

1 A New High Frequency, News Based, Indicator of Macroeconomic Activity

1.1 Introduction

The information encoded in text has been recently used in empirical economics research as a complement to the more structured macroeconomic and financial data traditionally used (Gentzkow, Kelly, & Taddy (2019)). Text selected from news, social media, reports and speeches contains “soft” information missing in the quantifiable variables. Unlike most of the headline macroeconomic data that are published at a relatively low frequency and for which past observation periods are revised as more accurate estimates become available, text such as news articles arrives daily and is not revised. These advantages make data extracted from text an ideal candidate to build more accurate nowcasting models about aggregate economic activity in real time.

In this paper I propose a text-based approach to create a high-frequency news-based sentiment indicator (NBSI) regarding aggregate economic conditions. Following dictionary methods in the Natural Language Processing (NLP) literature, NBSI is established from lead paragraphs of news articles that are related to economic activity. The corpus consists of 410,601 economic and financial news articles published at a daily frequency in the Wall Street Journal from January 1991 to March 2020.

To evaluate the predictive ability of the NBSI, this paper provides two applications. In the first application, I use NBSI to track a wide range of economic activity measures, such as consumer confidence index, total nonfarm payroll, unemployment rate, industrial production, manufacturing and industries trade sales, and initial claims to unemployment insurance benefits. In the second application, I incorporate NBSI in the business cycle phases nowcasting models developed in Chapter 3, and estimate the contribution of NBSI to identify U.S. recessions in real time.

This paper is organized as follows. In section 2 I review related literature. In section 3 I present a text-based approach to create a high-frequency news-based sentiment indicator (NBSI) regarding aggregate economic conditions. Section 4 presents preliminary results. Section 5 lays out a future

plan for this study.

1.2 Literature Review

Choi & Varian (2012) use Google Trends search engine data, which is a real-time daily and weekly index of the volume of queries that users enter into Google. The authors use Google Trend data from January 2004 to July 2011 to forecast economic indicators including initial claims for unemployment insurance and consumer confidence, and find that simple seasonal auto-regression models that include relevant Google Trends variables tend to outperform models that exclude these predictors. Choi & Varian (2012) concludes that Google Trends may help in predicting the present.

Kelly, Manela, & Moreira (2019) model the inclusion and repetition of words of some attributes at a monthly frequency in the Wall Street Journal front page news articles. The authors find out that the text from January 1990 to December 2010 documents contains additional information that is useful for forecasting monthly macroeconomic indicators beyond that of the benchmark dynamic factor model model.

Shapiro, Sudhof, & Wilson (2020) derive a daily and monthly time-series sentiment from economic and financial newspaper articles from January 1980 to April 2015 with dictionary methods. The authors demonstrate that news sentiment is predictive of consumer sentiment measured using surveys. The authors also find that positive sentiment shocks increase consumption, output, and interest rates and dampen inflation.

Bybee, Kelly, Manela, & Xiu (2020) use topic models to estimate a daily time-series news attention estimates to each theme from full text content of Wall Street Journal articles from January 1984 to June 2017. The authors show that the topic-based estimates accurately track a broad range of economic variables such as aggregate output (industrial production) and employment (non-farm payrolls). The authors also find that the estimate have incremental forecasting power above and beyond standard numerical predictors.

This paper makes contributions by demonstrating sentiment analysis that can be used in economics research to extract textual data from nontraditional data sources. The news-based sentiment index

should pick up wider range of information regarding economic activity compared with the survey-based consumer sentiment index. Therefore NBSI that arrives at a daily frequency should be a candidate as supplement to macroeconomic indicators that is available at lower frequencies in nowcasting macroeconomic activity. The second contribution of this paper is to explore the power of NBSI in nowcasting a wide variety of macroeconomic variables as well as for high-frequency monitoring of business cycle phases.

1.3 Methodology

From the Factiva Database, I collect a large sample of lead paragraphs of 410,601 articles published in The Wall Street Journal and The Wall Street Journal Online in the United States from April 2, 1991 to April 30, 2020 that have the following subjects:

- News about commodity, debt, bond, equity, money and currency markets.
- Analysts' comments or recommendations about corporates and industries.
- Economic performance or indicators, government finance, monetary policy, trade or external payments.

The document downloaded from the database is a Rich Text Format (RTF) file, which contains indexes for custom search fields, such as indexes for column name (CLM), section name (SE), headline (HD), author's name (BY), word count, publication date (PD) and time (ED), source name (SN), lead paragraph (LP), language code (LA), subject code (NS), region code (RE), publisher name (PUB), accession number (AN). For the purpose of this topic, I only extract lead paragraphs and publication dates by detecting the presence of the pattern of the index in each piece of news articles. Then I match the lead paragraph with the publication date such that the text is represented at each point in time. Table 1 is an example of this sample.

Before applying statistical methods, the raw text is extracted as a manageable high-dimensional numerical array using process criteria in the following order:

- Exclude lead paragraphs with less than 50 characters.

Table 1: Text Example

Date	Lead Paragraph
03-28-2006	Jon Corzine, New Jersey’s new Governor, isn’t the first politician not to follow through on a campaign promise. But rarely is such dishonesty later presented as a virtue. The question for voters to contemplate is whether this is also an indication of what to expect if Democrats gain control of Congress in November. Mr. Corzine won the Trenton statehouse last year by running as a tax cutter who’d raise property tax rebates by 40% over four years. "I’m not considering raising taxes. It’s not on my agenda. We have a very high-rate tax structure. I’m not considering it," the then-U.S. Senator had vowed in October.
03-29-2006	A growing number of homeowners, riding the crest of the real-estate boom, are getting hit by an unpleasant surprise when they sell: a hefty tax bill. This development stems from a 1997 law that Treasury Department officials said at the time would eliminate capital-gains taxes for nearly everyone selling their primary residence. Under that law, most married couples who file jointly can exclude as much as \$500,000 of their gain. For most singles, the limit is \$250,000.
03-29-2006	The airline most often viewed as the strongest, healthiest and best-run of the pack may be one of the weaker bets for investors hoping to profit from a budding turnaround in the industry. A strengthening market has sent some beleaguered airline stocks soaring in recent months, but shares in industry stalwart Southwest Airlines have been more sluggish. In the past year, the stock of American Airlines parent AMR Corp. has risen 168% on the New York Stock Exchange. Rival hub-and-spoke carrier Continental Airlines saw its stock price jump 142% on the Big Board during the 12-month period. But the same tide of good news has lifted Southwest’s stock just 25% on the NYSE, though that gain still handily outpaced the Dow Jones Industrial Average, which was up 7% over the same period.

- Tokenize the text into individual words and transform it to a tidy data structure.
- Strip stopwords such as ‘a’, ‘the’, ‘to’, ‘for’ out of token lists.
- Replace words with their root such that “economic”, “economics”, and “economically” are all replaced by the stem “economic”.
- Assign “positive” or “negative” sentiment to words based on a general-purpose “Bing” lexicon from Bing Liu and collaborators, which categorizes words in a binary fashion.
- Reverse sentiment of words preceded by negation words such as “no”, “not”, “never”, “without”.
- Measure sentiment scores for each lead paragraph with $\frac{n_{pos} - n_{neg}}{n_{words}}$, where n_{pos} , n_{neg} , and n_{words} represent separately the number of positive, negative, and total words in each lead paragraph.
- Average over articles published on the same day to obtain the daily sentiment index.
- Average over daily sentiment to obtain the weekly sentiment index.
- Decompose the weekly sentiment into trend, seasonality and remainder with a filtering procedure developed by Cleveland, Cleveland, E. McRae, & Irma Terpenning (1990), and remove seasonality from weekly series.

1.4 Preliminary Results

The weekly NBSI from April 02, 1991 (the fourteenth week of 1991) to April 30, 2020 (the eighteenth week of 2020) is shown in Figure 1. Shaded areas indicate U.S. recessions. The index drops sharply before the start of the two recessions in the sample period. This suggests that the index might be a leading indicator with respect to recessions and might be used to nowcast or even forecast recessions.

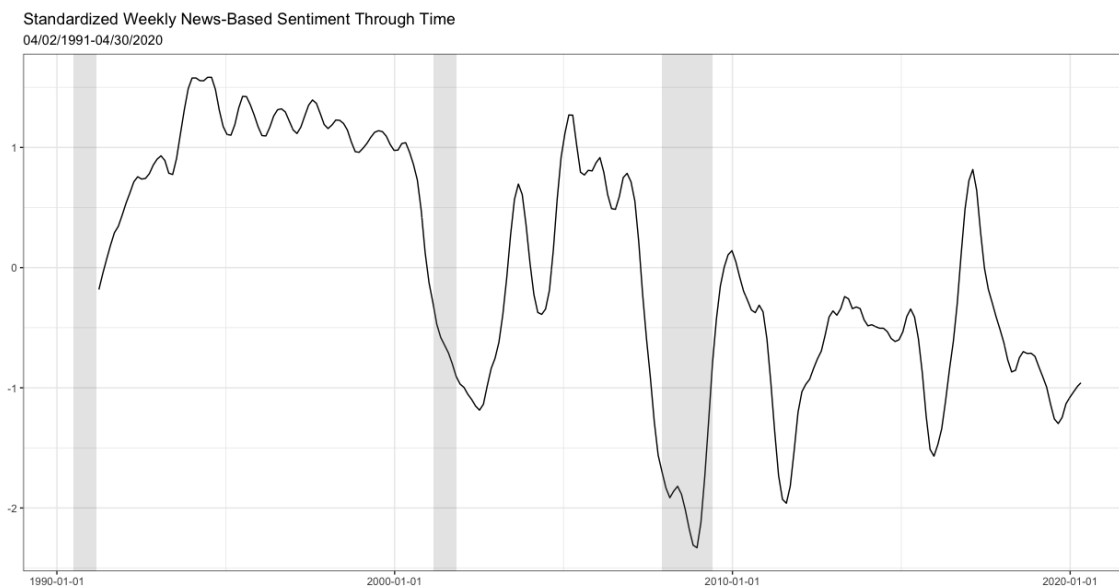


Figure 1: Weekly News-Based Sentiment Index (April 1991- April 2020)

Figure 2 shows the standardized weekly NBSI for January through April in 2020. It presents how NBSI picked up the bad economic outcomes in March 2020. It is also interesting that NBSI falls in February, prior to the real problems starting in the United States.

Next I study the correlation between the weekly NBSI and the weekly initial claims to unemployment insurance (ICSA). Due to the economic impact of Covid-19, ICSA has skyrocketed since late March 2020. Including these values makes the movement of NBSI and ICSA prior to late March less legible; therefore, Figure 3 only shows fluctuations of the weekly NBSI and ICSA as of March 14, 2020. How NBSI tracks ICSA contemporaneously is shown in the left panel and how 4-week moving average of NBSI tracks 4-week moving average of ICSA is shown in the right panel.

Table 2 shows Pearson correlations between weekly variables starting from April 02, 1991 (the

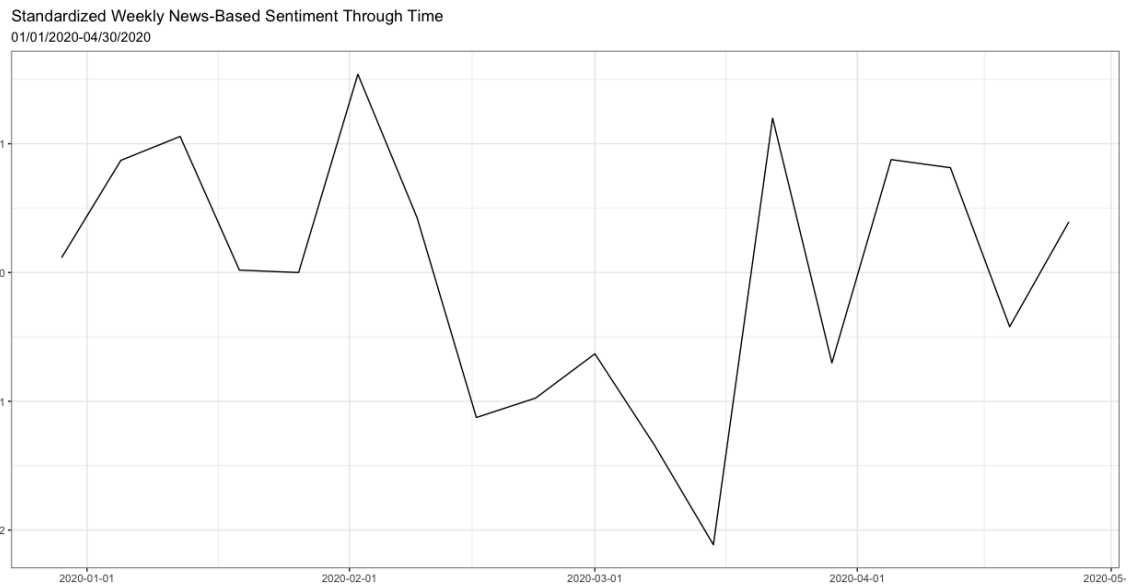


Figure 2: Weekly News-Based Sentiment Index in 2020

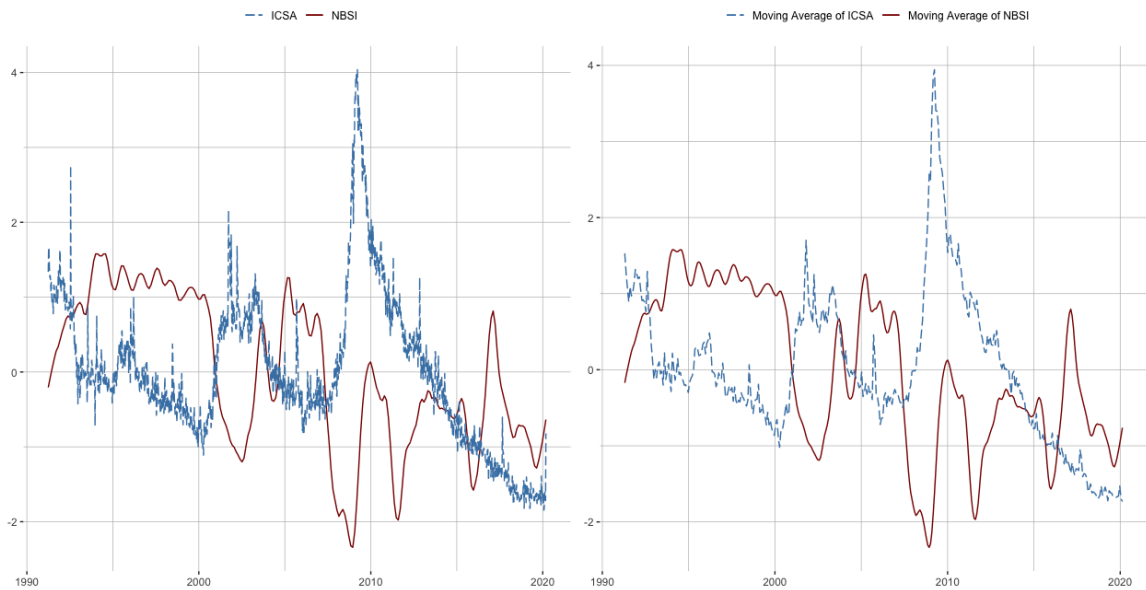


Figure 3: Weekly News-Based Sentiment Index and Initial Claims to Unemployment Insurance

Table 2: Correlation between Weekly NBSI and ICSA

	NBSI	Moving Average of NBSI	Lagged NBSI	Moving Average of Lagged NBSI
ICSA	-0.099(<.001)		-0.101(<.001)	
Moving Average of ICSA		-0.097(.060)		-0.099(.055)

fourteenth week of 1991) to April 30, 2020 (the eighteenth week of 2020). P-values are in parenthesis. There is sufficient evidence to conclude that there is a significant negative linear relationship between the weekly NBSI and contemporaneous ICSA at a significance level of 0.01. There is a significant negative linear relationship between NBSI lagged 1-week and ICSA at a significance level of 0.01. There is a significant negative linear relationship between 4-week moving average of NBSI and 4-week moving average of ICSA at a significance level of 0.1. There is a significant negative linear relationship between 4-week moving average of NBSI lagged 1-week and 4-week moving average of ICSA at a significance level of 0.1. All Pearson correlations have correct signs and fairly large magnitude.

Next, I analyze the ability of the monthly NBSI to track fluctuations in the economic activity. Monthly macroeconomic variables I use include unemployment rate, total nonfarm payroll, industrial production, real manufacturing and industries sales, and University of Michigan consumer sentiment. Figure 4 shows the movement between the monthly NBSI and monthly macroeconomic variables from April 1991 to April 2020. Contemporaneous Pearson correlations and p-values between the monthly NBSI and macroeconomic variables are shown in Table 3.

There is sufficient evidence to conclude that there is a significant negative linear relationship between the monthly NBSI and unemployment rate at a significance level of 0.05. There is also a significant negative linear relationship between the monthly NBSI and the change of unemployment rate at a significance level of 0.01. There are positive linear relationships between the monthly NBSI and all other macroeconomic variables at a significance level of 0.01. All Pearson correlations have correct signs and fairly large magnitude.

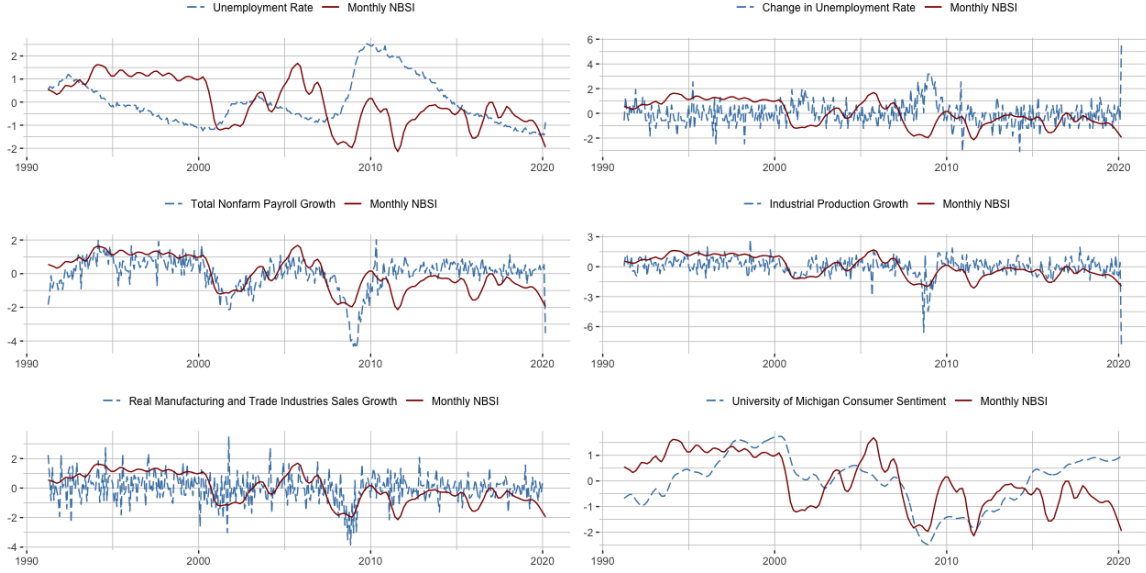


Figure 4: Monthly News-Based Sentiment Index and Macroeconomic Variables

Table 3: Contemporaneous Correlation between Monthly NBSI and Macroeconomic Variables

	Pearson Correlation	P-value
Unemployment Rate	-0.1347044	0.0118927
Change in Unemployment Rate	-0.2390528	0.0000065
Payroll Employment Growth	0.4907129	0.0000065
Industrial Production Growth	0.3747335	0.0000065
Real Manufacturing and Trade Sales Growth	0.2464932	0.0000065
University of Michigan Consumer Sentiment	0.4682164	0.0000065

1.5 Future Directions

There are shortcomings of dictionary methods. Words such as “federal funds rate”, “increase”, and “decrease” are not categorized, and words such as “unemployment” and “tax” are categorized as “negative”. In cases such as “the federal funds rate increase”, the dictionary method fails to contribute to the negative score; and in cases such as “the unemployment rate decreases”, the dictionary method wrongly contribute to the negative score. Sentiment index constructed from dictionary methods is the average of valence scores of each word within an article. It fails to account for the context of a word, and could wrongly decipher how a word was used within a sentence.

In order to account for the textual characteristics of words, “off-the-shelf” supervised machine learning techniques that are highly efficient in understanding the context and relation between words could be a potential solution. To achieve this, I plan to manually rate the sentiment of a fraction of news articles, which is taken as the “true” sentiment and used as training set. Then I plan to use Bidirectional Encoder Representations from Transformers (BERT), a more recently developed learning model developed at Google, to model context and sequential information in text.

For this paper, I plan to use the news-based sentiment index to the following applications:

- Exploring nowcasting power of NBSI for monthly macroeconomic variables used in the models developed in Chapter 3 and updating the nowcast as the month evolves
- Exploring nowcasting power for business cycle phases, within the framework proposed in Chapter 3

Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. 2020. The structure of economic news. *NBER Working Paper Series*, (26648).

Choi, H., & Varian, H. 2012. Predicting the present with google trends. *The Economic Record*, 88: 2–9.

Cleveland, R. B., Cleveland, W. S., E.McRae, J., & IrmaTerpenning. 1990. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 3–33.

Gentzkow, M., Kelly, B., & Taddy, M. 2019. Text as data. *Journal of Economic Literature*,

57(3): 535–574.

Kelly, B. T., Manela, A., & Moreira, A. 2019. Text selection. *NBER Working Paper Series*, (26517).

Shapiro, A. H., Sudhof, M., & Wilson, D. 2020. Measuring news sentiment. *Federal Reserve Bank of San Francisco Working Paper*, (2017-01).