of adjectives in comparative and superlative forms to all adjectives, the percentage share of adverbs in the whole text, percentage shares of adverbs in comparative and superlative form in all adjectives, and, respectively, the percentage share of verbs, nouns, foreign words and numbers in the whole text. We compared these metrics for two separate groups of weekly commentaries. The first group encompasses those commentaries written after a relatively poor WIG performance (1/3 of all commentaries with the lowest values of WIG weekly return rates), while the second group encompasses those written after a relatively strong WIG performance (1/3 of all commentaries with the highest values of WIG weekly return rates). The results are presented in Table 1. Commentaries written after the low-rate-of-return weeks tend to have more adverbs (at a significant level) and fewer numbers (non-significant) than those written after weeks with a high rate of return.

Table 1. *T* Test Comparison Results: Statistical Value and Relevant *p* Value between Commentaries Written after a Relatively Low WIG Rate of Return (*x*) and after a Relatively High WIG Rate of Return (*y*) Based on the Part-of-speech Derived Metrics

Metric	Statistic	*p* Value	*x* Mean	*y* Mean
WIG	−14.8939	0.0000	−1.59%	1.89%
Words number	0.625701	0.5335	701.36	693.97
pct_adjective	−0.06441	0.9488	9.40%	9.42%
pct_adjective_comp	0.965701	0.3377	9.38%	8.44%
pct_adverb	2.1015	0.0392	5.30%	4.84%
pct_adverb_comp	0.310867	0.7568	3.85%	3.60%
pct_noun	−0.03454	0.9725	32.52%	32.53%
pct_verb	0.707422	0.4817	12.98%	12.79%
pct_foreign	−0.58293	0.5621	0.02%	0.03%
pct_numbers	−1.21507	0.2285	3.68%	3.87%

Source: the authors' own calculations.

We verified this result based on other commentaries' classification in other groups. Namely, we used the Dow Jones Industrial Average (DJIA) index for classification into weeks with relatively low or high rates of return. The results are presented in Table 2. The differences are not significant (less significant than in the case of WIG-based commentary classification) which is to be expected as the commentaries contain a summary of the financial markets from the previous weeks on the world market but concentrate on the situation on the Warsaw Stock Exchange.

Table 2. *T* Test Comparison Results: Statistical Value and Relevant *p* Value between Commentaries Written after a Relatively Low DJIA Rate of Return (*x*) and after a Relatively High DJIA Rate of Return (*y*) based on the Part-of-speech Derived Metrics

Metric	Statistic	*p* Value	*x* Mean	*y* Mean
DJIA	−4.71202	0.0000	−0.44%	1.09%
Words number	0.668226	0.5062	701.70	694.34
pct_adjective	0.039847	0.9683	9.44%	9.43%
pct_adjective_comp	1.36671	0.1756	9.60%	8.35%
pct_adverb	1.647495	0.1040	5.20%	4.86%
pct_adverb_comp	0.737309	0.4631	4.32%	3.71%
pct_noun	0.109566	0.9131	32.59%	32.55%
pct_verb	0.460124	0.6469	12.93%	12.82%
pct_foreign	−0.23579	0.8144	0.02%	0.02%
pct_numbers	−0.83012	0.4090	3.73%	3.85%

Source: the authors' own calculations.

Table 3. *T* Test Comparison Results: Statistical Value and Relevant *p* Value between Commentaries Written before a Relatively Low WIG Rate of Return (*x*) and before a Relatively High WIG Rate of Return (*y*) based on the Part of Speech-derived Metrics

Variable	Statistic	*p* Value	*x* Mean	*y* Mean
wig_f	−17.4536	0.0000	−1.57%	1.99%
Words number	1.014274	0.3140	701.5556	691.5278
pct_adjective	0.440511	0.6611	9.47%	9.35%
pct_adjective_comp	0.002455	0.9980	9.92%	9.92%
pct_adverb	0.220446	0.8262	5.00%	4.95%
pct_adverb_comp	1.037893	0.3029	4.46%	3.60%
pct_noun	0.125138	0.9008	32.58%	32.54%
pct_verb	−0.51402	0.6090	12.83%	12.93%
pct_foreign	−0.45144	0.6531	0.02%	0.03%
pct_numbers	−0.82598	0.4116	3.69%	3.84%

Source: the authors' own calculations.

The last test verified if the linguistic content of the commentary may provide and disclose some information on the next week's rate of return. For this purpose, the classification of commentaries was put into a group that encompasses commentaries written before weeks of relatively low WIG rates of return and others written before weeks with relatively high WIG return rates. The results

are presented in Table 3. No significant differences between these two groups of commentaries were observed. The results are presented in Table 3.

We also compare how similar the commentaries are. For this purpose we first remove punctuation, stop words and numbers, change all the words to lower case and conduct a stemming procedure. We then construct the term document matrix, which contains for each word stem and each respective document the number of occurrences. We limit the number of word stems to those that had occurred at least 3 times on the global level. Then the differences in terms of the cosine and Jaccard distances are calculated, as recommended in the literature (Strehl, Ghosh & Mooney 2000). Then we compare the average (cosine and Jaccard) distance within the class of commentaries that differ by 1 week with the average distance within the class of commentaries that differ by 2 weeks and so forth. The results are presented in Table 4. We can observe that commentaries that are 1 week different are more similar to each other that the commentaries that are 2 or more weeks different.

Table 4. *T* Test Comparison Results: Statistical Value and Relevant *p* Value between the Cosine Distance of Commentaries that Differ by *I* Weeks and the Cosine Distance of Commentaries that Differ by *i* + 1 Weeks (the First Column Provides the Information on Particular *I* Value and Is in Form of: *i* / *i* + 1)

Weeks	Cosine Distance	*t* Test	*p* Value
1 / 2	0.318755	−2.21405	0.0285
2 / 3	0.34799	−0.06051	0.9518
3 / 4	0.348819	−0.11756	0.9066
4 / 5	0.350426	0	1.0000
5 / 6	0.350426	−0.08028	0.9361
6 / 7	0.351529	0	1.0000
7 / 8	0.351529	−0.0854	0.9321
8 / 9	0.352711	0	1.0000
9 / 10	0.352711	0	1.0000
10 / 11	0.352711	0	1.0000
11 / 12	0.352711	0	1.0000
12 / 13	0.352711	−0.03817	0.9696

Source: the authors' own calculations.

We also validated the result based on Jaccard distance metrics. The results presented in Table 5 corroborate the difference in the similarities between commentaries that differ by 1 week and those that differ by 2 or more weeks at a significance level of 0,1335.

Table 5. *T* Test Comparison Results: Statistical Value and Relevant *p* Value between Jaccard Distance of Commentaries that Differ by *I* Weeks and the Jaccard Distance of Commentaries that Differ by *i* + 1 Weeks (the First Column Provides the Information on Particular *I* Value and Is in Form of: *i* / *i* + 1)

Weeks	Jaccard Distance	*t* Test	*p* Value
1 / 2	0.639781	−1.50915	0.1335
2 / 3	0.659714	−0.13763	0.8907
3 / 4	0.66155	−0.07453	0.9407
4 / 5	0.66255	0	1.0000
5 / 6	0.66255	−0.08982	0.9286
6 / 7	0.663765	0	1.0000
7 / 8	0.663765	−0.08178	0.9349
8 / 9	0.664879	0	1.0000
9 / 10	0.664879	0	1.0000
10 / 11	0.664879	0	1.0000
11 / 12	0.664879	0	1.0000
12 / 13	0.664879	−0.03486	0.9722

Source: the authors' own calculations.

Table 6. A Comparison of the *T* Test Results: The Statistical Value and Relevant *p* Value between the Cosine Distance of Commentaries of Classes *i* and *i* + 1 (the First Column Provides the Information on the Particular *I* Value and Takes the Form: *i* / *i* + 1; the Description of Commentaries Classes in a Text)

Group	Mean *i*	Mean *i* + 1	*t* Test	*p* Value
1 / 2	0.314184	0.309868	0.198287	0.8438
1 / 3	0.314184	0.329076	−0.66034	0.5122
2 / 3	0.309868	0.329076	−0.99655	0.3237

Source: the authors' own calculations.

The third analysis compares the similarity of the commentaries with a one-week difference. Each pair of such commentaries is classified into three separate classes based on the previously observed WIG weekly rate of return for both commentaries of a particular pair. Namely, a given pair is classified as a 1 if one of the commentaries is preceded by a relatively low rate of return (1/3 with lowest return), while the second commentary is preceded by a relatively high rate (1/3 with highest return). A given pair is classified as a 2 if both commentaries are preceded by the same WIG rate of return (relatively low –1/3 with the lowest return, middle –1/3 with a middle return and relatively high –1/3 with the highest

return), respectively. Otherwise a given pair is classified as a 3. Table 6 compares the results, showing that there were no significant differences.

5. Conclusions

The growing amount of information distributed to stock market investors requires pertinent analytical and computational tools enabling data processing and supporting the decision-making process. Text mining can be a useful way to search for patterns and trends within textual stock market commentaries and analysts' opinions.

The analysis of the interdependencies between selected documents prepared by one of the institutional investors operating on the Polish stock market and stock price behaviour was carried out using basic text data processing operations (e.g. speech tagging) and similarity measures (cosine and Jaccard distance).

Our results indicate that: 1) there is no significant linguistic difference between market commentaries written after weeks of relatively low and relatively high rates of return on the Warsaw Stock Exchange; 2) the linguistic content of selected market commentaries does not have a predictive value for the Polish stock market; 3) commentaries with a one-week time difference linguistically differ less than the commentaries with two or more weeks' time difference.

Our further research will involve a larger set of textual data from different information sources (e.g. information services, social media) and apply advanced text mining tools to comprehensively verify their potential to provide informative value to investors. Having information gained from the text mining tools may translate into an advantage in the stock market.

Bibliography

Arbel A., Swanson G. (1993), *The Role of Information in Stock Split Announcement Effects*, "Quarterly Journal of Business and Economics", vol. 32, no 2.

Azar P. D., Lo A. W. (2016), *The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds*, "Journal of Portfolio Management", Special QES Issue, vol. 42, no 5, https://doi.org/10.3905/jpm.2016.42.5.123.

Basu S., Duong T. X., Markov S., Tan E. J. (2013), *How Important Are Earnings Announcements as an Information Source?*, "European Accounting Review", vol. 22, no 2, https://doi.org/10.1080/09638180.2013.782820.

Boya C. (2013), *Market Efficiency and Information: A Literature Review*, "Zagreb International Review of Economics and Business", vol. 16, no 2.

Dasilas A., Leventis S. (2011), *Stock Market Reaction to Dividend Announcements: Evidence from the Greek Stock Market*, "International Review of Economics and Finance", vol. 20, no 2, https://doi.org/10.1016/j.iref.2010.06.003.

Easley D., O'Hara M., Yang L. (2016), *Differential Access to Price Information in Financial Markets*, "Journal of Financial and Quantitative Analysis", vol. 51, no 4, https://doi.org/10.1017/s0022109016000491.

Fama E. (1970), *Efficient Capital Markets: A Review of Theory and Empirical Work*, "Journal of Finance", vol. 25, no 2, https://doi.org/10.2307/2325486.

Feinerer I., Hornik K. (2016), Wordnet: WordNet Interface, R package version 0.1-11, https://CRAN.R-project.org/package=wordnet (accessed: 1.08.2017).

Feinerer I., Hornik K. (2017), Tm: Text Mining Package, R package version 0.7-1, https://CRAN.R-project.org/package=tm (accessed: 1.08.2017).

Feinerer I., Hornik K., Meyer D. (2008), *Text Mining Infrastructure in R,* "Journal of Statistical Software" vol. 25, no 5, http://www.jstatsoft.org/v25/i05/ (accessed: 1.08.2017).

Fellbaum C. (1998), WordNet: An Electronic Lexical Database. Bradford Books.

Ishijima H., Kazumi T., Maeda A. (2015), *Sentiment Analysis for the Japanese Stock Market*, "Global Business and Economics Review", vol. 17, no 3, https://doi.org/10.1504/gber.2015.070303.

Jagadesh N., Kim W. (2006), *Value of Analysts' Recommendations: International Evidence*, "Journal of Financial Markets", vol. 9, no 3.

Kaestner R., Liu F. Y. (1998), *New Evidence on the Information Content of Dividend Announcements*, "The Quarterly Review of Economics and Finance", vol. 38, no 2, https://doi.org/10.1016/s1062-9769(99)80116-1.

Kalay A., Kronlund M. (2013), *The Market Reaction to Stock Split Announcement: Earnings Information after All*, Working Paper, Columbia University and University of Illinois at Urbana-Champaign.

Khadjeh Nassirtoussi A., Aghabozorgi S., Ying Wah T., Chek Ling Ngo D. (2014), *Text Mining for Market Prediction: A Systematic Review*, "Expert Systems with Applications", vol. 41, no 16, https://doi.org/10.1016/j.eswa.2014.06.009.

Kyle A. S. (1985), *Continuous Auctions and Insider Trading*, "Econometrica", vol. 53, no 6, https://doi.org/10.2307/1913210.

Mamaysky H., Glasserman P. (2016), *Does Unusual News Forecast Market Stress?*, The Office of Financial Research Working Paper no. 16-04.

Marcus M. P., Marcinkiewicz M. A., Santorini B. (1993), *Building a Large Annotated Corpus of English: The Penn Treebank*, "Computational Linguistics" vol. 19, no 2.

Mielcarz P. (2015), *Główne cechy rekomendacji a zmiany cen akcji na GPW w Warszawie*, "Zeszyty Naukowe Uniwersytetu Szczecińskiego", no 854, "Finanse, Rynki Finansowe, Ubezpieczenia" no 73.

Murg M. (2016), *Intraday Effects of Analysts' Recommendations on International Stock Markets*, Annual International Conference on Accounting & Finance.

Ramnath S., Rock S., Shane P. (2008), *The Financial Analyst Forecasting Literature: A Taxonomy with Suggestions for Further Research*, "International Journal of Forecasting", vol. 24, no 1, https://doi.org/10.1016/j.ijforecast.2007.12.006.

Strehl A., Ghosh J., Mooney R. (2000), *Impact of Similarity Measures on Web-page Clustering*, Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI 2000).

Sun A., Lachanski M., Fabozzi F. J. (2016), *Trade the Tweet: Social Media Text Mining and Sparse Matrix Factorization for Stock Market Prediction*, "International Review of Financial Analysis", vol. 48, https://doi.org/10.1016/j.irfa.2016.10.009.

Tetlock P. C., Saar-Tsechansky M., Macskassy S. (2008), *More than Words: Quantifying Language to Measure Firms' Fundamentals*, "The Journal of Finance", vol. 63, no 3, https://doi.org/10.1111/j.1540-6261.2008.01362.x.

Wang K. T., Wang W. W. (2017), *Competition in the Stock Market with Asymmetric Information*, "Economic Modelling", vol. 61, https://doi.org/10.1016/j.econmod.2016.11.024.

Wnuczak P. (2015), *Effectiveness of Recommendations Issued by Stock Market Analysts in Periods of Stagnation on Capital Markets*, Research Papers of the Wroclaw University of Economics, no 412.

Womack K. L. (1996), *Do Brokerage Analysts' Recommendations Have Investment Value?*, "Journal of Finance", vol. 51, no 1, https://doi.org/10.1111/j.1540-6261.1996.tb05205.x.

Komentarze rynkowe a ceny akcji w Polsce – analiza z wykorzystaniem metod eksploracji danych tekstowych
(Streszczenie)

Z teoretycznego punktu widzenia zakres i jakość dostępnych informacji determinuje efektywność rynku, a tym samym wpływa na decyzje inwestycyjne inwestorów. Jednakże duża ilość informacji nie przekłada się wprost na poprawę tej efektywności, wymaga natomiast odpowiedniej selekcji, segregacji oraz strukturyzacji w celu ekstrakcji potencjalnych sygnałów kupna lub sprzedaży instrumentów finansowych. W przypadku informacji tekstowych realizowanie tych działań może zostać usprawnione poprzez zastosowanie komputerowych metod eksploracji danych tekstowych (*text mining*).

Przedmiotem analizy jest weryfikacja parametrów lingwistycznych komentarzy rynkowych oraz ocena ich powiązań z wahaniami cen akcji notowanych na Giełdzie Papierów Wartościowych w Warszawie. Badania z wykorzystaniem wybranych metod i mierników analizy danych tekstowych (m.in. znakowanie części mowy, odległość cosinusowa, Jaccarda) prowadzone są na bazie korpusów tekstowych opracowań analitycznych, sporządzanych regularnie w odstępach tygodniowych przez profesjonalny zespół analityków rynkowych jednego z towarzystw funduszy inwestycyjnych dla polskiego rynku akcji.

Wyniki badań wskazują, że: 1) nie występuje statystycznie istotne zróżnicowanie językowe tekstów komentarzy sporządzanych po zakończeniu tygodni o relatywnie niskich oraz tygodni o relatywnie wysokich stopach zwrotu na Warszawskiej Giełdzie Papierów Wartościowych, 2) zawartość językowa wybranych komentarzy rynkowych nie ma wartości predykcyjnej dla polskiego rynku akcji, 3) komentarze z jednotygodniową różnicą czasową różnią się pod względem lingwistycznym mniej niż komentarze z różnicą czasu dwóch lub więcej tygodni.

Słowa kluczowe: informacja, przewidywanie cen akcji, analiza danych tekstowych, rekomendacje analityków finansowych, komentarze rynkowe.