News Analytics: Exploring Predictive Power of Aggregated Text Sentiment Measure

Caslav Bozic* and Detlef Seese*

News analytics and text sentiment detection were established in recent years as methods that can support forecasting of market movements. The body of literature exploring relations between sentiment measures and various financial indicators is rapidly growing. We contribute by taking a more global view and by proving that there is a positive and significant relation between average sentiment of all news published about one country and that country's GDP and GDP change. This relation holds also when observing aggregated sentiment in one year against GDPs and GDP changes in following years.

1. Introduction

The recently established discipline of Computational News Analytics offers a possibility to automatically analyse published news. One of the tasks of News Analytics (NA) is determining novelty of the published text, i.e. automatically deciding whether the news story at hand is reporting about some new event, or it is merely introducing new facts about an event that is already known. Another example of a NA task is detecting sentiment of the published news item. Analysing the language of a text, and the selection of the words author used, it is possible to determine whether the analysed text is positive or negative in sentiment, and also the degree of positivity or negativity. The texts lacking high levels of positivity or negativity are usually denoted as neutral. Such measure we call text sentiment measure.

We use the output of Thomas Reuters NewsScope Sentiment Engine, a commercially available News Analytics system to define a sentiment measure. The sentiment measures of individual news stories are then aggregated to a yearly level and according to the country they are related to. This aggregated sentiment measure is then related to the gross domestic product and the gross domestic product change for the country.

The dataset contains macroeconomic data from 181 countries in the period of 6 years (from 2003 to 2008) and an aggregated sentiment measure of over 6 million news records. Panel regression and ordinary least squares are used for determining relations between variables and the robust standard error is reported.

* Institute of Applied Informatics and Formal Description Methods, Karlsruhe Institute of Technology (KIT), Germany. Corresponding author's e-mail: bozic@kit.edu.
Financial support from the Graduate School 895 "Information Management and Market Engineering" funded by Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

We find that there is a positive and significant (to the level of 5%) relation between the average sentiment of all news items published in one year for one country, and that country's GDP, as well GDP change. We find that the relation is positive and significant for lagged values of sentiment, too. This fact supports the statement that aggregated news sentiments from one year have certain predicting power on GDP and GDP change in future years.

The paper is organized as follows: first we give an overview of existing work on sentiment detection and its predictive power on different financial indicators. Then, we describe the data we used and the pre-processing steps that were needed to prepare the dataset for the analysis. In the Methodology chapter, we explain the econometric tools we used, and the results of the analysis are presented in chapter four. The last chapter represents conclusion and some outlook.

2. Literature Review

In the current body of literature related to our work we can observe two groups of contributions. The first group contains systems designed with a goal to analyse published financial texts and to extract sentiment score or just classify them according to sentiment of the text. The goal of authors of such systems is to predict some of the financial indicators, so they implicitly relate their sentiment measure to these financial indicators, e.g. stock price trend or stock price volatility. The second group contains mostly papers with a methodology from empirical finance that show a significant relation between sentiment scores and different financial indicators. The methodology of sentiment extraction is usually out of scope of such works, and they often use commercial systems, with a proprietary methodology that has not been described in the academic literature.

An Overview of the systems falling into the first group is given by (Mittermayer and Knolmayer, 2006), while (Mitra and Mitra, 2011) list some additional systems that were not mentioned in (Mittermayer and Knolmayer, 2006) or were created after the year 2006. The articles from this category are, for example, (Wüthrich et al., 1998) that uses news in an attempt to forecast the trend of the index daily value one day ahead, or (Lavrenko et al., 2000) which determines the influence of sentiment of news articles from Yahoo!Finance to particular U.S. stocks. Further works include (Gidofalvi and Elkan, 2003) which determines the influence of sentiment on stock prices of constituents of the Dow Jones index, (Fung, Yu, and Lam, 2003) which has a goal to predict a price the trend for intraday market movements of some of the stocks listed on the Hong Kong Stock Exchange, and (Mittermayer and Knolmayer 2006b) which propose a high frequency price trend forecast system that classifies press releases of publicly traded companies in the U.S.

While most of the articles in this group focus on predicting price trends of a single stock or an index, there are publications that aim at determining the influence that news sentiment has to volatility. (Thomas, 2003) improves risk-return profile

by exiting the market in case of news that are predicting high volatility, while (Schulz, Spiliopoulou, and Winkler, 2003) attempt to classify press releases of German public companies according their influence on volatility of stock prices.

Some of the authors do not primary attempt to prove economical relevance of their published texts by evaluating specifically tailored trading strategies, but rather to find statistically relevant relations between financial indicators and sentiment extracted from the text. (Antweiler and Frank, 2004) classifies messages posted to Yahoo!Finance and Raging Bull and determine their sentiment. They do not find statistically significant correlations with stock prices, but they find sentiment and volume of messages significantly correlated to trade volumes and volatility. In their methodological paper (Das and Chen, 2007) offer a variety of classifiers, as well as a composed sentiment measure as a result of voting among classifiers. In the illustrative example they analyze Yahoo's stock boards and stock prices for eight technology companies, but they do not find clear evidence that the their sentiment index can be predictive for stock prices. (Tetlock, 2007) observes the Wall Street Journal's column "Abreast of the Market", uses the content analysis software "General Inquirer" together with a Principal Component Analysis approach and finds that high pessimism in published media can predict downward pressure on market prices. Authors of (Tetlock, Saar-Tsechansky, and Macskassy, 2008) succeeded to find that the rate of negative words in news stories about a certain company predicts low earnings of the company. In (Bozic and Seese, 2011) we show a significant relation between sentiment scores produced by our own system based on neural networks and a future daily returns with a forecast horizon of up to three days.

The second group contains articles relating media coverage with market movements, e.g. (Mitra, Mitra, diBartolomeo, 2009), (Da, Egelberger, and Gao, 2009), and (Dzielinski, Rieger, and Talpsepp, 2011). There are articles that explain the influence of market developments to macroeconomy, like (Atje and Jovanovic, 1993), but to the authors' knowledge no publication explores direct relation between news sentiment scores and macroeconomic variables.

3. Methodology

The data used in this research originates from two main sources: the source for quantified news data and the basis for sentiment score comes from Reuters NewsScope Sentiment Engine for historical data, while the public World Economic Outlook (WEO) database of the International Monetary Fund provided the macroeconomic data.

Data constituting the output of the Reuters NewsScope Sentiment Engine represents the author's sentiment measure for every English-language news item published via NewsScope in the period from 2003 to 2008 inclusive. The measure classifies a news item into one of three categories: positive, negative, or neutral. The probability of the news item falling into each of the categories is also given. Each record represents a unique mention of the specific company, with a

possibility of one news item relating to more than one company. In our dataset there are 6,127,190 records. Each news item can have multiple tags called topic codes, and topic codes are grouped into categories. One of the categories represents countries, and it can be used for determining what country is mentioned in the particular news item. Using this tagging feature, average sentiment per country can be calculated. The data is aggregated using the system described in (Bozic et al., 2010) to the level of country and year. The data provided by Reuters are outputs of the classifiers in the form of three probabilities: probability that the analysed news item is positive in sentiment P_{pi}, that it is neutral P_{oi}, or that it is negative in sentiment P_{ni}. We define *sentiment score* S_i by subtracting probabilities for positive and negative class, and by aggregation we get a *yearly sentiment score for a country* $S_{y,c}$.

$$S_i = P_{pi} - P_{ni}$$

In this way we get panel data with six consecutive years and 233 countries and territories. We combine these data with the data for the same period from WEO database. After excluding countries that are not a part of the WEO database, we get the data about 181 countries. The selected subset of variables from WEO dataset is given in Table 1.

Table 1: Selected variables from WEO dataset

Variable Description	Units	Scale
Gross domestic product, constant prices (change)	Percent change	
Gross domestic product, current prices	U.S. dollars	Billions
Gross domestic product per capita, current prices	U.S. dollars	Units
Gross domestic product based on purchasing-power-parity (PPP) valuation of country GDP	Current international dollar	Billions
Gross domestic product based on purchasing-power-parity (PPP) per capita GDP	Current international dollar	Units
Gross domestic product based on purchasing-power-parity (PPP) share of world total	Percent	
Investment	Percent of GDP	
Gross national savings	Percent of GDP	
Inflation, average consumer prices	Index	
Inflation, average consumer prices (change)	Percent change	
Inflation, end of period consumer prices	Index	
Inflation, end of period consumer prices (change)	Percent change	
Import volume of goods and services (change)	Percent change	
Export volume of goods and services (change)	Percent change	
Unemployment rate	Percent of total labor force	
Population	Persons	Millions

The estimation of panel regression coefficients on the data is performed with the help of OxMetrics and DPD software (Doornik, Arellano, and Bond, 2001) and the results are presented in the next chapter four.

4. Findings

The results are represented in Tables 2-5, and the explanation for the statistical significance of the results in Table 6. The first relation we explore is the intensity of the media coverage to the macroeconomic variables. The results that are statistically significant are presented in Table 2. The level of statistical significance is represented using asterisks, as explained in Table 6. As a proxy for intensity of the coverage we count the total number of published news items that are covering a story connected to a given country. We find the positive relation to Gross domestic product (GDP), Gross domestic product based on purchasing-power-parity valuation of country GDP (GDP-PPP), Gross domestic product based on purchasing-power-parity per capita GDP (GDP-PPP-PC), change of export volume of goods and services (CH-EX), and population (P). The negative and significant relation is found for change of inflation measured by average consumer prices (CH-INFL-AVG), change of inflation measured by end of period consumer prices (CH-INFL-EOP), and change of import volume of goods and services (CH-IM). All these result suggest that countries with better economies have a greater number of mentions in news. Additionally, the number of news items is negatively and significantly related to change of gross domestic product (CH-GDP). The observation is consistent with the fact that more media attention is payed to developed countries, which generally have positive macroeconomic development, except for the fact that their GDP is not growing as fast as in some developing countries.

Table 2: Estimated coefficients, number of items as independent variable

Variable Description	Coefficient
Gross domestic product, constant prices (change)	-0.00003998 ***
Gross domestic product, current prices	0.00444543 ***
Gross domestic product based on purchasing-power-parity (PPP) valuation of country GDP	0.00418583 ***
Gross domestic product based on purchasing-power-parity (PPP) per capita GDP	0.01198870 **
Inflation, average consumer prices (change)	-0.00001217 *
Inflation, end of period consumer prices (change)	-0.00000910 **
Import volume of goods and services (change)	-0.00001954 ***
Export volume of goods and services (change)	0.00001033 **
Population	0.00001473 ***

The previous results are rather telling us more about the structure of coverage of western media outlets, like Reuters, and their focus on strong economies. So we now try to explore the sentiment score itself and its relations to macroeconomic variables. The Table 3 shows estimated values of coefficients for OLS panel regression when only contemporaneous sentiment is used as independent variable. We find positive and significant relation of the average sentiment of all news items published in one year and mentioning a country, with that country's GDP in that year, and with that country's GDP change compared to the previous

year. Thus, the country with more positive news published in a year will have greater GDP and more positive GDP change by the end of the year.

Table 3: Estimated coefficients, sentiment score as independent variable

Variable Description	Coefficient
Gross domestic product, constant prices (change)	1.51385 **
Gross domestic product, current prices	44.3611 **
Gross national savings	6.29754
Import volume of goods and services (change)	4.67562
Export volume of goods and services (change)	1.88586
Population	1.18362 ***

This prediction goes beyond the time span of one year, as the data from Table 4 suggests. It shows that the sentiment score in one year can be significantly and positively related to GDP, GDP change, gross national savings, and import volume for the following years, up to three years ahead. The negative estimations for the coefficients in Table 4, like relation to export volume, or import volume three years ahead, are all not significant, so we can not make any statements on that relations.

Table 4: Estimated coefficients, lagged sentiment score as independent variable

Variable Description	S	S(t-1)	S(t-2)	S(t-3)
Gross domestic product, constant prices (change)	2.89904**	1.37394	1.65560*	0.96562
Gross domestic product, current prices	32.6456	67.7682*	52.5004	93.2768***
Gross national savings	14.6525*	14.6006**	8.52303*	6.20423**
Import volume of goods and services (change)	13.8133**	12.5976**	0.456780	-2.97500
Export volume of goods and services (change)	1.40626	-4.28914	-2.96447	-0.961789
Population	0.95896***	0.470776	1.03818**	1.23115***

In Table2, Table3, and Table 4 it can be noticed that the relation with the population of the given country is always positive and very significantly related to the sentiment score and the number of news items. So we repeat the regression, adding the population as the independent variable besides sentiment score. The Table 5 shows that the results for GDP hold also in that case and are robust enough.

Table 5: Estimated coefficients, sentiment score and population

Variable Description	Sentiment	Population
Gross domestic product, constant prices (change)	1.57371	-0.00905784
Gross domestic product, current prices	55.3923 *	58.0685 ***

Table 6: Statistical significance of the results

	p value
$***$	$< 1\%$
$**$	$< 5\%$
$*$	$< 10\%$
otherwise	$>= 10\%$

5. Conclusion

In the paper we firstly observe significant relations from the number of news for a country and that country's macroeconomic indicators. The relation is positive for macroeconomic variables where a high value is a sign of a strong economy, while it is negative where a high value is generally bad for the economy, as in the case of inflation change. There is an exception to this in the case of GDP change. This observation is consistent with the fact that more media attention is payed to developed countries, which generally have positive macroeconomic development, except for the fact that their GDP is not growing as fast as in some developing countries.

We also find that there is a positive and significant (to the level of 5%) relation between average sentiment of all news items published in one year for one country, and that country’s GDP, as well GDP change. We find that the relation is positive and significant for lagged values of sentiment, too. This fact supports the statement that aggregated news sentiments from one year have certain predicting power on GDP and GDP change in future years.

References:

Antweiler, W. & Frank, M.Z. 2004, „Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", *The Journal of Finance*, 59(3), pp. 1259-1294

Atje, R. & Jovanovic, B. 1993, "Stock markets and development", European Economic Review, 37(2-3), pp. 632-640

Bozic, C., Riordan, R., Seese, D. & Weinhardt, C. 2010, „Towards a Benchmarking Framework for Financial Text Mining", *Information Management and Market Engineering*, 2, KIT Scientific Publishing, pp. 21-36

Bozic, C. & Seese, D. 2011, *Neural Networks for Sentiment Detection in Financial Text*

Da, Z., Engelberg, J. & Gao, P. 2009, *In Search of Attention*

Das, S. & Chen, M 2007, "Yahoo! for Amazon: Sentiment extraction from small talk on the web", *Management Science*, 53(9), pp. 1375-1388

Doornik, J., Arellano, M. & Bond, S. 1991, *Panel Data estimation using DPD for Ox*

Dzielinski, M., Rieger, M.O. & Talpsepp, T. 2011, „Volatility, asymmetry, news and private investors", The handbook of news analytics in finance, Wiley Finance

Gidófalvi, G. & Elkan, C. 2003, *Using news articles to predict stock price movements*

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. & Allan, J. 2000, *Mining of Concurrent Text and Time-Series*

Mitra, L. & Mitra, G. 2011, "Applications of news analytics in finance: A review", *The handbook of news analytics in finance*, Wiley Finance

Mitra, L., Mitra, G. & Dibartolomeo, D. 2009, „Equity portfolio risk estimation using market information and sentiment", *Quantitative Finance*, 9(8), pp. 887-895

Mittermayer, M. and Knolmayer, G. 2006, *Text mining systems for market response to news: A survey*.

Mittermayer, M.-A. & Knolmayer, G.F. 2006b, *NewsCATS: A News Categorization and Trading System*

Pui Cheong Fung, G., Xu Yu, J. & Lam, W. 2003, *Stock prediction: Integrating text mining approach using real-time news*

Schulz, A., Spiliopoulou, M. & Winkler, K. 2003, „Kursrelevanzprognose von Ad-hoc-Meldungen: Text Mining wider die Informations\überlastung im Mobile Banking", *Wirtschaftsinformatik*, 2, pp. 181-200

Tetlock, P. 2007 "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *Journal of Finance*, 62(3).

Tetlock, P., Saar-Tsechansky, M. & Macskassy, S. 2008, "More Than Words: Quantifying Language to Measure Firms' Fundamentals", *Journal of Finance*, 63(3).

Thomas, J. 2003, *News and trading rules.*

Wüthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J. & Lam, W. 1998, *Daily prediction of major stock indices from textual www data*