NARODOWY
BANK POLSKI

NBP Working Paper No. 349

What news can really tell us? Evidence from a news-based sentiment index for financial markets analysis

Anna Marszal

NBP Working Paper No. 349

What news can really tell us? Evidence from a news-based sentiment index for financial markets analysis

Anna Marszal

Narodowy Bank Polski
Warsaw 2022

Anna Marszal – Narodowy Bank Polski; anna.marszal@nbp.pl

This paper represents the opinions of the author. It is not meant to represent the position of the NBP. Any errors and omissions are the fault of the author.

Published by:
Narodowy Bank Polski
Education & Publishing Department
ul. Świętokrzyska 11/21
00-919 Warszawa, Poland
www.nbp.pl

ISSN 2084-624X

Contents

Abstract

This study presents a state-of-the-art approach in measuring financial market sentiment, namely, extracting it from news headlines. The sentiment index is constructed by analysing over 124,000 news items for the 2020-2021 period using natural language processing methods. Its informational power is validated by the strong correlation with the VIX index as well as by the occurrence of common periods of higher volatility of both measures. These findings reinforce the treatment of the news-based index as a true sentiment indicator and contribute to its usage independently of any financial instruments. Additionally, a direction of significant correlation coefficients between the sentiment indicator and selected financial assets is consistent with the natural logic of capital flows in financial markets.

At the same time, the developed tool allows to identify not only market sentiment, but also the main factors contributing to its direction and time periods in which they are of most significance. It is necessary to understand that the analysed period is specific as it coincides with the outbreak and development of the COVID-19 pandemic. This was reflected in the results that highlight coronavirus as the dominant topic throughout the dataset.

JEL: C6, C8, G4

Keywords: market sentiment, natural language processing, lexicon-based models, VADER, risk aversion, risk appetite, VIX index, news, volatility

1. Introduction

Sentiment analysis has long entered the tools of investors and financial market analysts. Market sentiment is often referred to as a way of measuring investor behaviour in response to news or events [1]. The importance of this aspect results from its recognition as a factor responsible for the deviation of the real asset price from the fundamental valuation [2].

Specialized and academic literature presents various types of sentiment measurement. Currently, in practice, there are several approaches to measuring market sentiment, e.g. survey-based indicators obtained directly from market participants (e.g. the University of Michigan Index of Consumer Sentiment or Conference Board Confidence Index). The disadvantage of these traditional methods is that they are costly and time-consuming to conduct [3] and generally lead to a low frequency (e.g. monthly) of sentiment measure. That may constitute a crucial role in decision-making for monetary or government authorities, but it is less relevant to financial markets that absorbing every piece of information extremely fast.

Often, the role of sentiment is assigned to expected volatility measures (e.g. the VIX index), which are especially popular among financial market practitioners. In academic literature one may find a study [4] that gives the reasons why VIX index is referred to as the barometer of investor fear. In [5] modelling of expected volatility was presented as the gauge of investor fear in India.

Recently, alternative approaches in constructing sentiment measures based on published textual information (news) have gained interest. For this purpose it is necessary to resort to natural language processing (NLP) methods, which allow sentiment extraction and classification. NLP methods can be roughly categorized into machine learning and lexicon-based (dictionary-based) approaches [6].

Machine learning-based methods involve both supervised and unsupervised modelling to classify textual data into predefined sentiment categories [3]. For that purpose, standard algorithms solving the classification problem, such as, K-Nearest Neighbours, Support Vector Machine, Naïve Bayes method or Artificial Neural Networks, are advantageous. This is exemplified in [7], where the authors employ a deep neural network to detect news sentiment and use it then to improve stock market forecasts. The mining of social media data for discovering public sentiment and its application in prediction in the stock exchange is quite a popular issue that is well described in lots of papers e.g. [8], [9], [10], [11]. [3] computes the business sentiment index based on daily newspaper articles applying supervised machine learning modelling. The advantage of machine learning is the ability to train and customize models for specific purposes and domain. However, the limitation of this approach is the need of a sizeable training dataset and the low applicability to new data [12], [6]. Thus according to [12], machine learning sentiment analysis is more appropriate for document-level texts with more words, while for sentiment categorization at the sentence level (such as news headlines) a lexicon-based approach might be preferable.

Lexicon-based methods are easy to implement and do not require prior model training. Hence, they are typically much faster than machine learning algorithms. They use instead a predefined dictionary, in which every word is associated with a specific sentiment score. [13] showed that these methods have robust performance across domains and texts. Still, this approach is not without drawbacks, such as: (i) difficulties in creating an all-purpose dictionary that satisfies different text areas, (ii) finite number of words in the dictionary and the constant sentiment tags of words or (iii) multiple meanings of some words [12], [6]. The lexicon way of sentiment analysis is described in [14], applying it to financial news from the New York Times, in [15] on BBC news dataset or in [16] on web reviews.

Unlike survey-based measures or measures relying on financial instruments quotations, this study demonstrates a sentiment measure within the NLP framework using a lexicon-based approach. More specifically the rule-based sentiment model validated by the authors in [17] and in numerous studies is applied in order to classify news as positive, negative or neutral.

The added value of this research, that it also corresponds to its goal, is primarily the practical use of a modern method – such as NLP – in quantifying the impact of news on market sentiment. This study presents the methodological approach to the construction of such a measure with daily frequency and discusses the obtained results.

This paper is organized as follows. Section 1 describes the methodology of the sentiment index construction, while section 2 tries to explain the key changes of the presented gauge of sentiment. Section 3 compares the index with another measure often used by financial practitioners as a proxy of market sentiment, followed by Section 4, where possible examples of additional applications of the constructed sentiment index are discussed. The final section contains conclusions.

2. Index construction

In this study, sentiment is measured on the basis of relatively high frequency data, using a semantic analysis of headlines of the most important news during the day. The research focuses on the headlines analysis, omitting the full texts of news. This approach is determined by several aspects: (i) the headlines summarize the meaning of the entire text or indicate the topic of the article [17], (ii) processing and analysing of headlines is easier and therefore requires less computing power, (iii) high frequency automated trading systems are also likely to process the headline rather than the entire piece of news in fast market positioning.

2.1. Data acquisition and pre-processing

The corpus, or the dataset in other words, is composed of news headlines downloaded from the Eikon Refinitiv for the years 2020-2021. The beginning of the analysis period corresponds to the earliest date from which it was technically possible to download and cumulate data due to the limited amount of text data stored on the platform with

Before downloading, the “major breaking news” filter of the Eikon Refinitiv platform was used. Given the financial nature of this platform, most of the news already has some relation to the financial markets. Therefore, thematically, it was enough to exclude sports items to get the main news in the field of economics, finance, politics, statements of government representatives, central banks, etc. A more severe data narrowing (e.g. to forex or fix income markets) could lead to the loss of essential information and affect the daily sentiment. On the other hand, allowing to get all the pieces of news for a given day can bring a lot of unnecessary noise and also distort the daily sentiment.

All news was downloaded in English, and their source was Reuters – the largest international news agency. There are also other works that use Reuters news for building sentiment indicators (e.g. [18], [19]).

Ultimately, for each day a different number of headlines is downloaded (up to a maximum of 200) depending on how many news items were published on the platform and how many are left after filtering.

The raw text data was processed using an in-house Python algorithm, consisting of the following steps. Initially, the tokenization of headlines was performed, which means that sentences are split into single units such as words and punctuation marks. Secondly, punctuation marks were excluded as well as "stop words", which are the most common words for any natural language. While being sometimes necessary for the language syntax, they do not add much meaning to the sentence. Examples of such words in English are: "the", "is", "in", "for", "where", "when", "to". One important exception in the algorithm is the word "not", as it can change the meaning of a sentence to the opposite, and therefore, the sign of sentiment score. Subsequently, the words in the text underwent stemming – the process of reducing inflectional forms and derivationally related forms of a word, leaving only the word stem. Eventually, the words were joined back to sentences.[1]

2.2. Applying the VADER model in sentiment analysis

The VADER (Valence Aware Dictionary for sEntiment Reasoning) model was applied to analyse the content of each headline. This model is available in the Natural Language Toolkit (NLTK) package of the Python programming language. It is a rule-based model, which assesses individual words from a predefined English lexicon and classifies the text as positive, negative or neutral, determining the intensity of sentiment. The VADER model uses its own lexicon validated by independent human judges [20]. The authors of the VADER model proved its credibility by comparing

[1] For example, the original headline as of 30.09.2021 "*POLL-Oil to gain as demand recovery resumes, but virus risks remain*" vs text after processing "*poll oil gain demand recover resume virus risk remain*".

the results of the classification problem of short texts to the results of other popular models of sentiment analysis and machine learning (Naïve Bayes, Maximum Entropy, or Support Vector Machine), as well as to the human-made assessments [20]. That is why the VADER is considered to be a gold standard lexicon. There is wide variety of applications of the VADER model, e.g. in [21] for classification of movie reviews, in [22] for examining Bitcoin tweets, during the COVID-19 pandemic or in [23] for analysing financial social networks.

In the VADER model every single word is assigned a score ranging from -4 to 4, based on a predefined lexicon. The sign makes the model sensitive to polarity, while the magnitude makes it sensitive to the intensity of sentiment. For a better understanding of the judging criteria of the VADER model, let us have a look at examples of real words found in the analysed Reuters headlines. Words receiving a negative score include: *death* (-2.9), *violation* (-2.2), *turmoil* (-1.5), *debt* (-1.5), *deny* (-1.4), *loss* (-1.3), *impose* (-1.2), *risk* (-1.1), while the following obtained a positive score: *optimism* (2.5), *gain* (2.4), *valuable* (2.1), *support* (1.7), *integrity* (1.6), *ready* (1.5), *security* (1.4), *strengthen* (1.3). The compound score for the whole headline is calculated by summing up the results for each word, adjusting to the special rules and then normalizing this value to range (-1,1).

These special rules include so called generalizable heuristics that humans use to judge the sentiment intensity of the text. According to [17], the following methods are applied for text sentiment analysis in the VADER model:

- The contrastive conjunction “but” can change the polarity of sentiment. The sentiment magnitude of the text before “but” is reduced by 50%, while the sentiment magnitude of the text after the conjunction is increased by 50%
- Negation “not” reverses the polarity of the sentiment.
- Degree modifiers or booster words can either increase or decrease the intensity of sentiment (e.g. “extremely good” vs “marginally good”).
- Capitalization, especially “ALL-CAPS”, and

- punctuation, namely the exclamation point (!), both increase the intensity of sentiment.[2]

The last two heuristics are more specific for social texts rather than for headlines of financial news. Sometimes there are phrases consisting only of capital letters in the downloaded data (e.g. Reuters polls tag – see Footnote 1), but it makes no sense for the model to increase the sentiment intensity only because of that. Therefore, the last two heuristics are automatically ignored in model valuation after data pre-processing.

Finally the obtained compound score is, in fact, the sentiment of the headline, where -1 is the most negative, 0 is neutral and +1 is the most positive. The closer the compound score (in absolute terms) to zero, the more neutral the headline's sentiment. The sentiment results of each headline for a particular day were summed up, resulting in a daily sentiment time series.

2.3. Describing the sentiment index behaviour

In general, one can observe three different types of behaviour of the sentiment index in a day:

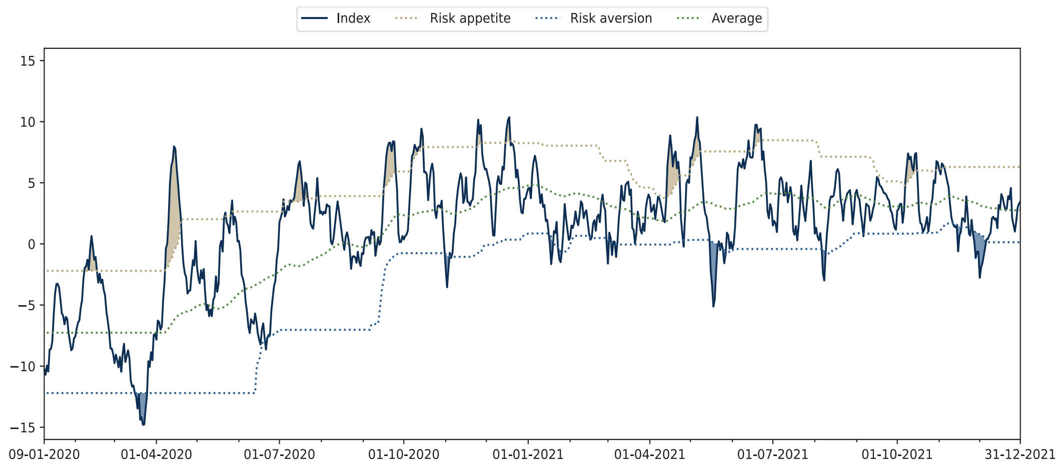
- The index is close to its average value in the case when the news published that day is neutral or when negative and positive headlines are balanced. Such behaviour of the index indicates neutral sentiment to risk.
- The index surges during a day in two cases: (i) when many headlines are published with small positive scores; however, their amount is big enough to result in a highly positive sentiment index for the whole day; (ii) when strongly positive scores are assigned to a few pieces of news, resulting in a strong positive index for the whole day. Such a change in the index corresponds to an improvement in market sentiment or in other words to an increase in risk appetite.
- The index drops significantly below its mean value in two cases, similar to the ones mentioned above, i.e. the accumulation of many slightly negative or a few intensely negative headlines. Such behaviour of the index should be

[2] For more details on how the intensity of the sentiment can be changed because of the listed heuristics look into [17].

interpreted as a worsening of sentiment, or in other words, an increase in risk aversion.

Thereafter, in order to reduce excessive volatility, the data was smoothed using a ten-day moving average. The resulting time series is a market sentiment index (see Figure 1). Its extreme values, such as the 90th and 10th percentiles of the empirical distribution, correspond to the periods of positive (risk-on) and negative (risk-off) sentiment, respectively. It is worth mentioning that the market tends to discount new information in asset valuation quickly and the topic becomes obsolete from a market perspective swiftly too. That can explain why the index does not stay beyond the defined bounds for very long.

Figure 1. Sentiment index



Source: Own work based on Refinitiv data.
Note: Percentiles and the mean are calculated in a rolling window of 90 days.

3. Explanation of changes in the sentiment index

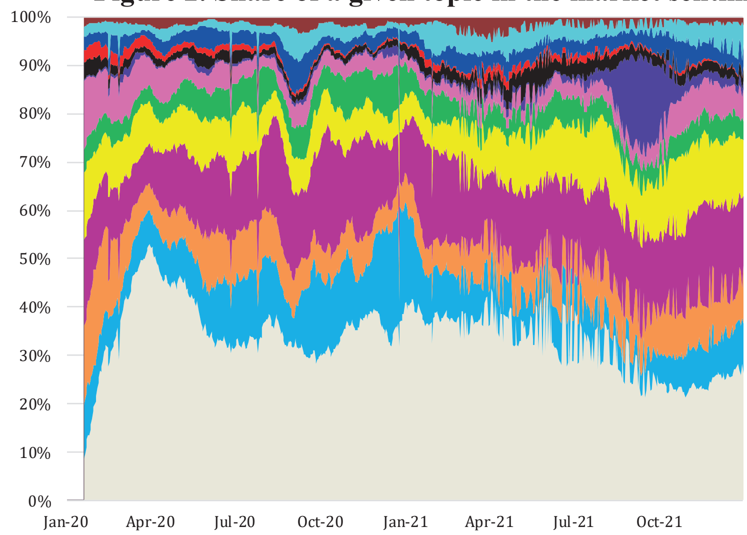
In this section, factors having an impact on the sentiment index during the analysed period are investigated. For that purpose, an analysis of the content of headlines was performed, resulting in finding 50 relevant[3], most frequent words and collocations for each half year, which can be classified into several topic groups (see Table A.1 in Attachment A).

[3] The results were subjectively filtered by the relevance of the words, because among the most frequent ones there were also words such as “update”, “minister”, “says”, “week” etc., which do not bring any informational value.

Briefly summarizing the list of the most popular thematic issues, it can be seen that most of the topics were repeated every half a year and the coronavirus thread was the most important topic in the whole investigated period. Keywords describing the vaccination process (such as “vaccine”, “dose”, “astrazeneca”) joined the coronavirus topic in the second half of 2020, while the mutation Omicron appeared in the top results at the end of 2021. It is interesting that previous variants of coronavirus did not appear in the results as frequently as Omicron. In the first sub-periods, the topic of the United Kingdom and Brexit was much more frequent than in the last ones. On the other hand, in the last half of 2021, news about the situation in Afghanistan, energy commodities and monetary policy was of great importance. In this context, it is also interesting that the word “inflation” was appearing in the results of the most common words throughout 2021.

The next stage of the analysis was to determine the frequency of the defined words (from Table A.1) for each of the selected topics on a single day during the entire dataset. This allowed to illustrate the relative importance of each of these topics to the sentiment index per day, as depicted in Figure 2.

Figure 2. Share of a given topic in the market sentiment index per day



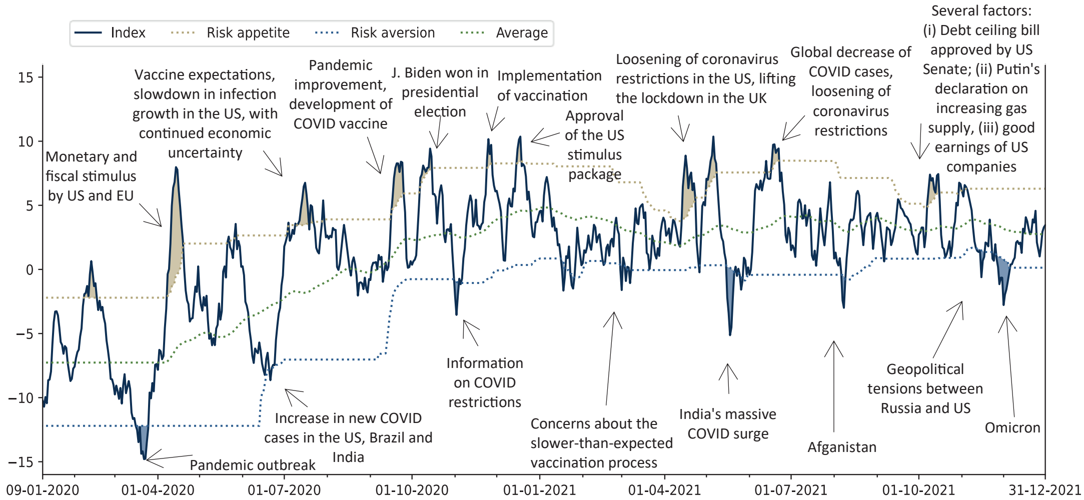
The average informational structure in years:

	2020	2021
COVID	35%	32%
USA	15%	16%
UK	12%	8%
China	9%	8%
Cen. Banks	8%	11%
EU	6%	5%
Commodities	5%	4%
Japan	3%	3%
Russia	2%	4%
India	2%	3%
Saudi Arabia	2%	2%
Brazil	1%	1%
Afganistan	1%	4%

Source: Own work based on Refinitiv data.
Note: Observation values were smoothed in the 21 day window to improve visualization.

Additionally, for a more precise analysis, Figures A.1 and A.2 are presented in Appendix A, showing the frequency of each topic and its share in the analysed period. Furthermore, Figure A.3 depicts the significance of each of the topics, defined as the relative position of the values in the dataset.[4] Such representation of results is helpful in assigning the emotional bias of information to the level of the sentiment gauge (see Figure 3).

Figure 3. Market sentiment index with events highlighted



Source: Own work based on Refinitiv data.

The analysis of the relative importance of an individual topic in the sentiment index (Figure 2) and the importance of their emotional bias for determining the level of the index (Figure 3) leads to the following observations:

- Coronavirus was by far the most important topic throughout the whole period. Importantly, however, the relative frequency of the appearance of COVID headlines indicated that it was not the only news category, nor outranked the other information in aggregate terms (except for the period of the pandemic outbreak in March 2020). It accounted for on average about 1/3 of the thematic structure of market sentiment. By the end of the analysis period, the relative importance of this topic (in information terms) had fallen to around 22%, to its lowest level since its peak in March 2020. The emergence of the Omicron mutation contributed to a renewed increase in the share of this topic in the overall information structure, but only to a moderate extent.
- At the same time, however, the emotional bias of the information on the coronavirus was so strong that it predominantly determined the level of the

[4] In other words, "significance" in this study is understood as the percentage of the topic's rank in the set of all topics. The ranks were assigned based on the frequency of occurrence of each topic on a given day.

sentiment index. In particular, three sub-periods should be indicated: March-April 2020, January-March 2021 and December 2021 (Figure 2). The impact of this factor is clearly visible in the drop in the sentiment index below the 10th percentile (the lowest value for this index was -14.77 on 23March 2020, see Figure 3). In turn, before the outbreak of the pandemic, at the beginning of 2020, the most popular topic, which was also related to COVID at that time, was “China” (see figures in Appendix A). In March and April 2020, the importance of crude oil also increased, caused by the sharp decline in the prices of this commodity due to the global pandemic outbreak.

- After reaching a negative extreme in March 2020, the gradual increase in the sentiment index was observed later on that month. This was related to an improvement in market sentiment due to monetary and fiscal stimulation by the US and EU authorities. Interestingly, in 2020 the importance of information on monetary policy accounted for an average of 8% of all information included in the sentiment index, while in 2021 it went up to 11%.
- The return to the risk aversion area in the index level in June 2020 was caused by the influence of two opposing factors – growing concerns about the re-increase of coronavirus cases (with the epicentre in the USA, Brazil and India), as well as macro data indicating a growing recovery of economic activity. Subsequently, expectations for the development of a vaccine and a gradual slowdown in the growth of infections in the US supported the market moods in July 2020. At the same time, however, a certain increase in new coronavirus cases in Western Europe and the continued economic uncertainty limited a further improvement of sentiment.
- An important topic in the analysed period was the departure of the United Kingdom from the European Union. Interest in Brexit was observed throughout 2020, especially at the end of the year, when the transition period of the UK’s exit from the EU was over. This information accounted for an average of 12% (maximum 24% on December 26, 2020) in the informational structure in 2020. In 2021, the topic of Brexit sharply decreased, and the UK

was mentioned in the news most often in the context of the COVID-19 pandemic.

- One of the most relevant topics concerned the United States. First, there is a sharp increase in the frequency of this topic in the period October-November 2020 (Figures A.1 and A.2), which can be associated with the presidential elections in this country. J. Biden's win contributed to risk-on sentiment (index increase) in November 2020 (Figure 3). The importance of the USA topic increased in 2021 and may be linked to a variety of political and economic developments.
- The topic of India was of greatest importance in the structure of the index in May 2021, when the country saw a massive COVID surge (over 400,000 cases per day). This resulted in a fall of the index significantly below the 10^{th} percentile level and was reflected in the frequency growth of the Indian theme in the news headlines (see figures in Appendix A).
- Another significant drop in the sentiment index was recorded in August 2021, which was related to the situation in Afghanistan (see Figure 3). The rapid impact of this topic is clearly visible both in the graphs in Appendix A and in Figure 2 (purple area).
- The topic of central banks appeared in the headlines steadily throughout the analysed period (with an average share of approximately 8% in 2020 and 11% in 2021, Figure 2). It decreased slightly at the turn of 2020-2021, and began to gain significance in the second half of 2021. This was also indicated by the appearance of words and phrases such as "inflation" and the ECB (in addition to the Fed and "central banks") in this thematic group.
- Russia is mentioned in the news in the context of various events, both political (e.g. sanctions and geopolitical tensions) and economic (e.g. activities of the central bank, oil production). Information on these issues constitutes on average 3% of all analysed headlines (Figure 2). In the second half of 2021, the topic grew in importance, as shown by its more frequent occurrence in the news (Table A.1 in Appendix A).

- The topic "Japan", despite its presence in the information structure (3% on average), had a minimal impact on the level of the sentiment index, and the occasional increase in its share was most often associated with the local waves of the COVID-19 pandemic. Its scope included information on the number of coronavirus infections, Japanese government stimulus packages, the publication of macroeconomic indicators, changes in stock market indices or individual companies.
- The last two topics concerning countries such as Saudi Arabia and Brazil contributed to the least extent to a sentiment movement. Saudi Arabia most often was found in the news in the context of geopolitical tensions in the Middle East (such as the conflict between Saudi Arabia and Iran, participation in the conflict between Israel and Palestine or in the civil war in Yemen) or of OPEC+ meetings. In turn, news about Brazil often concerned oil, coronavirus infections, the weakening of the Brazilian real, actions of the central bank (interventions, interest rate hikes) or anti-president protests.

The above observations focused mainly on the selected topics and do not describe all the changes in the sentiment index in the analysed period. More factors responsible for the index ups and downs are shown in Figure 3.

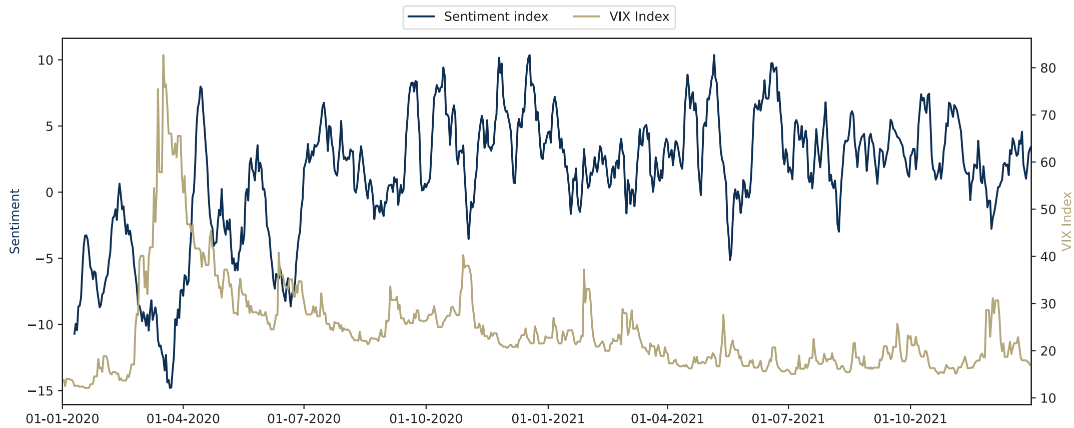
4. Comparison with the VIX index[5]

The sentiment index based on textual data quantification presented in this study is one of the possible measurement methods. Financial practitioners often use the VIX index as a measure of risk perception in financial markets. It estimates the implied volatility of the S&P 500 Index by using underlying option bid/ask quotes. This section presents the relationship between these two sentiment measures and Figure 4 presents their visual comparison. Contrary to the news-based index, the increase in the VIX index is interpreted as a deterioration in market mood, while its decrease is interpreted as sentiment improvement. Therefore, it can be noticed that both indices quite often

[5] Comparison of the constructed index with the sentiment measures such as the University of Michigan Index of Consumer Sentiment or Index of Global Economic Policy Uncertainty (https://www.policyuncertainty.com) is unreliable, due to: (i) different frequency – monthly, (ii) delays in publication (the latter index), (iii) lower degree of linkage with financial markets, more relevant to households in terms of their expenditures (the University of Michigan index).

indicated the same periods of negative sentiment, especially in the case of an intensification of risk factors.

Figure 4. The VIX index (right axis, %) vs the constructed news-based sentiment index (left axis)



Source: VIX index - Bloomberg, sentiment index – own study based on Refinitiv data.

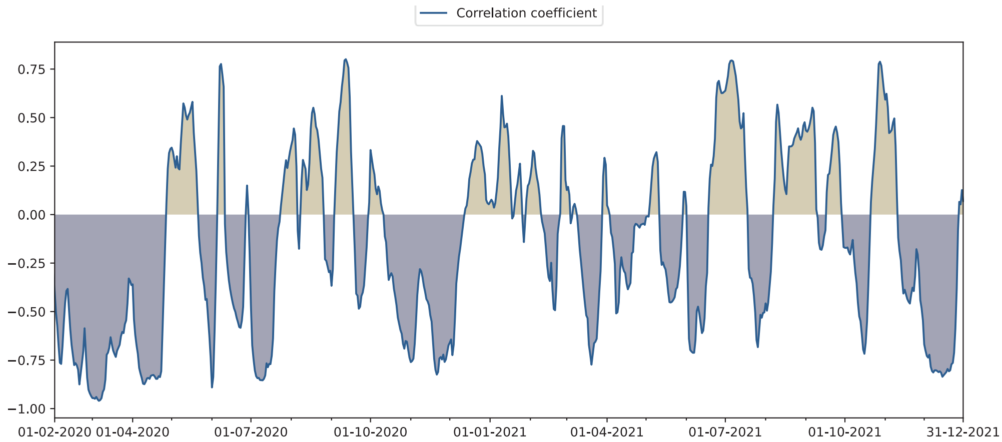
For a more formalized comparison of the two indices, a 21-day rolling Pearson's correlation coefficient was calculated[6] (Figure 5). The window length covers on average almost all working days in a month and seems to be enough to detect data dependency. A shorter rolling period can lead to misleading statistical recaps. According to [28], it is easier to obtain a strong correlation in a small sample rather than in a large one, as well as a strong correlation in small samples is often statistically insignificant. At the same time it was undesirable to use a longer rolling window because of the relatively limited entire dataset. Finally [29] recommends a sample size for Pearson's correlation coefficient close to 25 or more observations.

The periods of negative correlation between these two measures definitely prevailed, as both indices should react in different directions in the case of a change in market sentiment. Sometimes, however, there was a positive correlation between both indices. Nevertheless, the periods of negative correlation were longer than the periods when this measure took positive values. The correlation coefficient throughout the

[6] Pearson's correlation coefficient is used as it is an appropriate way of measuring a linear data relationship, which mostly occurs between two indices in such a small time window.

period amounted to -0.52%. The correlation analysis helped to validify that the presented sentiment index is capturing similar information to that of the VIX index.

Figure 5. Rolling correlation coefficient for VIX and news-based sentiment indices



Source: Own calculations.

The sporadic discrepancy in the market sentiment assessment is arising due to the methodological differences between the two measures, as well as the different interpretation of their levels. First, a decrease or increase in the sentiment news-based index from its average does not necessarily mean a deterioration or improvement in market sentiment, but, for example, that the sum of semantic evaluations of the news headlines turned out to be close to this value. On the other hand, the VIX index is a market measure of the expected volatility of the S&P 500 index and is calculated on the basis of bid/ask price quotes of options on that index. The methodology of the VIX index construction reflects how much the market expects the S&P 500 Index will fluctuate over the next 30 days [24], and not necessarily in every peak and valley of global risk reversion (e.g. the market may assume a lower volatility during the holiday season or expect a higher volatility ahead of the earning season for companies included in the S&P 500). [25] has also showed how to correctly understand the VIX index, placing stress on its property to reflect portfolio insurance price rather than the "investor fear gauge".

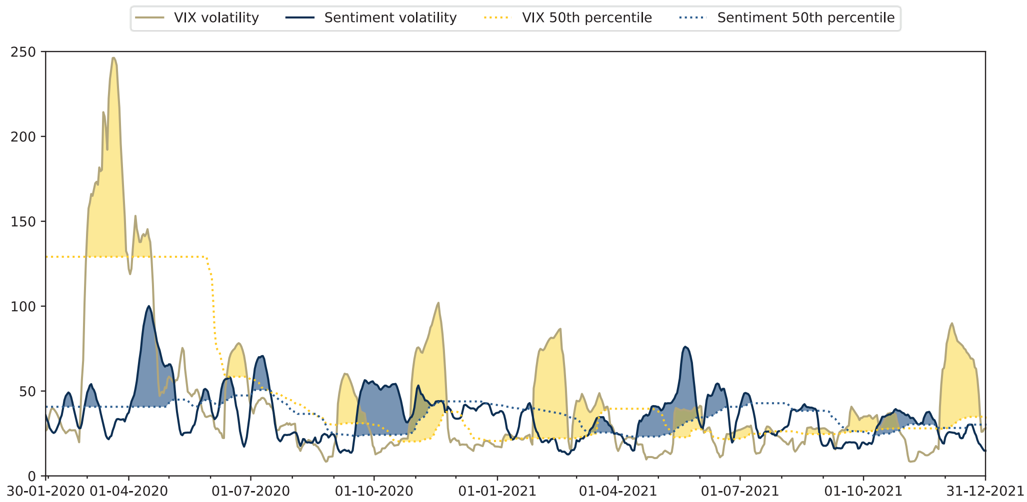
Secondly, the news-based sentiment index is global, while the VIX index – despite its crucial role for global markets – may take into account factors important for the US market to a greater extent. In other words, some events, although they are of a negative international nature, will not have a significant impact on US financial markets. For example, the VIX index remained immune to the withdrawal of US troops from Afghanistan and taking control over the part of the state by the Taliban, while a large volume of news in mid-July and August 2021 contributed to the decline of the constructed index. On the other hand, although the news search results were narrowed down, sometimes information of a local nature appeared among the downloaded headlines, which distorted the results of the sentiment index. Nevertheless, the influence of the latter factor was marginal.

Thirdly, the complexity of natural language can make it very difficult to access the semantics of a text [26], and algorithms for its analysis are sometimes not able to adequately assess the headline (especially of an economic or financial nature) as a human would do. For example, the word “aggressive” is assessed by the algorithm as negative with a valence of -0.6, when normalized -0.153. In combination with a degree modifier “more”, the sentence “more aggressive rate hike” will be rated at -0.23 (for phrases consisting of more than one word only normalized scores can be compared). While it should be evaluated in the context of the economic conditions or of the impact of this decision on other markets globally.

After correlation analysis, the next step was to check the volatility clustering of both indices. Figure 6 demonstrates the historical volatility for VIX and sentiment indices rolled in a 21-day window (the same as for the coefficient of correlation). When the volatility is above the corresponding 50th percentile[7], the area between the lines is coloured. At first glance, the VIX index is characterized by higher volatility. This looks plausible, since the volatility of financial instruments based on market price quotes is, as a rule, higher than that derived from estimation.

[7] The 50th percentile, or the median, represents the mid-point or central tendency of the sample. When the volatility is persistently above the ninety-day median, it indicates the period of higher volatility.

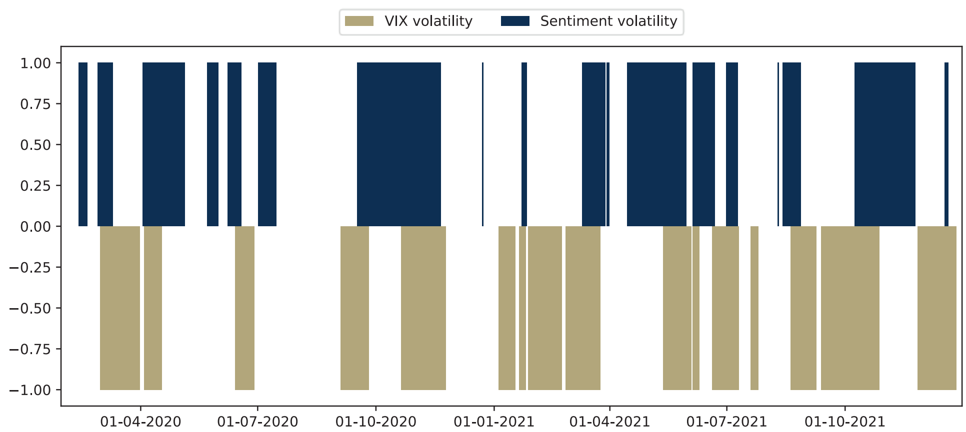
Figure 6. Historical volatility of VIX and news-based sentiment indices (%)



Source: Own work based on Bloomberg and Refinitiv data.
Note: Percentiles are calculated in a rolling window of 90 days.

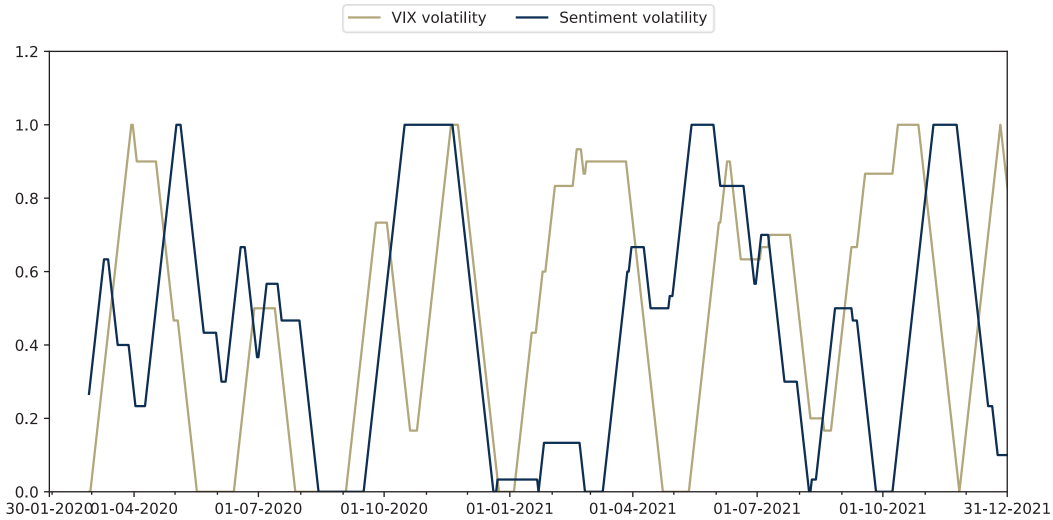
There are also observable periods with a consistent increase in volatility of both sentiment measures. In order to see it more clearly, in Figure 7 the growth of sentiment index volatility above the 50th percentile takes the value 1, while the analogous growth of VIX index volatility takes the value minus 1. At the same time, the share of the volatility that crossed the median in a given period for both indices is depicted in Figure 8.

Figure 7. Clustering of historical volatility of VIX and news-based sentiment indices



Source: Own work based on Bloomberg and Refinitiv data.

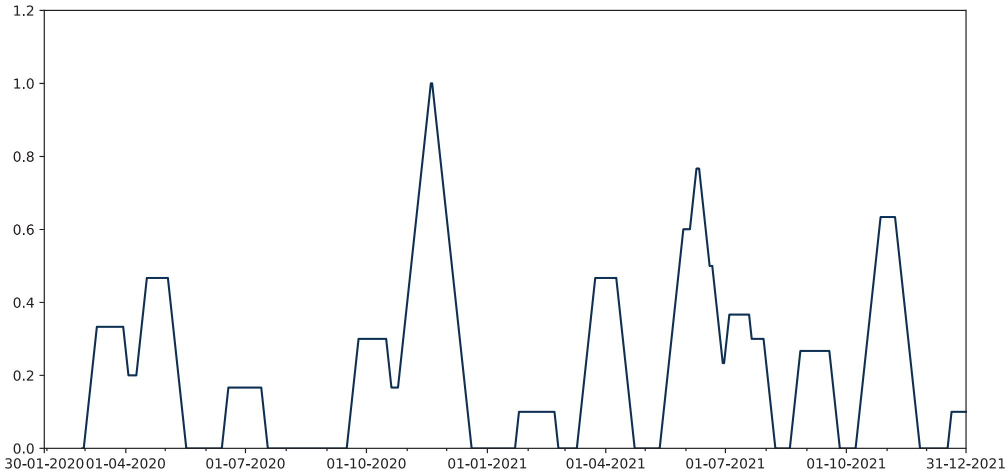
Figure 8. The share of VIX and news-based sentiment indices historical volatility above the 50th percentile (%)



Source: Own work based on Bloomberg and Refinitiv data.
Note: The share was calculated for the rolling period of 30 observations.

One can discover similar periods of higher volatility, with an average share of around 48% for both indices. In order to assess how much these periods are coherent, the share of the mutual volatility growth above the median (at the same period of time) has been calculated (Figure 9). There are unequivocal periods with mutual surge in volatility of both measures, with the highest one for November 2020. The majority of them correspond to negative global developments mostly related to the COVID-19 pandemic: (i) the pandemic outbreak (March – April 2020), (ii) the imposition of pandemic restrictions in November 2020 before the implementation of mass vaccination of the population, (iii) India's surge of coronavirus infections in May 2021 and (iv) geopolitical tensions between Russia and the US following the rapid spread of Omicron.

Figure 9. The share of VIX and news-based sentiment indices historical mutual volatility growth above the 50th percentile (%)



Source: Own work based on Bloomberg and Refinitiv data.
Note: The share was calculated for the rolling period of 30 observations.

To summarize, this study demonstrated that the constructed sentiment index can be used in analytical practice as a sentiment measure independently of the VIX index or other financial instruments, which at the same time allows us to identify the main factors influencing the direction of sentiment.

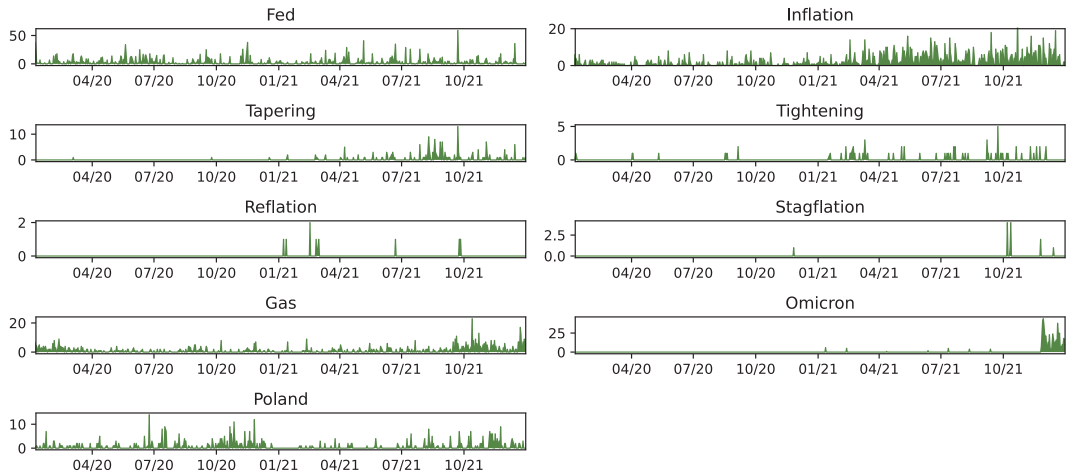
5. Examples of model applications

This part of the study contains the possible examples of additional use of the developed tool for the evaluation and quantification of textual data. At first, the relative importance of current economic issues was briefly examined and then the correlation relationship between the constructed measure of sentiment and the prices of various financial assets was presented.

5.1. Relative importance of current economic issues

Figure 10 shows how frequently the selected thematic issues appeared during the analysed period. The results indicate an intensification in 2021, especially in the second half of the year, of topics related to inflation, monetary tightening in the US and gas prices.

Figure 10. Frequency of appearance of the selected issues (words per day)



Source: Own calculations.

Reflation tendencies can be observed in the first quarter of 2021, and then reflation was mentioned seldom, e.g. in September 2021, when the Fed signalled the possibility of a reduction in asset purchases. In turn, growing concerns about stagflation appeared again in the last quarter of 2021, when energy prices rose and inflationary pressure intensified amid a deteriorating outlook for economic growth.

Poland most often appeared in the headlines in the context of the following events: (i) the presidential election in June 2020, (ii) the increase in new cases of coronavirus in October 2020, (iii) the Poland-Belarus border migration crisis in November 2021, or (iv) issues related to the freezing of recovery funding from the European Commission in December 2021.

5.2. Determining the correlation relationship with assets

This part provides a correlation analysis between the constructed index and various financial assets. For this purpose the following assets were selected: exchange rates of the US dollar against twelve emerging market currencies (for CE-3 currencies the euro crosses were also included), EUR/USD and the DXY dollar index, as well as the stock market MSCI EM index and gold prices.

The correlation coefficients for the majority of the mentioned assets (see Table 1) are statistically significant and its direction in most cases complies with the natural logic of capital flows in financial markets. Positive sentiment, reflected by an increase in the constructed index, translates into a flight of investors from safe assets (weakening of the dollar – decline in the DXY index and rise of EUR/USD exchange rate) to risky ones (strengthening of EM currencies against the dollar, growth of MSCI EM index), see the "General" column of Table 1.

Table 1. Correlation of the sentiment index with various financial assets

Asset	Correlation Coeffcient		
	General	Q1	Q3
DXY	-0,60	0,54	-0,49
EMCI	0,56	0,64	-0,47
EURUSD	0,59	0,49	-0,15
EURPLN	0,40	-0,19	-0,64
EURCZK	-0,12	0,51	-0,57
EURHUF	0,48	-0,32	-0,58
USDPLN	-0,34	0,37	-0,48
USDCZK	-0,12	0,51	-0,79
USDHUF	-0,28	0,48	-0,45
USDRUB	0,34	0,66	-0,46
USDTRY	0,34	-0,45	0,05
USDUAH	0,48	0,45	-0,65
USDKRW	-0,50	-0,19	-0,60
USDMXN	-0,24	0,48	-0,88
USDRON	-0,50	0,38	-0,55
USDINR	0,04	-0,16	-0,31
USDTHB	-0,02	0,53	-0,20
USDCNY	-0,55	0,05	-0,70
MSCIEM	0,56	0,64	-0,47
Gold	0,58	0,31	-0,62
Brent	0,36	0,68	-0,24

Source: Own calculations.

Additionally, the quartile correlations for the sentiment index and asset returns were calculated. The results obtained for the first quartiles (Q1) indicate that the movement of the sentiment index near or approaching the risk-off region – which reflects the deterioration of market sentiment – puts pressure on EM currencies (growth of DXY index, depreciation of EM currencies vs USD, rise of gold prices). On the other hand, the correlation for the third quartiles (Q3) – an improvement of market sentiment and the index approaching the risk-on region – is characterized by an appreciation of emerging market currencies (drop in EM FX crosses vs USD) and usually declines of gold prices.

In relation to the CE-3 exchange rates against the dollar, there is a negative correlation in the case of the overall and the third quartile correlation, and a positive correlation in the case of the first quartile, which is consistent with most other EM currencies.

In turn, the results for the zloty and forint crosses against the euro in the case of the overall and Q1 correlation can be assessed through the prism of the euro strengthening in the conditions of the dollar depreciation, or as a result of local factors (without significant contribution to the sentiment index).

The largest correlation coefficient (overall) was found between the sentiment index and the dollar index, EMCI index, the gold prices and the MSCI EM index. Interestingly, in the case of risk aversion periods (Q1), the largest correlation (over 64%) was recorded for the EMCI index, USD/RUB exchange rate and MSCI EM index. On the other hand, when the risk appetite was rising (Q3), significantly more assets recorded a correlation below -0.6, and the most correlated with the sentiment index turned out to be crosses USD/MXN, USD/CZK, USD/CNY, EUR/PLN, USD/UAH and the gold price.

6. Conclusion

This study presents a method for financial market sentiment measurement based on analysing the news headlines using a natural language processing tool. For this purpose over 124,000 news items were downloaded from Eikon Refinitiv and algorithmically evaluated, resulting in a daily sentiment index for 2020-2021.

As part of the validation process, the constructed index was compared with the VIX index, which is often used in analysts' practice and interpreted as a proxy for market sentiment. Both indices often showed the same shift in the direction of sentiment in response to major global events, which was confirmed by the occurrence of periods of strong correlation. Moreover, the periods of mutual increase in volatility of both indices were also revealed. The temporal divergence of sentiment assessment by both measures resulted mainly from the different methodology of indices construction.

Potential improvement can be obtained by applying a lexicon dedicated to the economic and financial domain.

The analysed period is specific as it coincides with the outbreak and development of the COVID-19 pandemic. This was reflected in the research results that highlight coronavirus as the dominant topic throughout the dataset. The coronavirus pandemic, directly or indirectly (e.g. through the actions taken by governments and central banks), is responsible for the vast majority of significant increases or decreases in the sentiment index (which was also the case for the VIX index in 2020-2021). As a result, the influence of other factors was much smaller. The exceptions were such events as J. Biden's victory in the US presidential election, the situation in Afghanistan, the geopolitical tensions between Russia and the USA, and the declaration of the Russian president on increasing gas supply.

The algorithm designed in this study not only extracts market sentiment, but also identifies the factor contributing the most to the sentiment and the moment when it begins to gain significance for the sentiment (e.g. the appearance of the words about vaccines under the COVID topic or the word "inflation" under the topic of central banks). Thus, the tool that was developed allows us to trace the impact of any thematic issue and set the moment of its intensification or fading. Moreover, a significant correlation was demonstrated between the sentiment index and selected financial assets, and its direction was consistent with the natural logic of capital flows in financial markets.

References

[1] M. González-Sánchez & M. Morales de Vega, "(2021) Influence of Bloomberg's Investor Sentiment Index: Evidence from European Union Financial Sector", *Mathematics,* p. 297, 2021.

[2] L. Smales, "The importance of fear: investor sentiment and stock market returns", *Applied Economics, 49:34,* pp. 3395-3421, 2017.

[3] K. Seki, Y. Ikuta & Y. Matsubayashi, "News-based business sentiment and its properties as an economic index", *Information Processing & Management,* 59, 2022.

[4] P. Carr, "Why is VIX a fear gauge?", *Risk and Decision Analysis,* pp. 1-7, 2017.

[5] I. Shaikh & P. Padhi, "The implied volatility index: Is 'investor fear gauge' or 'forward-looking'?", *Borsa Istanbul Review,* 15, pp. 44-52, 2015.

[6] A. D'Andrea, F. Ferri, P. Grifoni & T. Guzzo, "Approaches, Tools and Applications for Sentiment Analysis Implementation", *International Journal of Computer Applications,* pp. 26-33, 2015.

[7] X. Li, P. Wu, W. Wang, "Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong", *Information Processing &* Management,, 57, 2020.

[8] A. Derakhshan & H. Beigy, "Sentiment analysis on stock social media for stock price movement prediction", *Engineering Applications of Artificial Intelligence,* 85, nr 569-578, 2019.

[9] J. Bollen, H. Mao & X. Zeng, "Twitter mood predicts the stock market", *Journal of Computational Science,* 2, pp. 1-8, 2011.

[10] Y. Qian, Z. Li & H. Yuan, "On exploring the impact of users' bullish-bearish tendencies in online community on the stock market", *Information Processing & Management,* 57, 2020.

[11] X. Zhang, Y. Zhang, S. Wang, Y. Yao, B. Fang & P. S. Yu, "Improving stock market prediction via heterogeneous information fusion", *Knowledge-Based Systems,* 143, pp. 236-247, 2018.

[12] C. Khoo & S. Johnkhan, "Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons"" *Journal of Information Science,* pp. 491-511, 2018.

[13] M. Taboada, J. Brooke, M. Tofiloski, K. Voll & M. Stede, "Lexicon-based methods for sentiment analysis", *Computational linguistics,* 2, pp. 267-307, 2011.

[14] D. J. Garcia, "Sentiment during recessions", *Finance,* 3, pp. 1267-1300, 2013.

[15] A. Samuels & J. Mcgonical, "News Sentiment Analysis", 2020.

[16] L. V. Avanço & M. d. G. V. Nunes, "Lexicon-Based Sentiment Analysis for Reviews of Products in Brazilian Portuguese", w *Brazilian Conference on Intelligent Systems*, 2014.

[17] C. Hutto & E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", w *Proceedings of the 8th International Conference on Weblogs and Social Media*, 2015.

[18] T. Dijk, "News As Discourse", *Routledge Communication Series,* 2013.

[19] K. Goshima, D. Takahashi & T. Yamada, "Construction of business news index by natural language processing and its application to volatility prediction", *Financial Research,* 2019.

[20] J. Aukia, J. Heimonen, T. Pahikkala & S. T. , "Automated quantification of Reuters news using a receiver operating characteristic curve analysis: The Western media image of China", University of Turku, Finland., 2017.

[21] V. Bonta, N. Kumaresh & N. Janardhan, „A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis," *Asian Journal of Computer Science and Technology,* 8, pp. 1-6, 2019.

[22] K. R. Pano T, "A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19", *Big Data and Cognitive Computing,* 4, p. 33, 2020.

[23] A. Sadia, F. Khan & F. Bashir, "An overview of lexicon-based approach for sentiment analysis", w *3rd International Electrical Engineering Conference (IEEC 2018)*, 2018.

[24] "Cboe Volatility Index," [Online]: https://cdn.cboe.com/api/global/us_indices/governance/VIX_Methodology.pdf

[25] R. Whaley, "Understanding VIX", *Journal of Portfolio Management,* pp. 98-105, 2009.

[26] S. Bird, E. Klein & E. Loper, Natural Language Processing with Python, O'Reilly Media Inc., 2009.

[27] N. Mukhtar, A. M. Khan & N. Chiragh, "Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains", *Telematics and Informatics,* 35, pp. 2173-2183, 2018.

[28] D. de Vaus, Analyzing Social Science Data: 50 Key Problems in Data Analysis, SAGE Publications Ltd, 2002.

[29] F.N. David, Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples, Cambridge: Cambridge University Press, 1938.

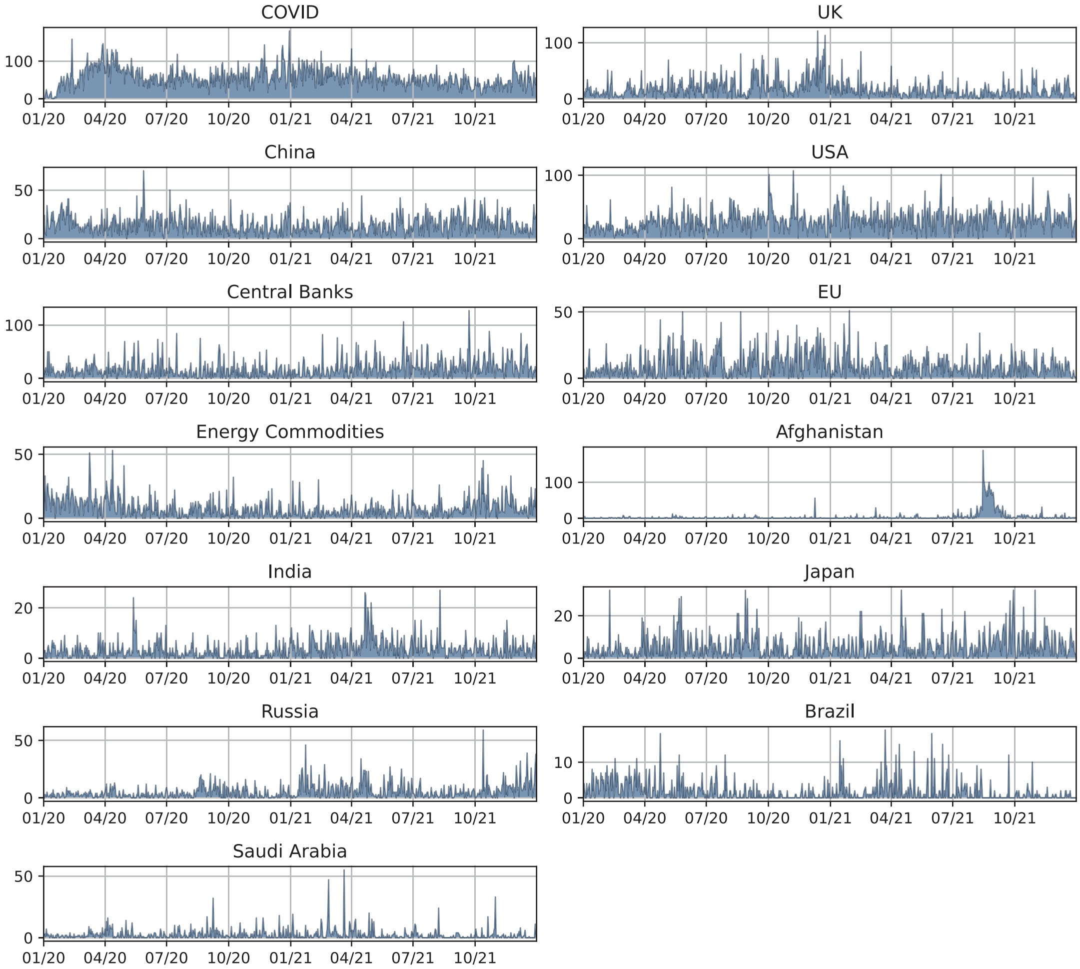
Annex A

Table A.1. The most popular topics in the analysed period

H1-2020		H2-2020		H1-2021		H2-2021	
topic	**words**	**topic**	**words**	**topic**	**words**	**topic**	**words**
COVID	*covid (1043)*	COVID	*covid (3552)*	COVID	*covid (4051)*	COVID	*covid (3190)*
	coronavirus (6102)		*coronavirus (2866)*		*coronavirus (1056)*		*vaccine (1364)*
	health (927)		*vaccine (1304)*		*vaccine (2993)*		*coronavirus (761)*
	virus (1166)		*health (787)*		*health (711)*		*cases (673)*
	lockdown (489)		*pandemic (433)*		*dose (1493)*		*omikron (594)*
					astrazeneka (548)		*health (523)*
UK	*uk (2409)*	UK	*uk (3494)*				
			brexit (1341)	UK	*uk (2246)*	China	*china (2716)*
China	*china (2788)*	China	*china (2244)*			USA	*biden (1793)*
				China	*china (1972)*		*white house (1098)*
USA	*trump (1933)*	USA	*trump (2096)*				*us(863)*
	us (738)		*biden (1028)*	USA	*biden (1960)*		
			white house (1154)		*white house (1670)*	UK	*uk (1957)*
Central Banks	*central bank (1153)*				*trump (558)*		
	fed (694)	EU	*eu (1676)*		*us (780)*	Afghanistan	*afghanistan (835)*
			german (534)				*taliban (674)*
Oil	*oil (1495)*			Central Banks	*fed (520)*		*kabul (426)*
		Central Banks	*central bank (986)*		*central bank (1159)*		*afghan (467)*
EU	*german (1047)*		*fed (430)*		*inflation (520)*		
	eu (813)					Central Banks	*central bank (1992)*
		Japan	*japan (931)*	EU	*eu (996)*		*ecb (665)*
Japan	*japan (806)*				*german (509)*		*fed (562)*
		Oil	*oil (647)*				*inflation (698)*
India	*indie (526)*			Indie	*indie (866)*		
		Russia	*russia (554)*			Russia	*russia (1360)*
Brazil	*brazil (524)*			Japan	*japan (850)*		
		Saudi Arabia	*saudi (505)*			Energy commodities	*oil (735)*
				Russia	*russia (797)*		*energy (489)*
					kremlin (467)		*gas (406)*
				Saudi Arabia	*saudi (574)*	EU	*german (541)*
							eu (748)
				Oil	*oil (571)*		
						Japan	*japan (1121)*
						India	*India (600)*

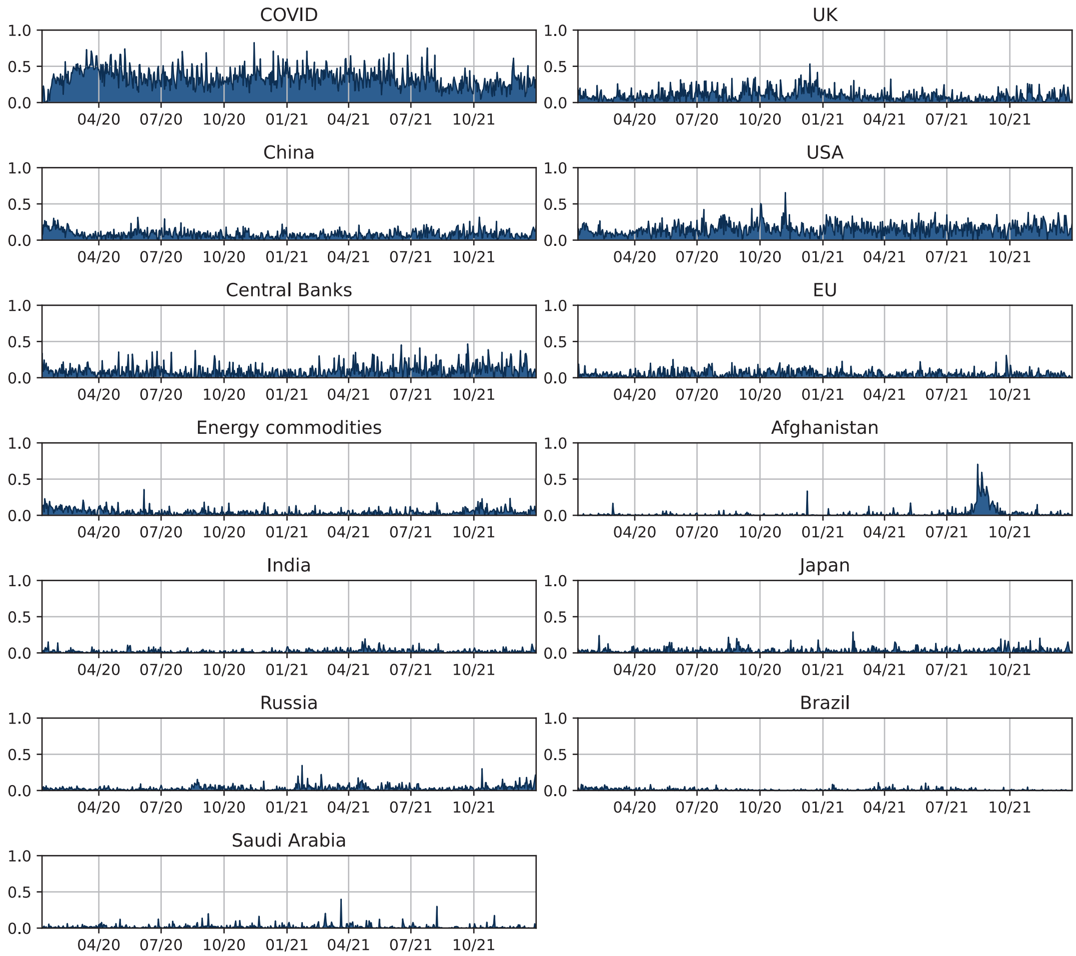
Note: Words found in headlines are written in italics, and the frequency of their occurrence in the headlines is given in parentheses.
Source: Own calculations based on Refinitiv data.

Figure A.1. Frequency of topics per day in the analysed period



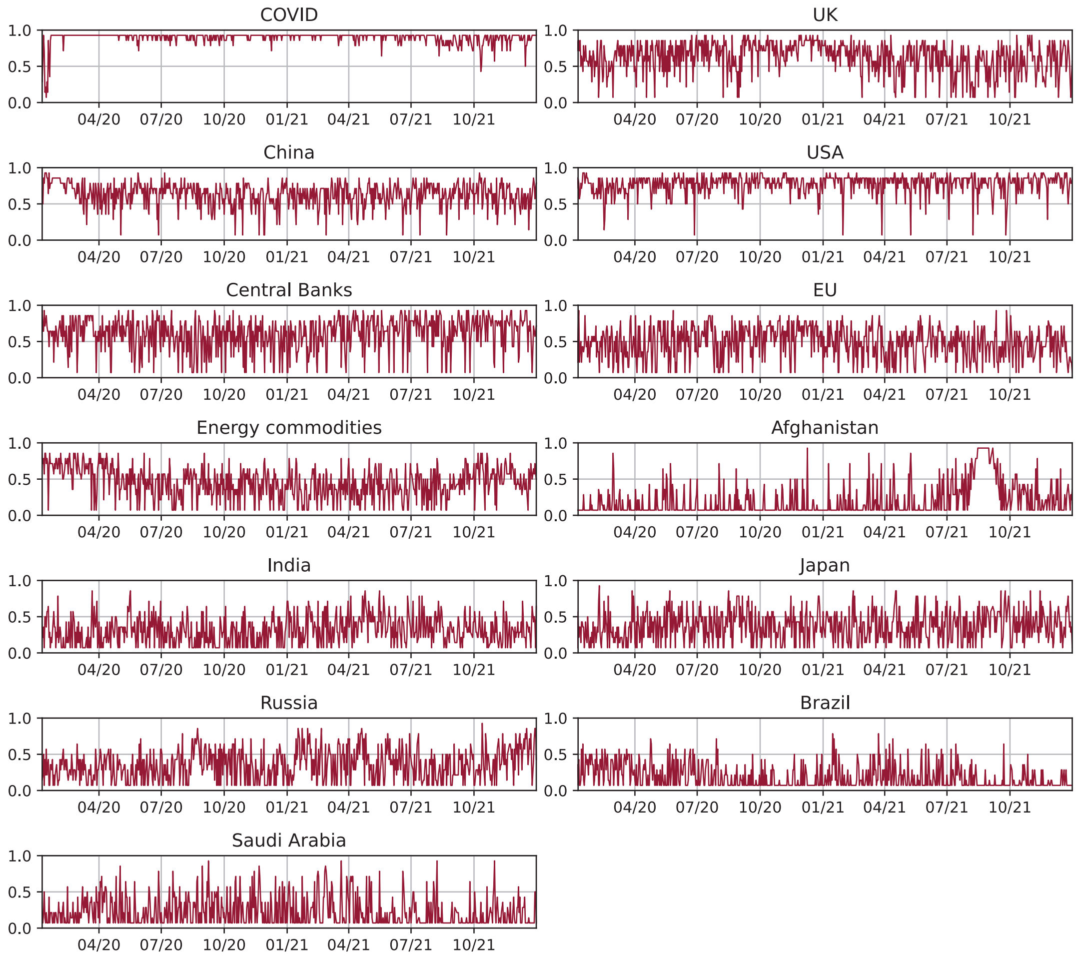
Source: Own calculations based on Refinitiv data.

Figure A.2. Share of topics per day in the analysed period



Source: Own calculations based on Refinitiv data.

Figure A.3. Significance of topics per day in the analysed period



Source: Own calculations based on Refinitiv data.

www.nbp.pl