The University of Chicago

Operationalizing the New York Times to Forecast Economic Indicators

By

Christopher Maurice

June 2022

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Jon Clindaniel
Preceptor: Sanja Miklin

1 Introduction

On October 24th, 1929, or Black Thursday, the Dow Jones Industrial Average fell 11 percent. The subsequent performance of the stock market was so dismal that historians continued to use the same name when referencing the following days: on Black Monday, or October 28th, the Dow dropped 13 percent; Black Tuesday, October 29th, another 12-point drop. By mid-November, the Dow lost almost half of its value (Richardson et al., 2013). The initial drop on Black Thursday is considered the start of the Great Depression; the rest is history: a decline in consumer spending, double-digit unemployment rate, plummeting wages, bank runs, lines for food, and foreclosures. Yet, in June 1930, while the economy was still on thin ice, President Hoover declared to a group advocating for a public works program, "Gentlemen, you have come six weeks too late. The depression is over" (Further, 1933). President Hoover believed the unemployment rate was exaggerated, stating, "Its peak had been reached and passed. The tide had turned," citing reports he received, noting the gentlemen in the room lacked access to such reports. Hoover concluded, "Yes, [The United States] is now to drift peacefully, if slowly, back to good times" (Pinchot, 1931). It would not take long for Hoover to be proven wrong, for the signs of a turnaround did not appear until late 1933. Whether Hoover sought to downplay the economic crisis to better position himself for his re-election in 1932 is beyond the scope of this paper. What is a central focus of this paper, however, is the use of economic data to assess the health of the United States economy.

What stifled the post-Black-Thursday recovery was the theories of classical economic thinkers, including Adam Smith, could not explain what was happening nor provide solutions to fix the problem. During this time, the economist John Maynard Keynes developed a new branch of economies, now known as macroeconomics, that pinpointed the root of this economic

downturn and guided the intervention of policymakers. President Roosevelt relied heavily on Keynes's theories to justify his New Deal policies that eventually led the United States out of the worst economic crisis. In February 1936, Keynes published the entirety of his groundbreaking theories in *The General Theory of Employment, Interest, and Money*, laying out the complex web of economic interactions influencing the economy. At the center of this web is the idea that consumption is the leading determinant of economic activity, and, as a result, "Opportunities for employment are necessarily limited by the extent of aggregate demand. Aggregate demand can be derived only from present consumption or from present provision for future consumption" (Keynes, 1936). Keynes forever changed the role of policymakers in the economy— moving from a laissez-faire attitude to a hands-on approach. Keynesian economics explicitly calls for the government to play a greater role in stimulating consumption in contractionary periods, yet proper action requires data to understand what is happening on the ground and untangle the intricate relationship between consumers, producers, and government. Monitoring the health of the economy does not begin and end with consumption. For example, if individuals or businesses think economic conditions will worsen, they will decrease their consumption and increase savings to weather what they believe will be an impending downturn. As Keynes puts it, "An act of individual saving means—so to speak—a decision not to have dinner today. But it does not necessitate a decision to have dinner or to buy a pair of boots a week hence or a year hence or to consume any specified thing at any specified date. Thus, it depresses the business of preparing today's dinner without stimulating the business of making ready for some future act of consumption… it is a net diminution of such demand," meaning on an aggregate scale, preemptively saving for an economic contraction can do more harm than good (Keynes, 1936). The same logic applies to an investor, who "Is forced to concern himself with the anticipation of

impending changes, in the news or in the atmosphere, of the kind by which experience shows that the mass psychology of the market is most influenced," and underscore why President Roosevelt used his Fireside Chats to address the public's concerns and reestablish confidence in the economy (Keynes, 1936). Monitoring the perception of economic health is just as important as the actual economy's health, giving policymakers time and reason to preemptively act.

The Founding Fathers viewed nationwide data collection as a necessary feature of a government that serves its people. The Constitution calls for a decennial census to determine population and appropriate congressional representation. The Census has expanded in scope beyond a simple population count since it was first fielded in 1790, now offering a more comprehensive understanding of the country. In 1810, for example, questions were added to obtain the quantity and value of manufacturers' products (Census, 2000). For more than one hundred years, the Census was the only source of nationwide data capturing the realities of Americans, but the infrequent collection limited the insights this data provided. Other economic indicator datasets did exist to fill in the gap between censuses; one of the earliest tracked the annual cotton production beginning in 1798, a testament to the contributions of enslaved peoples in the early years of this republic (National, 1951). Historian Caitlin Rosenthal argues in *Accounting for Slavery: Masters and Management* the strong parallels between "accounting practices used by slaveholders and modern business practices" (Rosenthal, 2018). Some metrics, such as Labor Productivity, used to assess the health of the U.S. economy were derived from the practices of slaveholders, who maintained records showing "how much cotton every enslaved person who was working on a plantation picked every single day. West Indian plantations actually produced these monthly reports that summarized all of the data from the plantation onto a single sheet" (Rosenthal, 2018). Meanwhile, similar data did not exist from Northern factories

because workers “were quitting all the time. A lot of Northern factories had 100 percent turnover” (Rosenthal, 2018). The Founding Fathers’ belief that a good government collects good data on its population was not a revolutionary idea. The word “census” comes from the Latin word “censere”, or to estimate, and a census was central to the administrative operations of the Ancient Roman Empire (Office, 2016). While some metrics and methods of data collection have a darker side to them, history shows that data will continue to be an important input in the role of government.

By the end of the 1930s, frustrated with the lack of tools to fully understand the scope of unemployment in the United States, statisticians, economists, and policymakers worked together to create the Monthly Report of Unemployment. The first survey was fielded in March 1940, aligning with the 1940 Census to test the accuracy of this newly developed sampling method (Dunn et al., 2018). Two years later, the survey was renamed the Monthly Report on the Labor Force and morphed “to underscore new wartime concerns about labor scarcity” (Dunn et al., 2018). Today, the survey is called the Current Population Survey (CPS), highlighting its encompassing nature, and includes additional questions on income and earnings. The invention of automobiles, telephones, and computers drastically speed up data collection. In 1994, the CPS implemented a fully automated computer-assisted interviewing system. This advancement in technology does not mean data collection is easy, underscored by the fact the CPS interviewing manual is 434 pages. The CPS relies on a sample of 60,000 occupied households interviewed in-person or via telephone, taking between ten to fifteen minutes to complete. Each year, the U.S. Census Bureau conducts over 130 different surveys and programs each with its own complex methodology, and while computers do most of the survey legwork, accurate results require an army of field representatives, thousands of hours, and billions of dollars (Dunn et al., 2018).

Despite the government's best efforts, collecting high-quality data is not getting easier; in fact, some have reported having an increasingly difficult time.

In 2019, Pew Research Center, one of the leading public opinion pollsters in the country, raised the alarm about their declining telephone survey response rate— falling to seven percent in 2017 and six percent in 2018, and down from thirty-six percent in 1997 (Kennedy & Hartig, 2019). They point to the surge in automated telemarketing calls as one reason why participants may be declining their calls. The U.S. Bureau of Labor and Statistics (2021) states CPS's nonresponse rate is less of a problem than its other surveys, and the October 2021 response rate stood at 75 percent of eligible households, significantly higher than Pew's. In June 2020, however, the CPS response rate cratered to 64%, and while it has since recovered, there is a clear downward trend since October 2011. The emerging problem is telephone surveys are still considered the gold standard for polling methods. When an increasing number of eligible households do not respond to survey inquiries, it becomes more challenging and expensive to run a high-quality survey, especially given the fact that "even small levels of nonresponse could have an effect on estimates, such as the unemployment rate, which is measured in tenths of a percentage point" (Bureau, 2021). Another downside of relying on surveys to calculate economic indicators is the lag between when surveys are fielded and when statistics are released. CPS is conducted during the calendar week that includes the 19th of the month and questions ask about the household's employment status from the week prior (the week that contains the 12th day of the month) (Census, 2015). Table 1 breaks down the difference in days between the reference week and when the unemployment rate is published, ranging between 20- and 26-days difference. The unemployment rate is not the only economic indicator with a lag; Table 2 lays out the reference period and the release dates of several measures calculated by the Bureau of

Labor Statistics and the Bureau of Economic Analysis. Each of these statistics offers insights into the economic reality of consumers, producers, and the overall health of the entire economy, yet, they all have a lag ranging from a few days to over a month. The economy moves quickly, as evidenced by the rapid succession of events after Black Thursday in 1929 that led the U.S. into the Great Depression or the once-in-a-generation pandemic shutting down businesses overnight. The lag between survey reference week and its release date is a loss of critical time for policymakers and economists to act, hindering their understanding of the on-the-ground reality of the economy. This paper will not explicitly nowcast economic indicators and instead seeks to serve as a methodological intervention, opening the door to the effectiveness of utilizing alternative data sources to predict economic indicators.

Table 1: Unemployment Rate Release Schedule.

Reference Week	Release Date	Difference (days)
10/12/21	11/5/21	24
11/12/21	12/3/21	21
12/12/21	1/7/22	26
1/12/22	2/4/22	23
2/12/22	3/4/22	20
3/12/22	4/1/22	20
4/12/22	5/6/22	24
5/12/22	6/3/22	22
6/12/22	7/8/22	26
7/12/22	8/5/22	24
8/12/22	9/2/22	21
9/12/22	10/7/22	25

Table 2: Release Schedule of Economic Indicators

Indicator	**Release**	**Reference Period**
Job Openings and Labor Turnover Survey	1/4/2022	November 2021
Employment Situation	1/7/2022	December 2021
Consumer Price Index	1/12/2022	December 2021
Real Earnings	1/12/2022	December 2021
Producer Price Index	1/13/2022	December 2021
U.S. Import and Export Price Indexes	1/14/2022	December 2021

Usual Weekly Earnings of Wage and Salary Workers	1/19/2022	Q4 2021
Gross Domestic Product (Advance Estimate)	1/27/2022	Q4 2021
Employment Cost Index	1/28/2022	Q4 2021
U of M Consumer Sentiment (Final)	1/28/2022	January 2022
Productivity and Costs (Preliminary)	2/3/2022	Q4 2021

As computers were introduced to everyday life, the actions of users and the content of websites became a source of data. Yet, while a plethora of data exists, the government continues to rely on data collection methods designed in the 1930s and lacks any real-time health assessment of the U.S. economy. On the other hand, researchers and companies are forging a new path using big data to create innovative metrics of economic health. The Billion Prices Project was an academic initiative at MIT and Harvard University that used prices from hundreds of online retailers to construct a price index to track inflation (Billion, n.d.). This measure parallels the monthly Consumer Price Index (CPI), calculated monthly using the prices of approximately 80,000 goods and services across 75 urban areas. The prices of goods and services are obtained primarily through personal visits or telephone calls by BLS data collectors and sometimes websites. Like the monthly Personal Consumer Expenditure figure, Mastercard's SpendingPulse service offers real-time consumer spending patterns across sectors and geographies by aggregating credit card transactions, creating a dataset of the entire population of Mastercard users instead of just a sample (Mastercard, n.d.). There are countless other innovative applications of big data to replace the outdated methods used by the U.S. government; in fact, two economists wrote in 2013, "over the next decades big data will change the landscape of economic policy and economic research" (Einav & Levin, 2013). Almost one decade later, there is still much progress to be made, but policymakers, researchers, and companies continue to see more sources and applications of big data.

While the term *big data* might be new, that does not mean decades-old *big* sources of data are non-existent. The National Bureau of Economic Research (NBER) maintains an extensive archive of macroeconomic data, including a monthly index of help-wanted advertising in several U.S. newspapers from 1919 to 1956. An analysis of this dataset found that "cyclical movements in help-wanted advertising conform very closely to cycles in the economy at large" (Boschan, 1966). This dataset compiled by Metropolitan Life Insurance Company may be one of the earliest examples of big data and sets the stage for this thesis, which explores whether the content within a newspaper can offer a more robust assessment of the U.S. economy. The New York Times, founded in 1851 as the New York Daily Times, operated alongside several local, daily newspapers circulating in New York City. Since then, the paper has grown in scope and prestige, becoming the go-to source for news, business, style, and opinions. The paper embodies its slogan, "All the News That's Fit to Print", by reporting on major world events such as World Wars I and II, national events like the passage of the Affordable Care Act, New York City-specific news— the election of Rudy Giuliani to Mayor of New York City— and events, like 9/11, that sits at the intersection of world, national, and local news. When the paper began publishing online in 1996, it was no longer confined to *Print.* The press is often referred to as authors of the first draft of history, reflecting their ability to capture present happenings that will, one day, line the pages of history textbooks. Given the scope of the New York Times, they not only report on the financial capital of the world but also offer a stethoscope on the health of the entire nation, capturing sentiments, statistics, and stories from California to Maine. This thesis will explore if changes in two national economic indicators— consumer sentiment and unemployment rate— can be predicted using the frequency of key words in the New York Times. This relationship relies on the assumption that both national economic indicators and the

New York Times captured the economic reality of Americans, one quantitatively, one qualitatively. By converting the qualitative nature of words in the New York Times to numerical frequencies, I can assess if the New York Times' stethoscopic reporting of the country accurately predicts changes in economic indicators and thus eliminate the lag between the reference period and release date. For example, if unemployment rises and more Americans find themselves out of a job, does the word "unemployed" appear more; is the pattern strong enough to build an accurate machine learning model? I hypothesize quantifying the text in New York Times articles can accurately predict consumer sentiment and unemployment rate.

This thesis will add to the literature in four ways. First, I offer a robust assessment of forecasting different economic indicators using textual data from the New York Times. I will show whether text data can monitor the health of the U.S. economy, independent of any other data. Second, this thesis will assess the predicting power of three different manipulations of text data seen throughout the literature, providing insights as to which metric is best in penalized regression models. Third, I lay out best practices and methods for researchers and policymakers to follow when turning qualitative data into quantitative data, beginning with a description of how to morph words in articles into a time series dataset for predictive analysis. Lastly, this thesis adds to a growing literature reliant on high dimensional datasets, making it transferable to other research using thousands of regressors in machine learning models. This paper is guided by the aspirations of economists and policymakers who seek the means and methods to accurately nowcast economic indicators. I show how the frequencies of specific words in newspaper text reflect current economic health, as well as different approaches to harnessing the predicting power text data for forecasting economic indicators.

When President Hoover incorrectly stated the Great Depression was over, there was no measure of national economic activity and determining the unemployment rate lacked the statistical knowledge and methodical rigor we currently use (Lohr, 2013). Today, any such statement would immediately be debunked and political, but that does not mean our data collection is flawless. Technology has advanced rapidly since 1940, yet the Bureau of Labor and Statistics and Bureau of Economic Analysis still rely on methods crafted in the 20th century— in part because technological advancements do not correspond to better survey collection. The remainder of this thesis will show how to employ use text data in machine learning models to closely match the performance of expert forecasts. Section two lays out a foundation of literature this thesis relies on for a robust analysis. Sections three and four provide the steps to collect data and the methods used to create predictive models. Section five describes the results, in which I find text data alone is ineffective at predicting these two economic indicators, but by including lagged economic indicators and reducing the number of coefficients, model performance can improve. Section six discusses the implications of these findings, and lastly, section seven offers some concluding thoughts.

2 Literature Review

Given the spanning disciplinary scope of this thesis, it is necessary to build upon a foundation of previously published research across the spectrum of social science and computation. This literature review focuses on nowcasting studies, the use of newspaper text as a source of data, and examples of nowcast models using newspaper text as the main predictors of economic indicators. A focus of this thesis is how to shorten the lag between the end of an economic indicator's reporting period and the release of that period's results. For some indicators, it takes days to release results, for others: weeks. Many researchers refer to this ability

as “nowcasting”; the term originated in meteorology to describe the ability to use data from the past to predict weather patterns of the present or very near future (Bok, 2018). It is one thing to warn people of a tornado after it hits the ground; it is another to predict when and where a tornado will form, allowing cities to prepare for what is to come. In economics, outdated survey methods govern what economists know about the current health of the economy. For example, in the United States, the monthly unemployment rate is calculated by interviewing 60,000 households about labor force activates over the phone or in-person; thus, a lag exists between when surveys are sent to the field and when results are released (U.S. BLS, 2015). Consequently, the insights we gain into the health of the United States economy are weeks old. It is one thing to warn policymakers of a recession when millions of Americans have already lost their jobs; it is another to take proactive steps to avert an economic crisis.

Economists, aided by big data and superior computation ability, saw the added value of nowcasting and began using the term to describe the ability to monitor macroeconomic indicators in real-time. Staff at the Federal Reserve Bank of New York note that nowcasting allows for smarter and more accurate insights into the state of economies (Bok, 2018). In the field of economics, “nowcasting” first appeared in a paper published by researchers from the European Central Bank, which defined the term as “the prediction of the present, the very near future and the very recent past” specifically related to economic indicators, such as GDP (Giannone, Reichlin, & Small, 2008). Their approach takes advantage of the skewed nature of the monthly data releases of around 200 macroeconomic metrics to paint a clearer picture of quarterly GDP as the months within a quarter of interest pass. Of course, researchers have been interested in the tracking of economic health for decades, and Giannone et al.’s paper marked the beginning of an era where regulators and policymakers saw the value and feasibility of a real-

time assessment of economies. While there was great fascination around the model presented by Giannone et al., it still relied on economic metrics and surveys— such as new residential sales, the Federal Reserve's Business Outlook Survey, and monthly sales for retail and food services— to predict GDP. Soon researchers were interested in operationalizing data that is not inherently related to the economy or released by a government entity to predict economic indicators.

Enter Google, where the world turns to for answers to nearly all their questions. What is the time in Paris? Do I have the flu? How do I get my unemployment check? The introduction of the Google Trends feature, which provides an index of the volume of Google queries by geographic location and category, gave researchers access to the type of big data that would allow them to nowcast just about anything. Case in point, when more and more individuals begin searching questions related to unemployment, there might be an indication that the unemployment rate will rise. In fact, that is what several researchers found. For example, Choi & Varian (2009), two researchers from Google, find a positive correlation between initial claims reported by the United States Department of Labor and Google Trends search query data related to Jobs and Welfare & Unemployment. The literature became inundated with other articles using Google search trend data to nowcast private consumption (Kholodilin, 2010), unemployment (Askitas & Zimmermann, 2009; Choi & Varian, 2009), stock market behavior (Preis, 2013), auto sales (Choi & Varian, 2009), and new house sales (Choi & Varian, 2009)— all of which convey some assessment of economic health. Choi & Varian (2009) also found that adding relevant Google Trends variables to both a seasonal autoregressive model and a fixed-effect model "[outperformed] models that exclude these predictors. In some cases, the gain is only a few percent, but in others can be quite substantial, as with the 18% improvement in the predictions for Motor Vehicles and Parts and the 12% improvement for New Housing Starts". Researcher's

use of Google Trends was just big data's opening act, and soon enough, researchers found other tools to strengthen their forecasting ability.

The proliferation of social media gave researchers access to one of the largest, most robust data sources; one that captures the daily attitudes, opinions, and experiences of users. When individual data is aggregated, trends among the population can be identified, allowing for accurate nowcasting on just about anything. One study uses Twitter to forecast a movie's box-office revenue, finding a significant correlation between opening weekend revenue and the number of tweets referencing a movie before its release. The authors, Asur and Huberman (2010), boldly title their paper *Predicting the Future with Social Media*, highlighting the sheer power of nowcasting with big data. Another group of researchers used a combination of social media and internet search data to predict covid-19 outbreaks across China. Using keyword searches as an indicator that an individual is beginning to display symptoms of covid-19, Li et al.'s model could predict where and when covid-19 outbreaks are bound to occur. They found a strong correlation between the normalized count of two keywords— coronavirus and pneumonia— and data published by the National Health Commission of China (NHC) on the daily count of laboratory-confirmed and suspected cases of covid-19 (Li et al., 2020). Going a step further than other researchers who utilize social media data to forecast, Li et al. (2020) present a lag correlation to show where the accuracy of their model peaks by revealing the offset between official data and what is occurring on the ground. They found "a maximum correlation at 8–12 days for laboratory-confirmed cases and 6–8 days for suspected cases" meaning individuals were using the keywords of interest on a search engine or social media several days before those cases were confirmed by the NHC. These studies highlight how operationalizing the right big data sources and using proper computational methods allow for more accurate insights

into what is happening in the world around us in a timelier manner than official government data. While social media provides a plethora of data to analyze, other researchers what some might call the *original social media*: the newspaper, a daily summation of the on-the-ground-reality of the region said newspaper covers.

Newspaper text serves as a primary data source for social scientists to analyze a wide range of activities for decades, given their continuous nature, providing rich data on events across the world. The literature consists of countless examples, including civil strife and protests (Gurr, 1968; Jenkins & Eckert, 1986), revolutions and interstate war (Tanter, 1966), and stock market movement (Ammann, 2011; Wan, 2021). Researchers have elected to use newspaper text as a predictor because "by quantifying language, researchers can examine and judge the directional impact of a limitless variety of events, whereas most studies focus on one particular event type" (Tetlock, 2008). In other words, newspapers detail a collection of events that, when aggregated, present a clear understanding of how the intersection of events influences the lives of individuals. The New York Times mission statement of values states "our breadth of coverage reaches well beyond news and politics. Since our inception, The Times has been a thoughtful and deep resource for topics that touch our readers' daily lives" (The New York Times). By relying on a population of newspaper text and not just a sample of specific events, researchers are taking the opposite approach of that laid out by Eugene Fama who coined the term "dredging for anomalies" to describe running studies that look at single events until one obtains a significant result when predicting the prices of assets (Fama, 1996). The literature has seen "an explosion of empirical economics research using text as data" given "the information encoded in text is a rich complement to the more structured kinds of data traditionally used in research," according to Gentzkow, Kelly, and Taddy (2019). Their article *Text as Data* offers a thorough assessment of

studies that employ text as data and the necessary steps researchers should take to properly process and utilize when building a dataset of textual variables. An entire section of *Text as Data* is dedicated to nowcasting with text data, where the authors state "text produced online such as search queries, social media posts, listings on job websites, and so on can be used to construct alternative real-time estimates of the current values of these variables" (Gentzkow, 2019). The literature contains countless studies relying on newspaper text to nowcast an economic variable, several of which are included in Gentzkow et al.'s article, but there appears to be three main approaches to modeling the relationship between the two.

The first approach is running a sentiment analysis on articles, which is rooted in a hypothesis that consumers and businesses rely on the news when making decisions. If a consumer sees a bad news story about the economy, maybe instead of going out to dinner, they will stay in and save that money. Thus, negative sentiments in newspapers can lead to less spending and investment. In June of 2020, The Reserve Bank of Australia laid out a method of assessing the health of their economy using newspaper text called News Sentiment Index. The authors state the need for a new approach to monitoring the economy in real-time arose during the onset of the covid-19 pandemic, which revealed gaping holes in our current methods of assessing economic health. Using three Australian newspapers from September 1987 to April 2020, Nguyen and Cava construct their daily index by taking the number of positive words subtracted by the number of negative words divided by total words. The 30-day moving average of the index "captures key macroeconomic events, such as economic downturns, and typically moves ahead of survey-based measures of sentiment" (Nguyen & Cava, 2020). Other studies that employ news sentiment to nowcast economic indicators include Tetlock (2008), who calculates a "pessimism score" of a Wall Street Journal column to forecast stock market returns.

The second approach seen in the literature to nowcasting economic indicators using newspaper text involves running analyses on word counts or the frequencies of specific words. Ammann, Frey, and Verhofen (2011) take words frequently used in newspaper articles and count the number of articles in a leading German financial newspaper that contain at least one appearance of those words. After standardizing the word counts, the authors use a stepwise linear predictive regression on their highly multivariate dataset— 236 variables (word counts) to be exact— to accurately predict future German Stock Exchange returns. The Economist presents another example of this approach in a short article published in 1998 titled *The Recession Index*, which laid out a theory and methods for what would later become known as the R-word index database. The idea was to count the number of stories in the New York Times and Washington Post included the word "recession" each quarter (The Economist, 1998). The index accurately pinpoints the start of a recession in 1981, 1990, and 2001; after subsequently falling post-2001, the r-word index began to rise in the second half of 2007 and soared in early 2008 (Economist, 2008). It took until December 1, 2008, for the National Bureau of Economic Research stated that the United States entered a recession in December 2007 (Isidore, 2008), underscoring the successful forecasting ability of a simple word count index. The creators of the R-word index state the New York Times and Washington Post are "instantly available, unlike official statistics, which are always out of date" (the Economist, 2001), highlighting one advantage of using newspaper text in their approach to monitoring the health of economies.

The last approach involves processing newspaper text with NLP methods and statistics to predict economic indicators. Using Le Monde, the largest newspaper in France, researchers begin by taking a similar approach Nguyen and Cava to improve GDP nowcast predictions by using a sentiment index of newspaper text but go a step further by calculating the term

frequency-inverse document frequency (TF-IDF) of words to run an elastic-net regression using 6,000 potential independent variables to predict France's Real GDP growth rate (Bortoli, Combes, & Renault, 2018). TF-IDF is a statistic that measures how important a word is to a document in a collection of documents. Bortoli et al. then average the TF-IDF values for each word by month or quarter, so the variables are at the same frequency as the dependent variables. The results of the elastic net regression using TF-IDF, however, performed comparably to an autoregressive model. This approach is also taken by Kalamara et al. (2020) who finds TF-IDF metrics from three popular UK newspapers can provide robust forecasts of economic indicators when included in in several prediction methods. Despite the mixed performance of these TF-IDF regressions, both these papers lay out a clear approach for incorporating NLP methods into nowcasting economic indicators using national newspaper text.

Each of the approaches for nowcasting using newspaper text laid out in this literature review offer a solid foundation for this thesis to build on. Nowcasting requires robust, frequently released data that is representative of the entire population, hence why economists who nowcast GDP do not rely on one variable but hundreds to capture the feelings of consumers, producers, and other actors. Giannone et al. (2008), for example, use 200 macroeconomic metrics in their model to forecast quarterly GDP to assess the health of manufacturers and consumers, understand the current state of trade, corporate dept, inventories, and home sales. Newspapers, like social media and economic surveys, offer a collection of experiences that when aggregated can offer a clear picture of what consumers and businesses are facing. Further, word counts provide a high-dimensional dataset that is a staple of other nowcast models like Giannone et al.

The final topic to cover in this literature review is accuracy and how previous studies have assessed the accuracy of their nowcasting models. The simple way is to train a model and test the

accuracy using a testing sample and assessing accuracy based on the difference between the predicted outcome and the actual outcome. This is done using root-mean squared error (Giannone, 2008), BIC (Askitas & Zimmerman, 2009), or r-squared (Li et al. 2020). The second approach, taken by Bortoli et al (2018), Kalamara et al. (2020), Choi & Varian (2020), Kholodilin (2010) assess the difference in root mean square forecast error (RMSFE) between a baseline simple autoregressive model (AR) and a different model that includes text variables. The last method of assessing model accuracy is not seen explicitly in literature on nowcasting but is highlighted elsewhere in economics literature is comparing a model's performance to the consensus of professional forecasters. In fact, there is an entire industry of bankers, consultants, journalists, and other economists creating their own models and methods to forecast metrics like GDP, unemployment rate, or Consumer Price Index. By comparing a model's predictions and subsequent error to that of the average error of professional forecasters' models, it is possible to determine whether a model using text variables can accurately nowcast economic indicators.

3 Data

The source of newspaper text data comes from the New York Times. I retrieved every single published article from January 1981 to December 2020 through the New York Times' archive application programming interface (API), which returns the metadata of all articles grouped by month. In total, there are over 1 million pieces written during this period. The returned metadata includes several essential features, including the section and medium of an article. I exclude any piece in the Business, Style, or Real Estate sections, among several others, and any multimedia or audio works. While I want to include the business section in my analysis, given its focus on all aspects of the economy, this section contains upwards of 900,000 additional articles— a majority of which were the quarterly earnings reports of publicly-traded

companies. As a result, I include only articles published in the U.S. or Today's Paper sections. The API also returns the URL of articles, which allows me to computationally scrape text from the New York Times website using the Requests and Beautiful Soup Python packages. Once the text of each article is scraped, I convert all words to lowercase and clean the text of punctuation, numbers, and single-lettered words. Two subsequent transformations of data are run to finalize this thesis's dataset. The first collection ignores a list of stop words from the NLTK package and 72 other words commonly seen in the testing phases to avoid counting meaningless ones, such as *the*. Scraping all articles in the U.S. and Today's Paper sections published in the period of interest results in a count of 402,489 different words used in the New York Times from 1981 through 2020. To narrow this list, I select words that appear at least 1,000 times, which reduces the number of total words to 10,902.

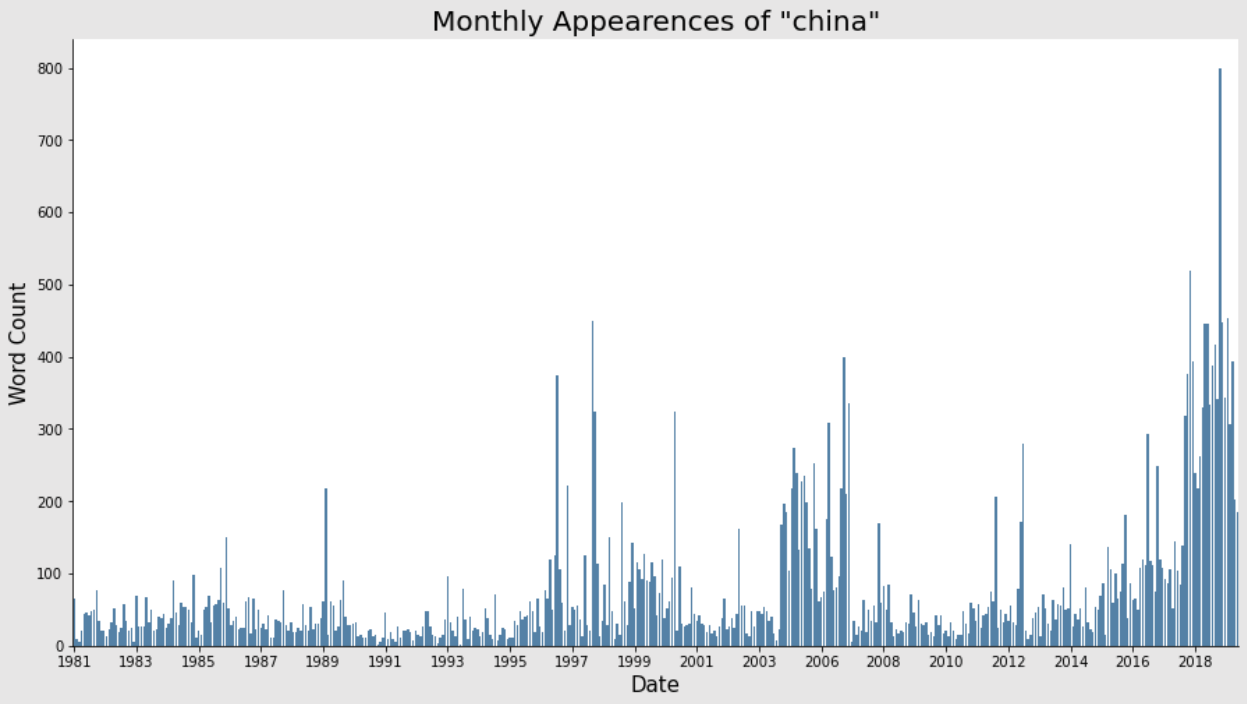
To further narrow the list of words included in the final dataset, I manually go through all the remaining words and remove ones that would not be beneficial for analysis. In the initial scrape, the most frequently appearing word that I subsequently removed from the dataset is *bush*, likely referencing Presidents George H. W. Bush and George W. Bush. Nearly all the removed words are first and last names. The reason for eliminating names is because this analysis focused on the underlying pattern of words used in news coverage to predict the status of the economy and not the predicting power of references to individuals making decisions. For example, *bernanke*, referring to the former Chair of the Federal Reserve, is used in the New York Times 293 times, a majority during the 2008 Great Recession when he navigated the United States out of the worst financial crisis since the 1930s. Yet, additional references to Ben Bernanke in 2016, when Janet Yellen was the Chair of the Federal Reserve, would bias the model given the negative correlation between the appearances of *bernanke* and economic growth. Other words

removed include the names of cities, counties, and states in the U.S., often features of a dateline— a brief piece of text at the start of an article that describes where a story is written. If the New York Times publishes a story about families struggling to pay their bills, this analysis does not seek to differentiate where those families live; what matters is that families are struggling, which is the substance of the article. The frequency of some states in the New York Times is seasonal, such as Iowa, one of the most referenced states, which draws substantial attention every four years by holding the first Presidential Primary. In total, 391,686 words were cut, appearing in the New York Times 10,895,416 times—averaging out to about 26 appearances per word, or 0.7 appearances per year. The top 10 most frequent words removed are shown in Table 3. Countries and international cities remain given the globalized economy and how the coverage of certain countries provides additional insights into the underscoring narrative of an article. Figure 1 shows the number of monthly appearances of *china*, which tells a fascinating story of China's rise from a secluded, agrarian country to an economic powerhouse. For example, the spike around 2006 coincides with discussions eventual passage of the 11th Five-Year Program, which laid out "a program of government action designed to ensure that rapid growth will be sustainable over the long term, and that the fruits of growth will be more equitably shared" (Naughton, 2005). Its growth, however, is not without a few speed bumps; the spike in the appearances of *china* in the late 1990s is likely due to the regional ripple effects of the Asian financial crisis.

Table 3: Most Frequent Words Removed

Word	Count
Bush	183,056
May	162,350
Washington	158,648
Clinton	149,664
Us	106,553
John	103,176
California	94,141
Obama	92,252
Reagan	90,913
Texas	74,638
South	73,807

Figure 1: Word Count of China from Jan. 1981 to Dec. 2019



Lastly, I stem all the remaining words to their root to further reduce the feature space. I do this after picking the most common words because the process to ensure accurate stemming is a manual and time-consuming endeavor. Stemming is a process of removing the suffixes and prefixes of a word; thus, the final dataset is the count of all root words and their unstemmed counterparts. For example, the root of *discovered*, *discovers*, and *discovering* is *discover*, therefore stemming reduces the number of features while also increasing the robustness of my

dataset. In the example above, the number of features is one, *discover*, but the appearance count is four. For this analysis, I will use three different stemmers— Lemmatization, Porter, and Krovetz— to create a hierarchy of checks to return accurate roots. First, if at least two stemmers return the original word, indicating a word is already a root: 2,924 words fall into this category. Next, I look for cases where all three stemmers return the same root, finding 916. Then, I check if two stemmers return the same word, using a series of elif statements and prioritizing the more accurate Lemmatization and Krovetz stemmers. I stem the remaining 902 words using the simple and least aggressive Krovetz method. Then, I create a python dictionary where the key equals one of the 9,352 words of interest, and the value equals that word's root. Stemming reduced the total number of words to 6,307, which parallels the 6,000 coefficients used in Bortoli et al.'s (2018) nowcast of French GDP using Le Monde. This process efficiently returns the most accurate roots, yet it is not perfect. For example, this process does not correctly stem *profitable* to its root of *profit*. These limitations are discussed in the conclusion section, but this process correctly stems a vast majority of words and reduces the feature space by 32 percent.

I conduct a final scrape of the articles in the U.S. and Today's Paper sections of the New York Times from January 1981 to December 2020, this time not only counting the number of appearances but also the proportion of a month's articles containing each of the 6,307 words and their unstemmed counterparts. I will henceforth refer to this metric as article count. Lastly, I calculate the modified term-frequency inverse document frequency (TF-IDF), which measures a word's importance, and is used in previous studies with some success (Bortoli et al., 2018 & Kalamara et al., 2020). In a given month, a word used frequently in many articles will have a low TF-IDF value; similarly, a few articles making few mentions of a word will also result in a low TF-IDF value. For example, a word appearing frequently in several articles, such as *the*, is just as

unimportant as a word that appears rarely in articles, such as *hippopotamus*. This occurs because in both scenarios *the* and *hippopotamus* do not add any substance to the month's collection of articles. There are two parts of the TF-IDF calculation: augmented frequency and inverse document frequency, displayed in equations 1 and 2 respectively. In equation 1, $f_{t,d,m}$ is the number of times term *t* appears in all articles in month *m* over the maximum number of times any term appears in month *m*. This equation scales term frequencies to prevent bias towards longer documents and words that appear frequently, given the wide range of word frequencies. In Equation 2, *D* is the total number of articles in month *m*, and $\{d \in D_m : t \in d\}$ represents the total number of articles term t occurs in month m. Equation 3 shows the final calculation, which is the product of equations 1 and 2. The three datasets — word count, article count, and TF-IDF— contain the monthly metrics for all 6,307 words across 40 years, for a total of 480 observations. If a word is not used in a month, its word count and article count metric will equal zero. Yet, TF-IDF presents an interesting dilemma. If a word does not appear in the New York Times during a month, does that mean it is unimportant and thus its TF-IDF value should be 0? Or does that imply a greater significance and thus TF-IDF should equal the maximum TF-IDF value? I determine training RMSE is minimized when null TF-IDF values are changed to 3.63— just over the maximum value of 3.6257. In total, 217,569 null values are altered, equating to just over seven percent of the entire dataset. I create twenty-four different models using the three different text metrics to predict consumer sentiment and unemployment rate over four time periods.

Equation 1: Augmented Frequency Calculation

$$tf_{t,d,m} = 0.5 + 0.5\left(\frac{f_{t,d,m}}{\max(\{f_{t',d,m} : t'_m \in d\})}\right)$$

Equation 2: Inverse Document Frequency Calculation

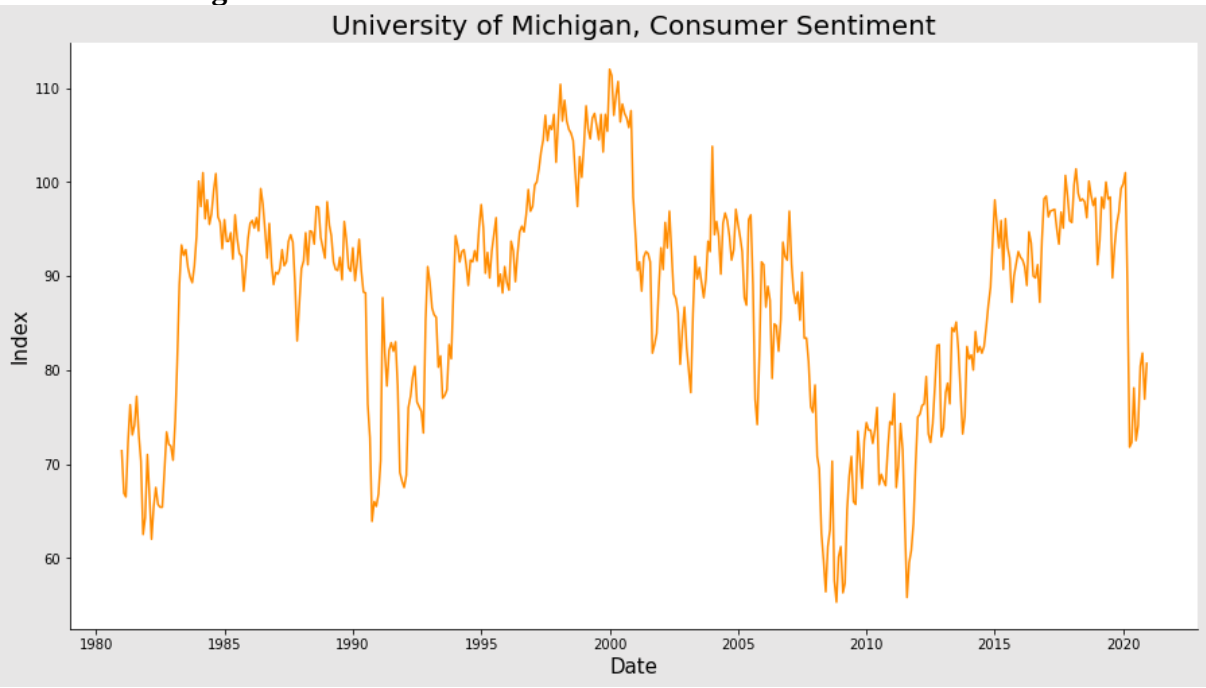
$$idf_{t,D,m} = log\left(\frac{D_m}{\{d \in D_m : t \in d\}}\right)$$

Equation 3: TF-IDF Calculation

$$tfidf_{t,d,D,m} = tf_{t,d,m} \times idf_{t,D,m}$$

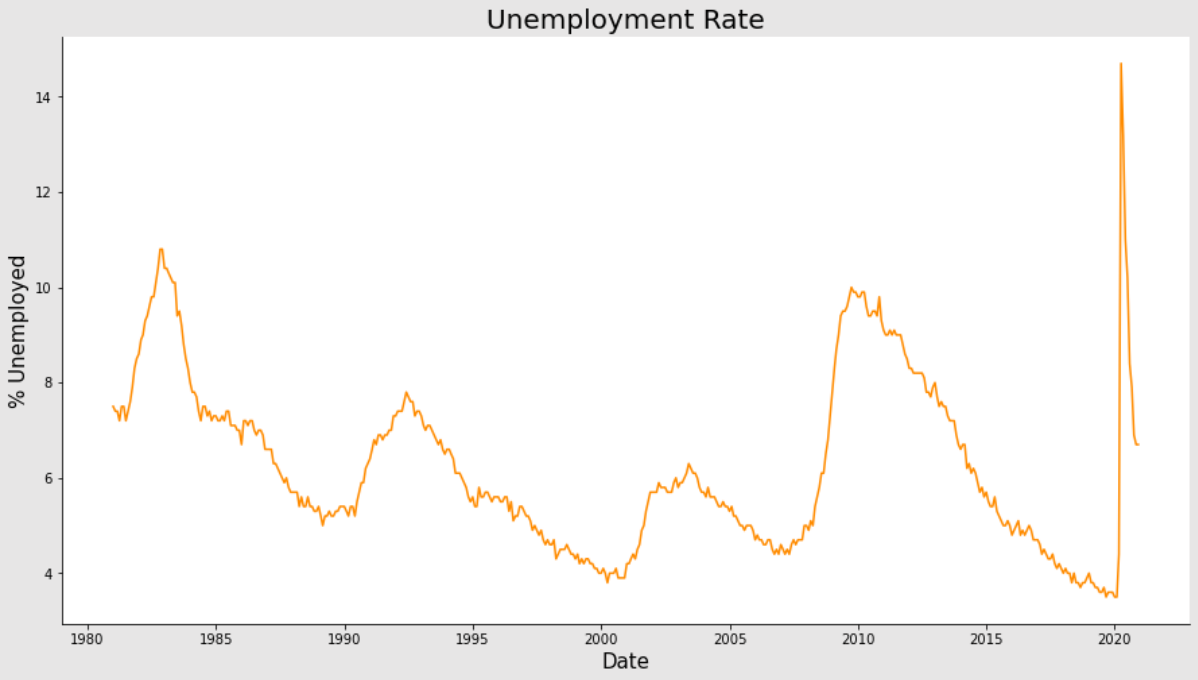
Consumer sentiment is an index of the results of the University of Michigan's monthly Survey of Consumers, which helps to understand and forecast changes in the national economy. The core survey questions focus on three areas: personal finances, business conditions, and buying conditions for large household durables, vehicles, and houses. The Survey Description states, "Economic optimism promotes consumer confidence and a willingness to make large expenditures and debt commitments, while economic uncertainty breeds pessimism and a desire to curtail expenditures and rebuild financial reserves" (University, n.d.), which connects back to Keynes's theories of macroeconomics. Economists and policymakers value this metric because consumer spending accounts for around 67 percent of U.S. economic activity (U.S. Bureau).

Figure 2: U.S. Consumer Sentiment from Jan. 1981 to Dec. 2020



The second dependent variable is the unemployment rate, which is the percent of the labor force without work. High unemployment coincides with economic downturns when demand for goods decreases, and fewer workers are needed to produce goods. As Keynes noted, this cycle can spin out of control: when more people are out of work, the demand for goods decreases, resulting in more people out of work. The unemployment rate is one of the core economic indicators used by economists, and an increase can send stock markets tumbling due to its all-encompassing nature. On the consumer side of production, or the employees, if an individual is out of a job, they will decrease their spending and contribute less to the greater economy. On the supply side, when firms who see a drop in demand will lay off some of their workforces to reduce cost and the quantity of goods produced.

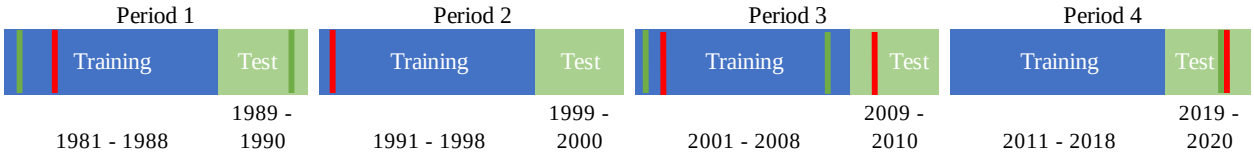
Figure 3: U.S. Unemployment Rate from Jan. 1981 to Dec. 2020



I choose to include these two measures in my analysis because both provide different lenses to assess the health of the U.S. economy. Collecting data to accurately calculate consumer sentiment and the unemployment rate takes a great deal of time and money; to provide efficient insight into the U.S. economy, I will explore whether textual data from the New York Times can

forecast these economic indicators. Data is split into four periods of equal length with models trained on datasets equaling 96 months and their performance assessed on sets equal to 24 months, as shown in Figure 4. Economies tend to follow the business cycle where economic expansion begets a recession. A peak in the business cycle refers to when economic growth maxes out, resulting in a decrease of Gross Domestic Product (GDP). Once GDP rises again, the business cycle bottoms out in a trough; however, the economy is still relativity weak even after it reaches its trough and recovery can take months or years. According to the NBER's Business Cycle Dating tracker, a peak occurs in July 1981 and troughs in November 1982, both in training period 1. The next peak occurs in July 1990, testing period 1, and bottoms in March 1991, training period 2. The U.S. economy expanded through March 2001, troughs shortly after in November 2001, and peaks in December 2007, all occurring in training period 3. Then the worst financial crisis since the Great Depression occurs, worsening until June 2009 in testing period 3. Following the Great Recession, the U.S. experienced its longest period of economic growth, peaking in February 2020 when the Covid-19 pandemic shut down the economy almost overnight. Yet, the recession did not last long and troughs in April 2020. The green lines in Figure 4 correspond to peaks while the red lines are troughs in the business cycle.

Figure 4: Visualizing Training and Testing Sets



4 Methods

The literature details several approaches to predicting economic indicators using textual data. Ammann et al. (2011) use a stepwise regression; Bortoli et al. (2018) rely on a penalized regression; Kalamara et al. (2020) use a lasso, ridge, elastic-net, support vector regressions,

artificial neural network, and random forests. Following Bortoli and Kalamara, this paper will deploy an elastic-net regression, given penalized linear models are "the most widely applied text regression tools due to their simplicity," (Gentzkow et al., 2019). The elastic-net, shown in Equation 4, was introduced by Zou and Hastie (2005) as a compromise between L1 and L2 penalties. The model selects variables like lasso and shrinks the coefficients of correlated independent variables like ridge. Further, the elastic-net is valuable when the number of predictors is greater than the number of observations—applicable for the 6,307 predictors observed over 480 months. The goal of the elastic net is to shrink the regression coefficients by imposing a penalty on their size. The coefficients come from minimizing the penalized residual sum of squares using centered, or standardized, inputs of x_{ij} and where $\sum_{j=1}^{p}|\beta_j|$ is the L1 regularization penalty used in lasso regressions and $\sum_{j=1}^{p}\beta_j^2$ is the L2 regularization penalty used in ridge regressions. Using the sklearn package in python, the λ represents the ratio of $\lambda_1 : \lambda_2$ such that if $\lambda = 0$ a ridge regression is run and if $\lambda = 1$ a lasso regression is run. The constant α weighs the contribution of the L1 and L2 regularization; when alpha = 0, an ordinary least square regression is run. Using the three different datasets— word counts, article counts, and TF-IDF— I evaluate the accuracy of predicting consumer sentiment and the unemployment rate over 4 time periods, thus creating a total of 24 different models.

Equation 4: Elastic Net Penalized Regression

$$L_{Elastic\ Net}\left(\hat{\beta}\right) = \left\{\frac{\sum_{i=1}^{N}\left(y_i - \hat{\beta}x'_i\right)^2}{2n} + \alpha\left(\lambda\sum_{j=1}^{p}|\beta_j| + 1 - \lambda\sum_{j=1}^{p}\beta_j^2\right)\right\}$$

The first step in model construction is to ensure variables are stationary and cointegrated, given the time-series nature of my independent variables and the data obtained from the New York Times. I use an augmented Engle-Granger two-step cointegration test from the statsmodels

python package. The null hypothesis for this test is no cointegration between a feature and outcome variable across all 40 years. By selecting words that return a p-value less than 0.1, I reject the null hypothesis that there is no cointegration and further reduce the feature space for each model. This results in a unique set of features for each of the 24 models, meaning the word count dataset used to predict consumer sentiment differs from the word counts used to predict unemployment. Elastic net regressions require standardized data, given the model penalizes coefficients for the penalty to accurately promote or punish coefficients. Using the Standard Scaler method from sklearn, each predictor is normalized with a mean of zero and a standard deviation of one, and test sets are subsequently transformed. For each model, I determine the optimal alpha and l1-ratio values by testing 100 different combinations of alpha and l1-ratio ranging from 0.1 to 1 and selecting the parameters that minimize the root mean squared error (RMSE) of the training set. Across all 24 models, RMSE is minimized when alpha and l1-ratio both are equal to 0.1.

Table 4: Average Number of Coefficients Across All 4 Time Periods

	Word Count (a)	Article Count (b)	TF-IDF (c)	Average
Consumer Sentiment (1)	954	928	924	936
Unemployment Rate (2)	219	201	193	204
Average	587	564	559	

The number of words with predicting power varies significantly depending on whether the elastic model predicts consumer sentiment or the unemployment rate; Table 4 breaks down the cross-sections. On average, models predicting consumer sentiment rely on 936 coefficients and just 204 coefficients for the unemployment rate. Models using the word count dataset rely on the most coefficients, followed by article count and TF-IDF. Since elastic net regressions are a combination of both lasso and ridge regressions and the l1-ratio for all 24 models equals 0.1, models select coefficients with a significant weight on the l2 penalty, but they are still slightly

influenced by the l1 penalty, allowing for the feature space to be further reduced by equating several coefficients to zero.

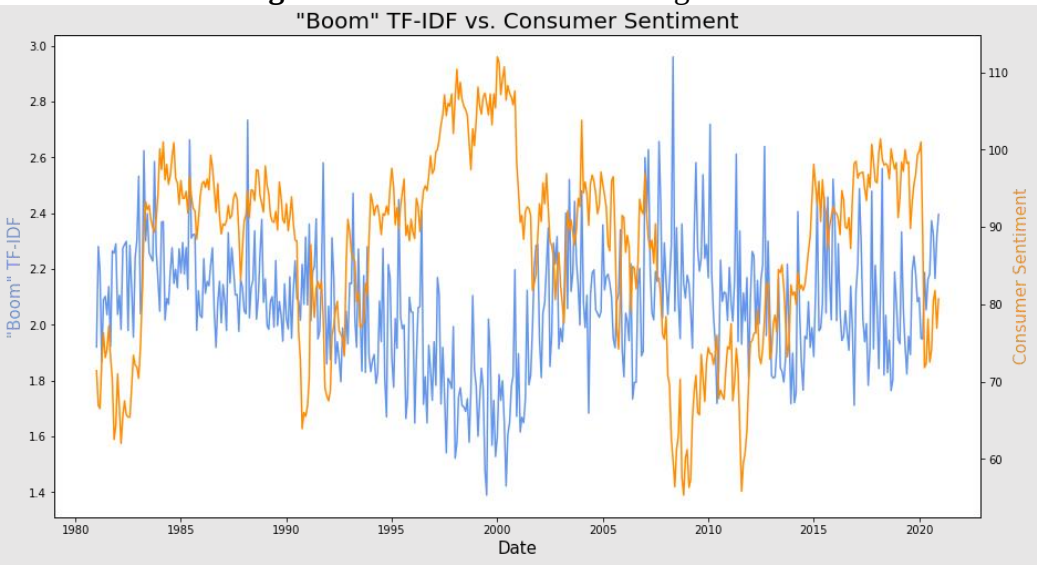
Table 5: 10 Most frequently used words in 24 models

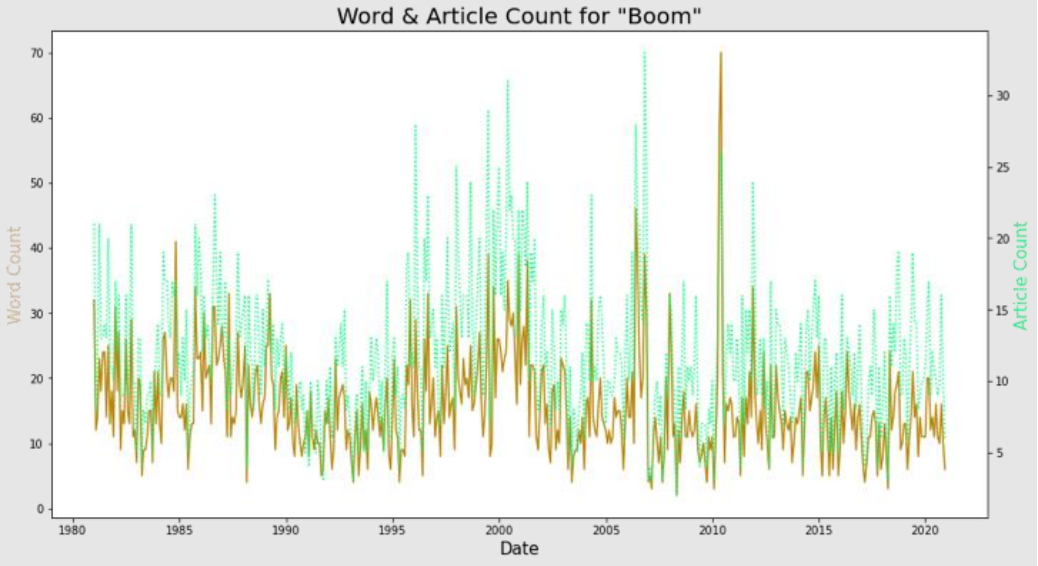
Word	# of models
unemployed	20
unemployment	18
jobless	17
economy	16
economic	15
midterm	14
recession	13
disciplinary	13
payroll	12
construction	12

Table 5 offers additional insight into how elastic net models use textual data to predict economic indicators. Specifically, which words do the models rely on most. Across all 24 models, there are 4,342 unique words used or roughly 68 percent of the entire dataset. First and foremost, this list clearly shows the challenges with properly stemming words to their root. *Unemployed*, *unemployment,* and *jobless* are all synonyms for the same thing. Yet it makes sense why these words are used more frequently; twelve of the models are predicting the *unemployment* rate. Of these twelve, eleven use *unemployed*, eight use *jobless*, and six use *unemployment*. The average coefficient of *jobless* for all models using the word count dataset to predict the unemployment rate is $\beta = 0.048$, meaning as the frequency of the word *jobless* increases in the New York Times, the unemployment rate will also increase. The models using the word count dataset to predict consumer sentiment also rely on *jobless* with an average β of -0.082, so as *jobless* increases, consumer sentiment decreases. Interpreting coefficients for word count and article count metrics are self-explanatory— a positive coefficient means an increase in the word count correlates to an increase in the corresponding economic indicator. Understanding

TF-IDF coefficients is somewhat more nuanced, given TF-IDF is a measure of a word's relevance in the collection of articles published each month. To better understand this relationship, we will look at the relationship between consumer sentiment and the TF-IDF of *boom*, which is used to describe periods of rapid economic growth. For example, a New York Times article published in June 1998 states, "A booming stock market is changing the face of American philanthropy…" (Miller, 1998). *Boom* is also the root of *Boomers*, as in Baby Boomers as seen in another New York Times article published in 1995: "retiring baby boomers dread the end of the boom times." (Sullivan, 1995). The average coefficient of *boom* for models predicting Consumer Sentiment using the TF-IDF dataset is $\beta = -0.182$. Figure 5a shows the relationship between the TF-IDF values of *boom* and Consumer Sentiment. From 1990 to 2000, there is a clear downward trend in the TF-IDF of *boom* coupled with an increase in consumer sentiment. Meaning, during this time *boom* loses its importance because, as shown in Figures 5b, there is an increase in both the word count and article count of *boom*.

Figures 5a and 5b: Understanding TF-IDF





The average absolute coefficient values in Table 7 also convey valuable information about potential model accuracy. There appears to be no significant difference in coefficient size across models using word count, article count, or TF-IDF datasets. On the other hand, there is a significant difference between the coefficients for consumer sentiment and the coefficients for the unemployment rate, suggesting consumer sentiment is more predictable than the unemployment rate. The following section contains a deeper dive into these coefficients, but Tables 4 and 6 provide preliminary insights into the accuracy of these models in that consumer sentiment is more predictable than the unemployment rate when using New York Times textual data.

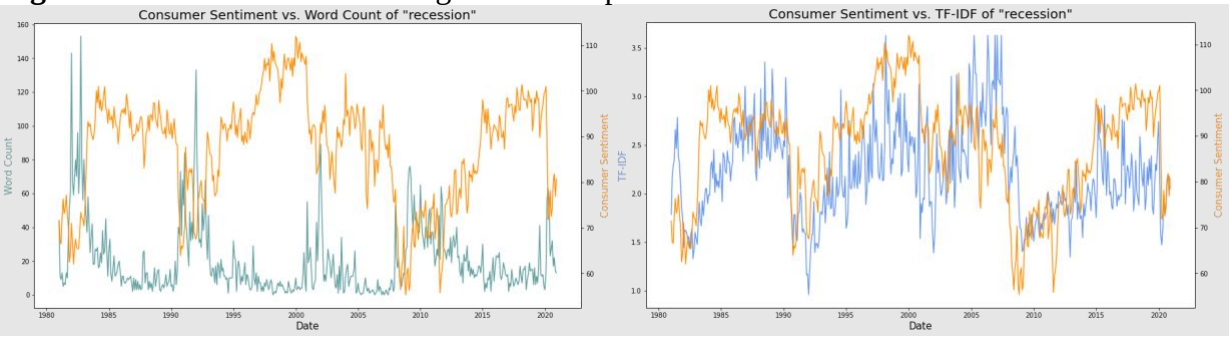
Table 6: Summary of Average Absolute Coefficient Values

	Word Count (a)	Article Count (b)	TF-IDF (c)	Average
Consumer Sentiment (1)	0.039	0.034	0.037	0.036
Unemployment Rate (2)	0.012	0.012	0.013	0.012
Average	0.034	0.030	0.033	0.032

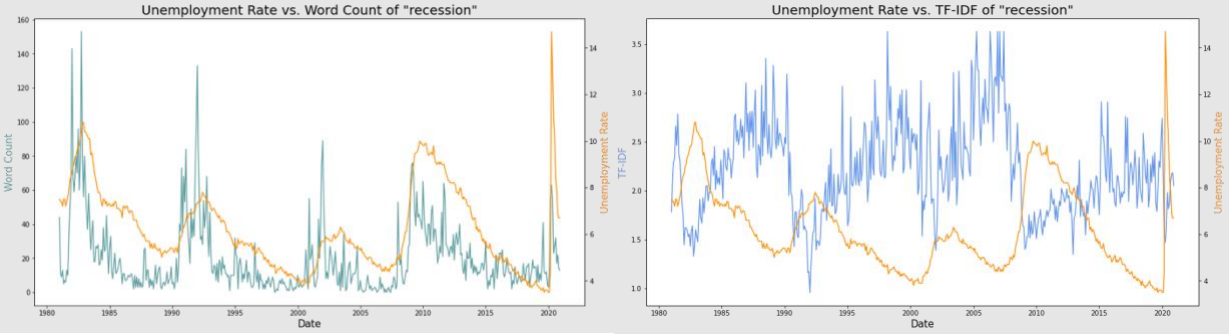
Across all 24 models, *recession* has the second-largest sum of absolute coefficient values, behind *economic*. Figures 6a, 6b, 7a, and 7b visualize the relationship between word count and

TF-IDF values of *recession* and the consumer sentiment and unemployment rate. When a recession ends and the economy expands, *recession* rarely appears in the New York Times, such as during March 1998, April 2005, May 2006, December 2006, February 2007, and June 2007. This highlights the benefit of including many coefficients in a model; during booming economic times when *recession* is not in the New York Times, models are reliant on other features to correctly predict economic indicators. As a result, strong forecasting models relying on text data will need to balance the effects of coefficients associated with bad economic times with coefficients associated with good economic times.

Figures 6a and 6b: Understanding Relationship Between Consumer Sentiment and Text Metrics



Figures 7a and 7b: Understanding Relationship Between Unemployment Rate and Text Metrics



5 Results

In this section, I present the results of all 24 models. Further, I compare the predictability of consumer sentiment and unemployment rate using text data in addition to an evaluation of the predicting power of the three different text metrics. Throughout this section, I use RMSE to assess model performance, a measure of accuracy thoroughly used in the literature. After

presenting the results for each model, Section 5.1 will provide additional information on how "good" the RMSEs are and offer comparisons between the RMSE values obtained for predicting consumer sentiment and the unemployment rate. This section also contains a thorough discussion of results by providing greater context to model accuracy and additional details underscoring why some models perform better than others. I then dive into decadal differences in words used to predict economic indicators to determine if there are shifts in the predicting power of words over time. Building off some of the revelations in the discussion, I show the progression of steps taken to improve model accuracy, providing a road map for future researchers and economists to follow when using text data to predict economic indicators.

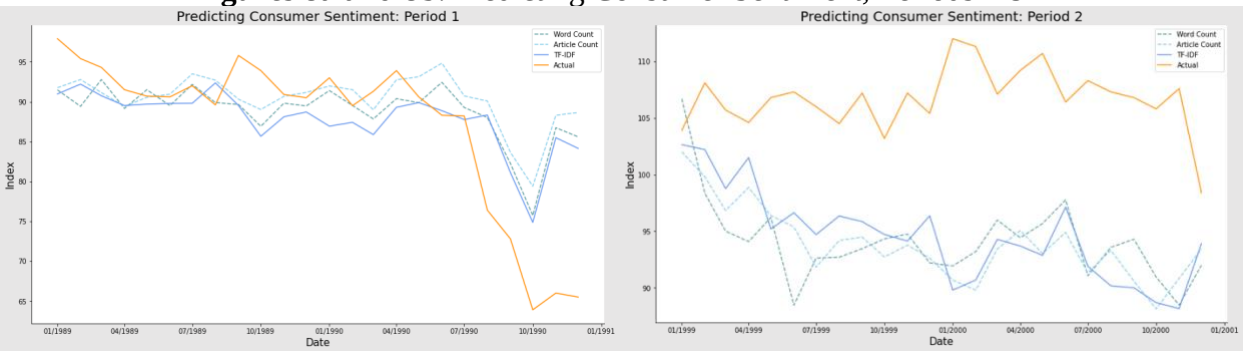
Model 1a uses a standard scaled word count dataset to predict consumer sentiment, as measured by the University of Michigan Survey of Consumers, using an elastic net penalized regression with an alpha of 0.1 and l1-ratio of 0.1. The alpha value denotes a slight weight on the penalized coefficients making models somewhat similar to a linear regression, which occurs when alpha equals 0. The l1-ratio for model 1a implies the l2 penalty used in ridge regressions is favored, reducing some coefficients to zero, yet still reliant on a large number of features. Four models are trained independently of each other, and their performance is assessed using four test datasets. The results of each testing period are as follows: the RMSE of period equals 7.69; period 2 equals 13.46; period 3 is 9.60; period 4 is 15.52. Over the four test dataset periods, the RMSE of predicting consumer sentiment using the word count dataset equals 11.97.

Model 1b uses a standard scaled proportion of total monthly articles that include a feature word to predict consumer sentiment, using an elastic net penalized regression with an alpha of 0.1 and l1-ratio of 0.1. Four models are trained and assessed independently of each other. The results of each testing period are as follows: the RMSE of period 1 equals 8.82; period 2 equals

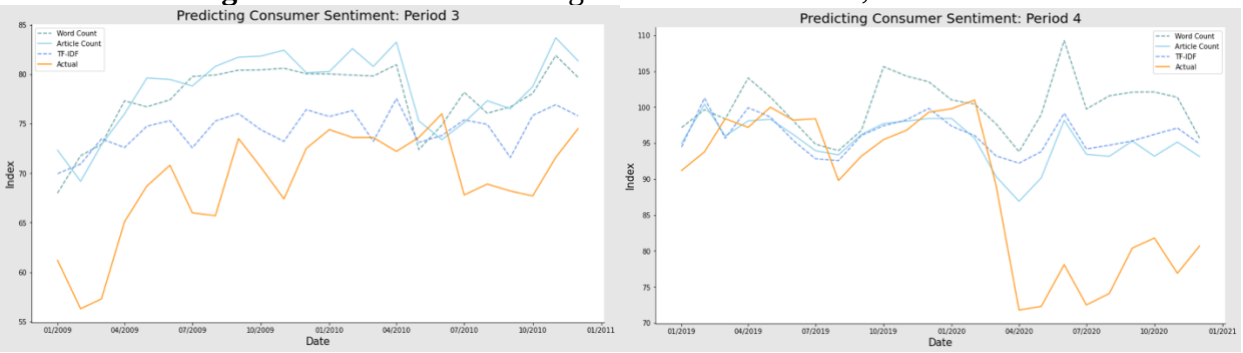
14.09; period 3 is 10.71; period 4 is 10.91. Over the four test dataset periods, the RMSE of predicting consumer sentiment using the article count dataset equals 10.94, which is roughly 8 percent lower than models using the word count dataset.

Model 1c uses a standard scaled TF-IDF dataset to predict consumer sentiment using an elastic net penalized regression with an alpha of 0.1 and l1-ratio of 0.1. The results of each testing period are as follows: the RMSE of period 1 equals 7.27; period 2 equals 13.72; period 3 is 7.10; period 4 is 12.25. In 3 out of the four time periods, the TF-IDF dataset is performs best. Over the four test dataset periods, the RMSE of predicting consumer sentiment using the TF-IDF dataset equals 10.29, which is the lowest RSME across all models predicting consumer sentiment. Further, in three out four periods, the best performing model is obtained by predicting using the TF-IDF dataset.

Figures 8a and 8b: Predicting Consumer Sentiment, Periods 1 & 2



Figures 8c and 8d: Predicting Consumer Sentiment, Periods 3 & 4



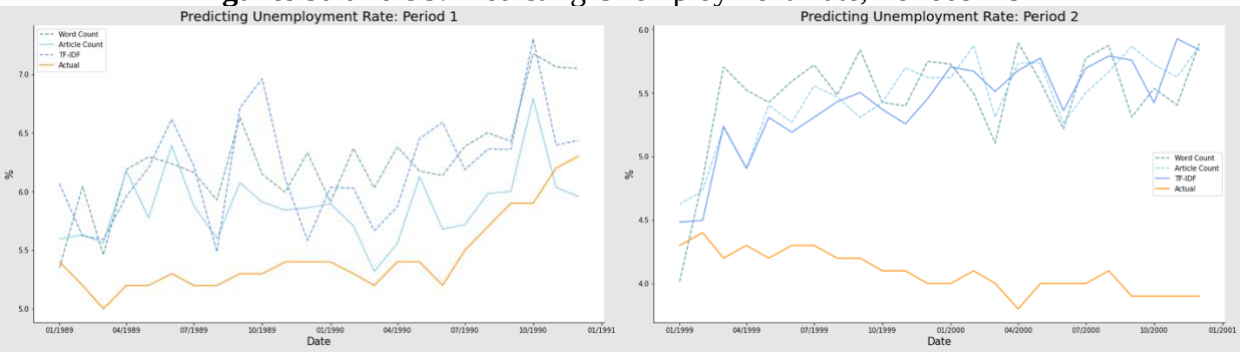
Note: for all charts, the solid line corresponds to the most accurate model for that period.

Model 2a uses a standard scaled word count dataset to predict the unemployment rate, using an elastic net penalized regression with an alpha of 0.1 and l1-ratio of 0.1. Four models are trained independently of each other, and their performance is assessed using four test datasets. The results of each testing period are as follows: the RMSE of period 1 equals 0.87; period 2 equals 1.47; period 3 is 3.88; period 4 is 4.32. Over the four test dataset periods, the RMSE of predicting consumer sentiment using the word count dataset equals 3.02.

Model 2b uses a standard scaled proportion of total monthly articles that include a feature word to predict consumer sentiment, using an elastic net penalized regression with an alpha of 0.1 and l1-ratio of 0.1. Four models are trained and assessed independently of each other. The results of each testing period are as follows: the RMSE of period 1 equals 0.55; period 2 equals 1.44; period 3 is 3.865; period 4 is 3.43. The RMSE of predicting consumer sentiment using the article count dataset from the New York Times across all four test datasets periods equals 2.695.

Model 2c uses a standard scaled TF-IDF dataset to predict the unemployment rate using an elastic net penalized regression with an alpha of 0.1 and l1-ratio of 0.1. The results of each testing period are as follows: the RMSE of period 1 equals 0.87; period 2 equals 1.42; period 3 is 3.64; period 4 is 2.65.

Figures 9a and 9b: Predicting Unemployment Rate, Periods 1 & 2



Figures 9c and 9d: Predicting Unemployment Rate, Periods 3 & 4

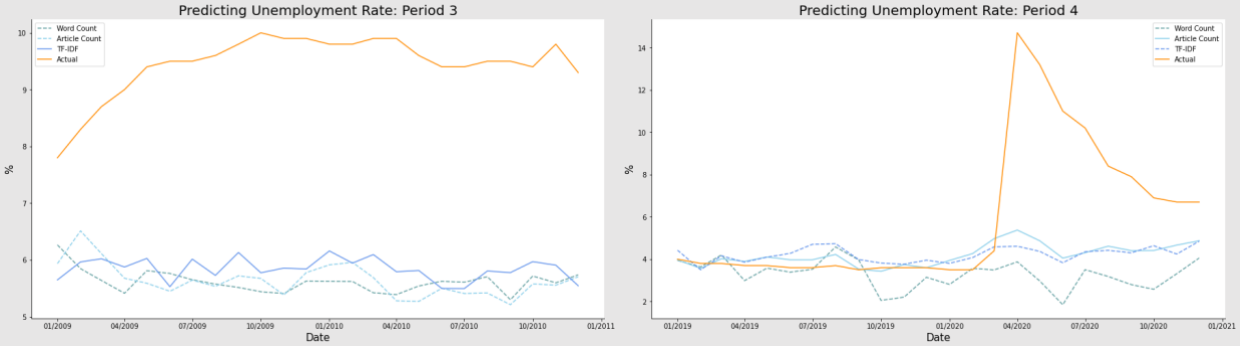


Table 7: RMSE of Elastic Net Models

	Period	Word Count (a)	Article Count (b)	TF-IDF (c)
Consumer Sentiment (1)	1	7.86	8.83	7.73
	2	13.80	14.09	13.72
	3	10.26	10.71	7.10
	4	15.76	12.25	12.23
	Avg.	**11.97**	**10.94**	**10.29**
Unemployment Rate (2)	1	0.87	0.55	0.87
	2	1.47	1.44	1.42
	3	3.88	3.86	3.64
	4	4.32	3.42	3.61
	Avg.	**3.02**	**2.695**	**2.695**

5.1 Assessing performance of Best Performing Models

To understand the effectiveness of using text data from the New York Times to predict economic indicators, I must scale the average results to compare the RMSE of consumer sentiment models to models predicting the unemployment rate. The lowest average RMSEs across all four periods for both consumer sentiment and the unemployment rate are obtained from the TF-IDF dataset. I will continue to refer to these models as 1c and 2c. Table 8 shows model 1c returns a significantly lower scaled RMSE than model 2c. I have evidence that quantifying the text in New York Times articles into TF-IDF values predicts consumer sentiment better than the unemployment rate.

Table 8: Scaled RMSE for Best Performing Models

	RMSE	Median test set	RMSE/Median
U of M Consumer Sentiment (1c)	10.29	90.8	11.3%
Unemployment Rate (2c)	2.695	5.3	50.8%

Following the methods of Kalamara et al. (2020), I can answer whether the text in New York Times articles accurately predicts economic indicators. I run a Diebold-Mariano Test to compare the predictive accuracy of all the best performing models for each economic indicator in each period to a baseline Autoregressive model with a lag of one— AR (1) — to determine if there is a statistically significant difference between the two. The results of the Diebold-Mariano Tests are displayed in Table 9 alongside the corresponding RMSE of each baseline model. There is only one model— model 1c in period 1— that performs better than a baseline model at a statistically significant level. These results counter the findings of Kalamara et al., who state, "forecasts of macroeconomic variables are improved by the addition of text," and Bortoli, who similarly find their penalized regression demonstrates superior performance compared to the autoregressive model. Therefore, it is reasonable to conclude predicting consumer sentiment and the unemployment rate using exclusively New York Times textual data is not feasible.

Table 9: Results of Diebold-Mariano Test

RMSE	Period	Dataset	AR (1) RMSE	Elastic Net RMSE	DM-Test P-value	Elastic Net Better?
U of M Consumer Sentiment (1)	1	TF-IDF	10.82	7.73	0.09*	Yes
	2	TF-IDF	10.14	13.72	0.00***	No
	3	TF-IDF	5.06	7.10	0.07*	No
	4	Article Count	11.03	12.25	0.86	
Unemployment Rate (2)	1	Article Count	0.73	0.55	0.26	
	2	TF-IDF	0.08	1.42	0.00***	No
	3	TF-IDF	1.41	3.64	0.00***	No
	4	Article Count	4.153	3.42	0.12	

Note: *** significant at 0.01 level, ** significant at 0.1 level

5.2 Comparing performance of models 1 & 2

As mentioned above, across all three datasets, text data from the New York Times is best at predicting consumer sentiment, followed by the unemployment rate. Connecting the findings of table 5, table 6, and table 8, I see models using a large number of strong coefficients will perform better than models using fewer, weaker coefficients. Model 1 uses an average of 936 words to make stronger predictions of consumer sentiment than the 204 coefficients used to make weaker predictions of the unemployment rate. This is due to the fact more coefficients can balance the bias of words correlated with economic contraction and words correlated with economic expansion. A similar pattern emerges when looking at the average penalized coefficient in Table 6, where the worst performing models, those predicting the unemployment rate, have the lowest average coefficient size, followed by consumer sentiment. Yet, because I only test two different economic indicators, it is challenging to reach a solid conclusion.

5.3 Comparing performance across models A, B, & C

Across the three datasets, one does not consistently outperform. Using RMSE to assess model performance, the TF-IDF best predicts consumer sentiment. For predicting the unemployment rate, TF-IDF and article count perform equally, which parallels the findings of Ammann et al. (2011), who construct an article count dataset to predict stock market returns; however, they did not utilize a TF-IDF dataset. The only consistent finding across all 24 models is that the word count dataset is never the best predictor of either economic indicator.

5.4 Comparing performance across periods

Predicting both consumer sentiment and the unemployment rate across all three datasets, the lowest RMSE was obtained in period one. Figures 8a and 9a show predictions that closely follow the actual test set values, including the drastic changes seen in both economic indicators

around 1990 when the global economy took a turn for the worse. Figure 4 shows training period 1 is exposed to every stage of a cyclical economy from a peak in July 1981 and a trough in November 1982. Further, Figure 3 shows the unemployment rate peaking in December 1982 and bottoming out in December 1998. Thus, when the economy begins to slow in July 1990, occurring in period one's test set, the predictions of all 4 models closely follow the changes in economic indicators (Hamilton, 2020). The worst performing models occur in period 4, likely due to the economic fallout from the coronavirus pandemic. The best predictor of a stalling economy during this test period would probably be *coronavirus*, but since it was rarely used before 2019, it was not included in the model. Interestingly, Figures 8c and 9c show in the runup to March 2020 predictions for consumer sentiment and the unemployment rate began to reverse course in a way that does not line up with where the direction of the actual economic indicators. Thus, while these models do not perform well, there is evidence that subtle changes in the monthly collection of words in the New York Times could prove powerful in forecasting economic indicators. That said, much of the poor performance of period four models has to do with the unexpected economic standstill brought by the coronavirus. The testing period ranging from 2009 to 2010 is the second-worst performing period. Figure 4 shows training period 3 is exposed to the entire business cycle, like training period 1, but the downturn in training period 3 lasted only eight months (March 2001 – November 2001) and neither consumer sentiment nor the unemployment rate reach levels realized in any other economic downturn. In the words of the St. Louis Federal Reserve, the 2001 recession "was relatively short and, by some measures, shallow" (Kliesen, 2003). The test set, however, covers the worst months of the Great Recession. By looking at the coefficients of the six models predicting consumer sentiment and the unemployment rate in period 3, it is clear they are ill-equipped to accurately predict a severe

downturn. Only model, 2a, uses the word *recession*, yet its coefficient is near-zero ($\beta = 0.004$). As discussed in the methods section, an increase in the frequency of *recession* in the New York Times is a pretty good indicator that a recession is going to occur. In period 1, the best performing period, coefficients for regression are significantly higher: in model 1a, $\beta = -0.147$; in model 1b, $\beta = -0.136$; in model 2a $\beta = -0.09$; in model 2b $\beta = -0.114$. Furthering the poor performance of models in period 3 is that additional words that are strong predictors of the two economic indicators, such as *unemployment*, have similarly low coefficients: $\beta = -0.085$ for model 1a and $\beta = -0.007$ for model 1b.

The reason for splitting up the data into four different training and testing datasets is to understand more about how the power of words shifts over time. Language evolves, evidenced by the fact we no longer speak in a Shakespearian manner. The coefficients of all 24 models further confirm this fact. Table 10 shows a breakdown in the percent of coefficients used in three or more periods. Models predicting consumer sentiment rely on more words over multiple periods than models predicting the unemployment rate. Due to the low l1-ratio of all models, thousands of words are used in one to two models, which inhibits models from dropping coefficients that may not have any predicting power.

Table 10: Percent of Coefficients Used in 3 or 4 Periods

	Word Count (a)	Article Count (b)	TF-IDF (c)
Consumer Sentiment (1)	18%	15%	16%
Unemployment Rate (2)	3%	3%	2%

5.5 Improving Model Performance

The reliance on only text data is an approach not seen frequently in the literature or the methods of professional forecasters, hence my curiosity to determine if it is possible. The results highlight the complicated and unreliable nature of using only text data to nowcasting economic

indicators. Other researchers, instead, use text data to complement other quantitative data. For example, Kalamara et al. (2020) use machine learning methods to predict several economic indicators using text data in conjunction with 33 macroeconomic factors. Other approaches, including, those taken by professional forecasters, rely on using lagged economic indicators variables of interest, meaning models use economic indicators from previous periods (Ammann et al., 2011; Bok et al., 2017; Bortoli et al., 2018; Choi, 2009; Giannone et al., 2010; Kalamara et al., 2020; Kholodilin et al., 2010). Further, the fact autoregressive models with lag equal to one perform better than the elastic net models reveals that the elastic net models may need some assistance from lagged economic indicator variables to steer predictions in the right direction. To further improve elastic net model performance, I also need to clear the road of distracting obstacles as well, namely the thousands of words with little-to-no predicting power. I can do this by increasing the l1-ratio to 1, so coefficients are determined entirely from the l1 penalty, which will reduce the coefficient of poor predictors to zero. The best alpha value is obtained by testing values between 0.1 and 1, selecting the value that minimizes training MSE when the l1-ratio equals 1. For models predicting consumer sentiment, both alpha and l1-ratio equal 1 while for models predicting unemployment, alpha equals 0.1 and the l1-ratio equals 1.

The impact of increasing the l1-ratio to one is seen in Table 11, which shows the number of coefficients used in each model drops from the hundreds to tens. Now, there are only 454 unique words used in these models, as opposed to 4,342 words when the l1-ratio equals 0.1. The lagged economic indicator is the most frequent feature, appearing in 21 of the 24 models.

Table 11: Average Number of Coefficients Across All 4 Time Periods

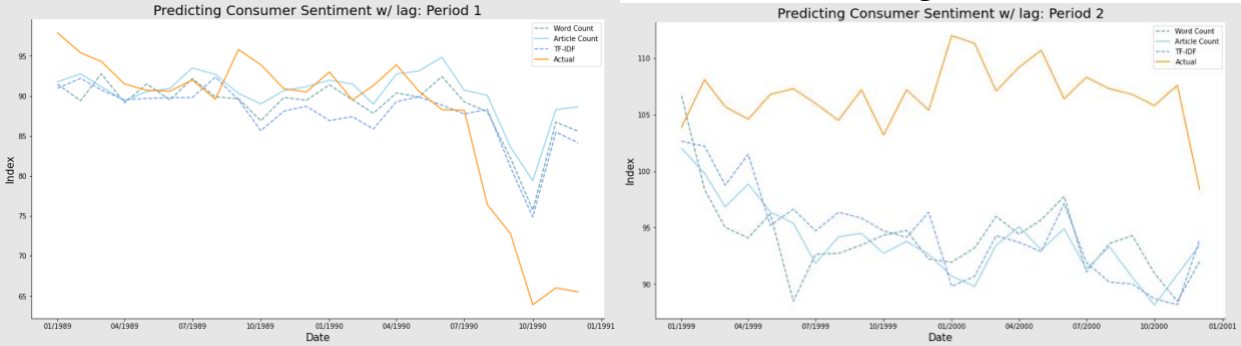
	Word Count (a)	Article Count (b)	TF-IDF (c)	Average
Consumer Sentiment (1)	20	18	21	20
Unemployment Rate (2)	40	31	33	34
Average	30	25	27	

Table 12: Results of Elastic-Net models with Lagged Economic Indicator.

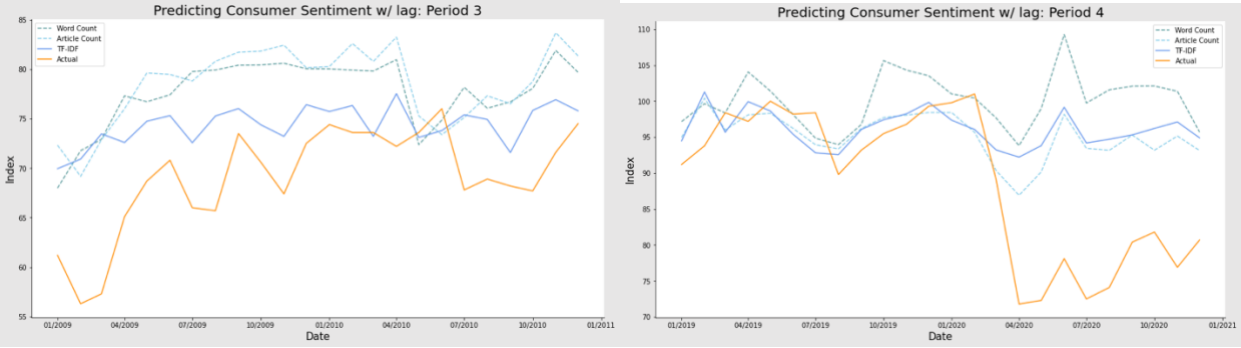
RMSE	Period	Word Count (a)	Article Count (b)	TF-IDF (c)
Consumer Sentiment (1)	1	4.84	4.61	4.69
	2	4.65	6.44	5.99
	3	4.61	4.17	3.96
	4	7.87	7.07	6.01
	Avg.	**5.66**	**5.70**	**5.23**
Unemployment Rate (2)	1	0.80	0.58	0.73
	2	0.90	0.90	0.96
	3	3.74	3.64	3.54
	4	3.51	3.02	3.05
	Avg.	**2.64**	**2.42**	**2.47**

The new results of all 24 models are shown in Table 12 and display a drastic drop in the RMSE of nearly every model and all model averages. After scaling the RMSE value, the results of which are displayed in Table 13, we again see textual data from the New York Times effectively predicts consumer sentiment and there is still a lackluster performance when predicting the unemployment rate. As mentioned in the literature review section, I propose an alternative measure of accuracy by comparing a model's performance to the consensus of professional forecasters. This approach will provide further evidence of whether the text in New York Times articles accurately forecast economic indicators. Each month, Reuters surveys top forecasters in financial markets on their predictions for several economic indicators, including the University of Michigan's Consumer Sentiment Index. Unfortunately, polling began in 1999, so I can only match 2/3s of the dates in my test set. The RMSE of professional forecasters for consumer sentiment is 3.42. While the RMSE for elastic net models with lagged economic indicators is 1.81 points higher than the accuracy of the professional forecasters, the results of further confirm the strong relationship between words in the New York Times and consumer sentiment.

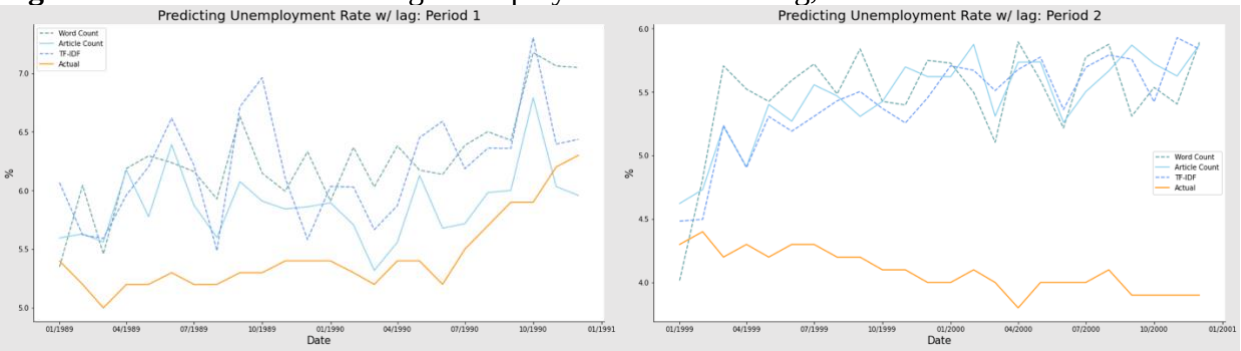
Figures 10a and 10b: Predicting Consumer Sentiment with lag, Periods 1 & 2



Figures 10c and 10d: Predicting Consumer Sentiment with lag, Periods 3 & 4



Figures 11a and 11b: Predicting Unemployment Rate with lag, Periods 1 & 2



Figures 11c and 11d: Predicting Unemployment Rate with lag, Periods 3 & 4

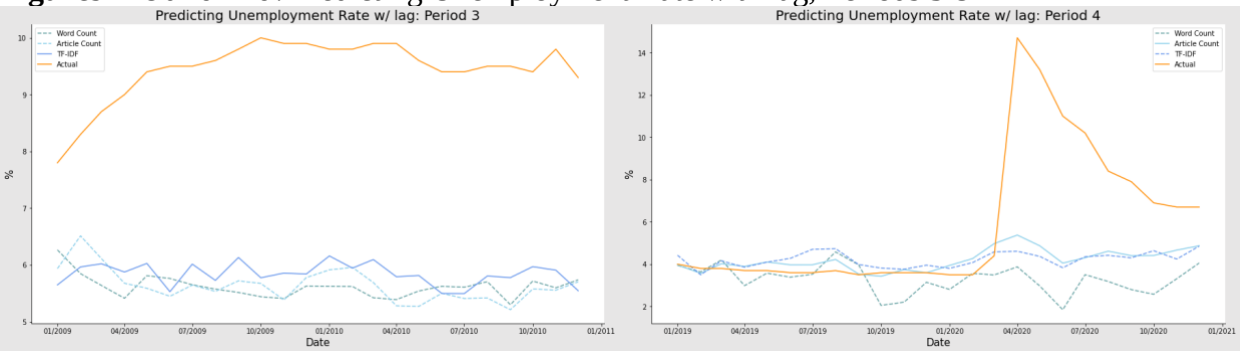


Table 13: Scaled RMSE for best performing models

	RSME	Median test set	RMSE/Median
U of M Consumer Sentiment (1c)	5.23	90.8	5.7%
Unemployment Rate (2c)	2.42	5.3	45.6%

In looking at the spread of words used across periods, there is only one model, model 2c, that uses a single word, *deficit*, more than 2 times. Meaning the remainder of the 453 unique words are used in only one or two periods. This shows that over time, there is a clear transformation in the meaning of words in the New York Times and the strength of those words to predict economic indicators. Thus, to create accurate nowcasting models using text data, researchers need to account for such shifts when training their models. That said, these new models reveal similar problems with training periods 3, in successful model performance is dependent on exposure to a broad range of economic conditions.

5.6 Training Models with Several Business Cycles

Table 10 shows that very few words are used to predict more than three periods of consumer sentiment or unemployment rate. After training models using only words included in 3 or more periods for their respective economic indicator and dataset[1], Table 14 shows the RMSE in nearly every period drops. There is a drastic drop in period 4 for both consumer sentiment and the unemployment rate, which, as previously discussed, is the worst-performing period because its training data does not encompass an entire business cycle. This proves a need to train models on the entire spectrum of the cyclical economy, which allow models to recognize and operationalize words that are strong predictors of good economic times, such as *construction* or *boom*, and bad economic times, such as *recession* or *downturn.* It does not matter how many

[1] These models do not include a lagged economic indicator

words are included in the model, but the right words, so when predicting unseen data, models are adequately balanced to predict the spectrum of changes an economy can experience.

Table 14: RMSE of Elastic Net Models using words that appear in 3 or more models

	Period	Word Count (a)	Article Count (b)	TF-IDF (c)
Consumer Sentiment (1)	1	10.02	8.17	8.70
	2	15.77	12.06	12.36
	3	8.783	5.37	6.43
	4	12.302	9.36	10.62
Unemployment Rate (2)	1	1.57	0.70	1.07
	2	1.52	1.59	0.89
	3	3.83	3.29	3.68
	4	2.23	2.12	2.59

So, what would happen if instead of splitting data into four different periods, models are trained over a longer period. To test this, I train models on the same data, except now it ranges from January 1981 to December 2012 and assessed using data from January 2013 to December 2020. Figure 11 provides a clear visualization showing models are trained over stretches of economic contraction and growth and assessed using data that covers the longest economic expansion in U.S. history and the fallout from the coronavirus pandemic (NBER, 2020). Given the improvement in model performance discussed in Section 5.5, these long-term models are trained using data that includes the lagged economic indicator and will have an l1-ratio of 1 to reduce the feature space. Alpha values are selected by minimizing the training RMSE and are equal to 0.1 for all six models.

Figure 12: Visualizing Long-Term Training and Testing Sets

Table 15: Results of Single Period Elastic-Net Model

RMSE	Word Count (a)	Article Count (b)	TF-IDF (c)
Consumer Sentiment (1)	5.23	4.26	4.37
Unemployment Rate (2)	1.84	1.67	1.57

The results are striking and present a clear path forward for future researchers. While the previous section highlights a clear difference in the words models rely on to make predictions across decades, we now see that when using text data to predict economic indicators what matters more is exposing models to a range of economic conditions. Comparing the results in table 15 to the average RMSE values in Table 12 shows improvement in every model. The RMSE of the professional forecasters' predictions of consumer sentiment over this same time is 3, so these results are still not nearly as accurate, but it is a drastic improvement from the results obtained when splitting data into four periods. In closing, the findings discussed throughout this section highlight the superiority of using textual data to predict consumer sentiment over the unemployment rate and the strength of the article count and TF-IDF datasets. While the last models created in this discussion section do not match the performance of professional forecasters, these results shed light on the power of using textual data to forecast economic indicators.

Figure 13a: Predicting Consumer Sentiment, Jan. 2013 to Dec. 2020

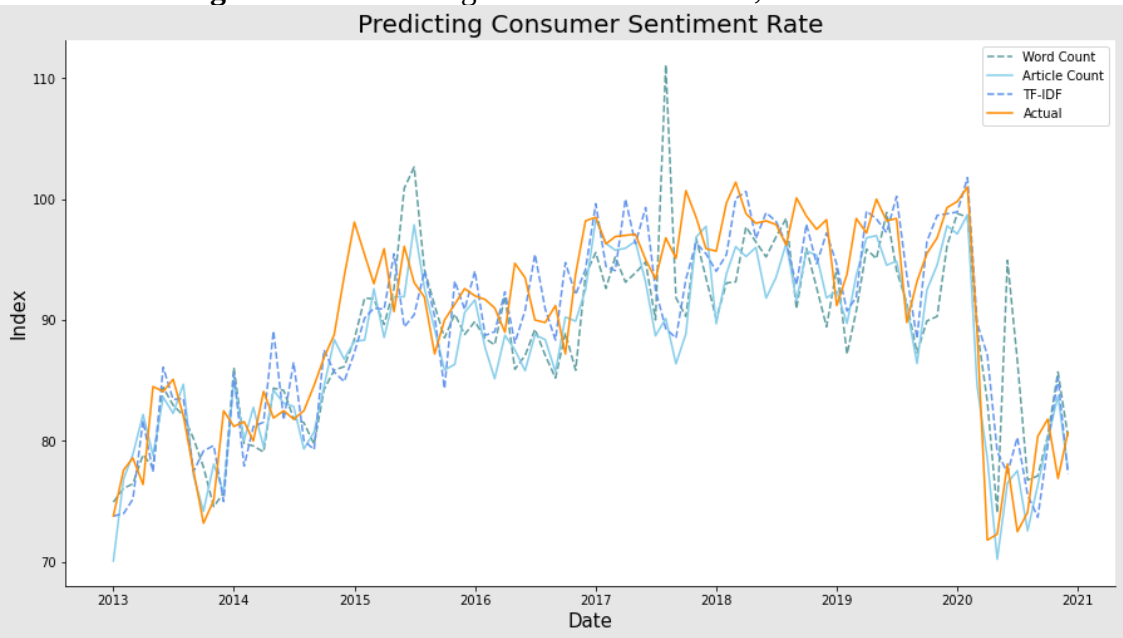
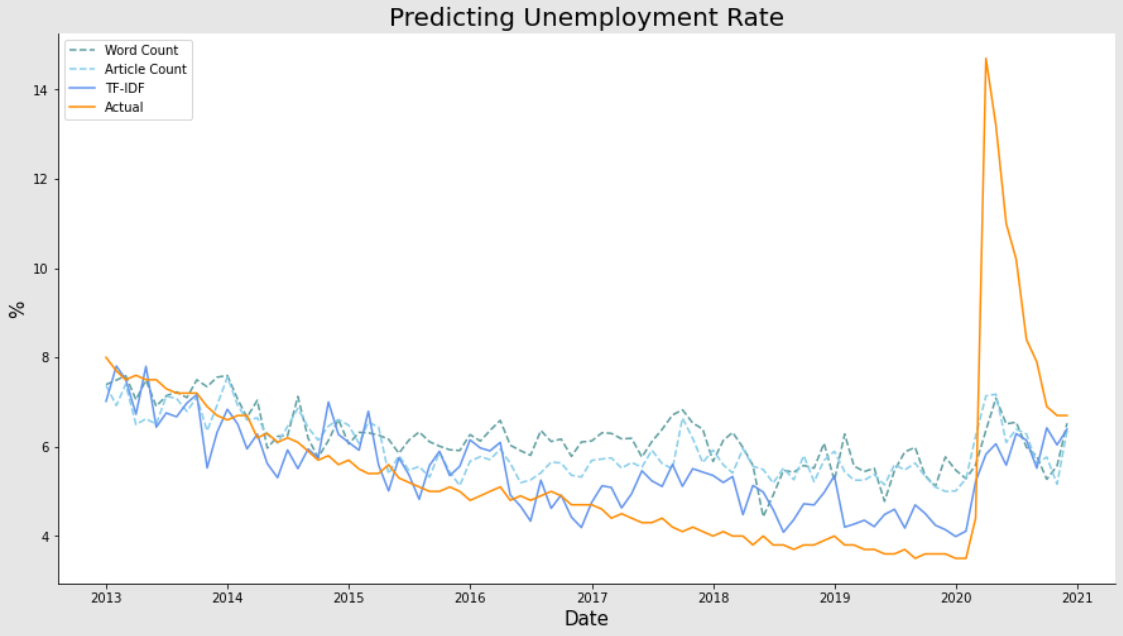


Figure 12b: Predicting Unemployment Rate, Jan. 2013 to Dec. 2020



6 Discussion

This paper, in positioning itself as a methodological intervention, seeks to shed light on a new source of data to assess the health of the economy. While the models in this paper do not explicitly nowcast, I do offer a robust assessment of the metrics and methods future researchers can use to nowcast economic indicators and reduce the length of time between a survey's reference period and its release. I chose to forecast consumer sentiment and the unemployment rate due to their holistic insight into the U.S. economy. To get here, I tested 20 different indicators and even more manipulations of those indicators to uncover which would be the best performing. I found no clear pattern on what makes an economic indicator more predictable using text data: consumer sentiment is a referendum of consumers, but similar metrics focused on consumers, such as personal expenditure, did not perform well in trial runs. While both Bortoli et al. (2018) and Kalamara et al. (2020) successfully forecast GDP using textual data, I was unable to produce any high-performing model. My results show that models using only text

data, text data with lagged economic indicators and a high l1-ratio, and models trained with long-term datasets are better at predicting consumer sentiment than the unemployment rate. This in part, may have to deal with the fact that consumers can be influenced by what they read in the New York Times—it presents a *what came first, the chicken or the egg* problem. This paper explores operationalizing textual data as a proxy for other economic data used in forecasting models and does not seek to understand the causality of these relationships. Consumers' beliefs on the health of the economy are important to monitor, given they account for over 67 percent of the United States' GDP. If the New York Times influences consumers, economic policymakers should be tuned in to such a relationship, for a shift in rhetoric could push the economy over the edge.

One takeaway from initial trial runs is the strength of the elastic net regression, which consistently performed better than ridge, lasso, partial least squares, and principal components regressions on nearly every trial. Elastic net regressions are a combination of ridge and lasso, allowing for some coefficients to drop to zero while exposing the models to a broad range of words that when taken as a whole can be useful in making predictions. This is similar to how a gradient boosted tree makes strong predictions from a collection of weak learners. Including thousands of words in the training process, however, can also introduce distracting noise that impacts model performance as seen when comparing the results of models with an l1-ratio of 0.1 to models with an l1-ratio of 1. Further, another takeaway from the results section is the superiority of using article count, which is the proportion of articles including a word in a given month, and TF-IDF datasets to predict economic indicators. Across all 48 models, the best-performing models never use the word count dataset. In the data section, I presented a hypothesis that using a large number of coefficients assists models in making predictions. Well, that may be

true when using just textual data, models that include the lagged economic indicator reveal that more words don't matter, but rather including the right words with strong coefficients will perform better on unseen data.

The guiding question of this thesis was to determine whether text data and text data alone could be used to predict national economic data. My findings show that autoregressive models with a lag of one perform better than elastic net models reliant solely on text data from the New York Times. Comparing the accuracy of models using just text data to the results of professional forecasters is like comparing apples to oranges: while both are forecasts, they take fundamentally different approaches to reach a conclusion. Hence my decision to introduce lagged economic indicators into the training process. While these models perform better than models without lagged economic indicators, they fail to match the accuracy of the Reuters poll. For models to be strong predictors, they need to be exposed to periods of economic growth and contraction when trained, hence why models predicting economic indicators in periods 3 and 4 perform terribly. Building upon each of the results across datasets, economic indicators and probing my findings, I conclude by training an elastic-net model using a lagged economic indicator in conjunction with textual data over a period that spans multiple cycles of economic growth and contraction with a high l1-ratio to reduce the number of coefficients is most effective at predicting consumer sentiment and the unemployment rate.

7 Conclusion

This thesis investigates the predicting power of textual data from the New York Times to nowcast two different economic indicators. The government relies on a wide range of methods to gather data that goes into calculating economic indicators, helping economists, policymakers, and researchers understand the health of the U.S. economy. Collecting good data is an expensive

and time-consuming process; as a result, there is a delay between when data collection occurs and when results go public. With an economy that contracts after a stock sell-off, the collapse of an investment bank, or a once-in-a-generation pandemic, not having a real-time assessment of economic health to guide monetary and fiscal policy can be the difference between a contraction lasting a few months to a few years. Economists have developed methods for nowcasting economic activity using other economic indicators (Giannone et al, 2008 & Bok, 2018). This thesis nowcasts using textual data, an approach rooted in the assumption that national economic indicators and the New York Times both capture Americans' economic reality, one quantitatively, one qualitatively.

The results of nowcasting using only textual data are mixed. Consumer sentiment proves most predictable, likely due to the New York Times' focus on stories that engage and relate to their subscriber base of consumers. Along these lines, I would assume a text from a publication such as the Financial Times, which focuses on business and economic current affairs and caters to an audience of corporate decision-makers, could predict metrics measuring business sentiment. On the other hand, the elastic-net models using only text data to predict the unemployment rate perform weakly. Further, I find using word count is a weak predictor for both economic indicators; future researchers should prioritize using the proportion of articles or a TF-IDF metric in text-reliant forecasting models. The discussion of the results explores how using test datasets that only cover periods of economic growth hampers the accuracy of predicting consumer sentiment and unemployment rate. I also discuss the improvements in consumer sentiment and unemployment predictability when training models with lagged economic indicator values. Finally, I show how the convergence of both these findings drastically improves model performance and rivals the performance of professional forecasters, setting a path for

future researchers to follow. Overall, the progression of thought in the results and discussion sections show text data has potential but is most powerful when used alongside lagged indicators and trained on data covering all stages of the cyclical economy.

There are three limitations in this study to note. First, the process of stemming words is not perfect. For example, the root of the word *economic* is *economy*, but both works remain in the final set of 6,307 words, and both have two of the largest coefficients used across all nine models. Second, in trial runs to determine the optimal alpha and l1-ratio combinations, I never test alphas greater than 1; some models might perform better if penalties are multiplied by a larger constant. Lastly, I only assess the performance of 5 different models— lasso, ridge, principal components, partial least squares, and elastic-net— and some other types of models could perform better than an elastic-net model. Kalamara et al. (2020) find "(non-linear) neural networks consistently perform the best across text sources, time horizons, and target variables", due to the fact they "are not as constrained in how they use the features fed into them", as opposed to elastic-net models.

Despite these limitations, this thesis hopes to advance the conversations on how textual data can predict economic indicators to reduce lags and inefficiencies in data collection. Researchers interested in using text data in their analysis, but specifically for predicting economic indicators, should consider taking additional steps to improve accuracy. One obvious step would be to explore the differences in performance between elastic net penalized regression and neural networks. Additionally, collecting N-grams to use when training models gives more context to singular words. For example, the positive coefficient for *unemployment* implies as the frequency of "unemployment" in the New York Times increases, the unemployment rate also increases. Using N-grams could split this trend by showing when the word *high* or *low* precede

unemployment. These two N-grams would likely have an inverse relationship with the unemployment rate and provide the model with extra information to improve predictions. Next, as you could see from many of the figures, the text data metrics, and especially TF-IDF, are very volatile with steep increases and decreases month to month. On the other hand, economic indicators are rather smooth. Using moving averages, running mean, or kernel methods to transform textual data could remove much of the noise and improve model accuracy. The last recommendation is to assess if text data can forecast economic indicators further out than this paper attempted. I use text data from t=1 to predict economic indicators from t = 1; it would be fascinating to explore the accuracy of text data from t = 1 in predicting economic indicators at t > = 2. Similarly, including broader lagged economic indicators— from two, three, or four periods back— in models should also be assessed to determine the impact on error reduction. In closing, this thesis takes the first step in showing the capabilities of textual data, opening the door to a future where newspaper articles, or other sources of data not collected by the federal government, can be assessed daily to nowcast economic indicators. Researchers are pushing the bounds of what data can be; forecasters must continue to follow these trends and operationalize new data that can further improve our insights into economic health.

Works Cited

Askitas, N. and K. F. Zimmermann (2009). Google Econometrics and Unemployment Forecasting. Applied Economics Quarterly, 107–120.

Asur, S., & Huberman, B. A. (2010). Predicting the Future with Social Media. *IEEE*.

Billion Prices Project (n.d.). About. www.thebillionpricesproject.com/.

Boschan, C. (1966). Job Openings and Help-Wanted Advertising as Measures of Cyclical Fluctuations in Unfilled Demand for Labor. National Bureau of Economic Research, 491-518.

Choi, H., & Varian, H. (2009, April). Predicting the Present with Google Trends.

Choi, H., & Varian, H. (2009, July). Predicting Initial Claims for Unemployment Benefits.

Dunn, M., Haugen, S., Kang, J. L. (2018, January). The Current Population Survey—tracking unemployment in the United States for over 75 years. *Monthly Labor Review.* U.S. Bureau of Labor Statistics. https://doi.org/10.21916/mlr.2018.4.

The Economist (1998). The Recession Index. www.economist.com/britain/1998/12/10/the-recession-index.

The Economist (2008). R-word Index Warning Lights. www.economist.com/finance-and-economics/2008/01/10/warning-lights.

Einav, L., & Levin, J. (2013, May). The Data Revolution and Economic Analysis. *National Bureau of Economic Research*, working paper 19035. www.nber.org/papers/w19035

Fama, E. F., & Kenneth R. French. (1996). Multifactor Explanations of Asset Pricing Anomalies. *Journal of Finance.*

Further Unemployment Relief through the Reconstruction Finance Corporation, S. 5336: A Bill to Amend the Emergency Relief and Construction Act of 1932, 72nd congress. 142-144 (1933, February) (testimony of John A. Ryan).

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 535-574.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *Journal of Monetary Economics*, 665-676.

Gurr, Ted R. (1968). A Causal Model of Civil Strife: A Comparative Analysis Using New Indices. *American Political Science Review*, 1104-1124.

Hamilton, J. (2020). Dates of U.S. recessions as inferred by GDP-based recession indicator [JHDUSRGDPBR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/JHDUSRGDPBR

Isidore, C. (2008, December 1). It's official: Recession since Dec. '07. www.money.cnn.com/2008/12/01/news/economy/recession/.

Jenkins, J. Craig, and Craig M. Eckert. (1986). Channeling Black Insurgency: Elite Patronage and Professional Social Movement Organizations in the Development of the Black Movement. *American Sociological Review*, 812-829.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2020). Making text count: economic forecasting using newspaper text. *Bank of England*.

Kennedy, C., Hartig, H. (2019, Feuruary). Response rates in telephone surveys have resumed their decline. Pew Research Center. www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/.

Keynes, J. M. (1936). *The general theory of employment, interest and money*. Macmillan Publishers.

Kholodilin, K., Podstawski, M., & Siliverstovs, B. (2010). Do Google Searches Help in Nowcasting Private Consumption?: A Real-Time Evidence for the US. *German Institute for Economic Research*.

Kliesen, K. (2003). The 2001 Recession: How Was It Different and What Developments May Have Caused It? *Federal Reserve Bank of St. Louis Review*.

Li, C., Chen, L. J., Chen, X., Zhang, M., Pang, C. P., & Chen1, H. (March 2020). Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data. *Euro Surveill*. https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000199.

Lohr, S. (2013, September 6). A Brief History of Data Revolutions in Economics. *New York Times*. www.bits.blogs.nytimes.com/2013/09/07/a-brief-history-of-data-revolutions-in-economics/

Mastercard (n.d.). Mastercard SpendingPulse. www.mastercardservices.com/en/solutions/data-networks/spendingpulse.

Miller, J. (1998, January 26). Booming Stock Market Changes Rankings in Philanthropy. *The New York Times.* Retrieved from https://www.nytimes.com/1998/01/26/us/booming-stock-market-changes-rankings-in-philanthropy.html?searchResultPosition=18

National Bureau of Economic Research (1951). U.S. Cotton Crop 1798-1955 [Data file]. www.nber.org/research/data/nber-macrohistory-i-production-commodities.

National Bureau of Economic Research. (2020, June). Business Cycle Dating Committee Announcement June 8, 2020. www.nber.org/news/business-cycle-dating-committee-announcement-june-8-2020.

Naughton, B. (2005, Fall). The New Common Economic Program: China's Eleventh Five Year Plan and What It Means. *China Leadership Monitor*, No. 16.

The New York Times. Journalism. Retrieved From https://www.nytco.com/journalism/.

Nguyen, K., & Cava, G. L. (2020, June). News Sentiment and the Economy. *The Reserve Bank of Australia*. www.rba.gov.au/publications/bulletin/2020/jun/.

Office for National Statistics. (2016, January). Census-taking in the ancient world. www.ons.gov.uk/census/2011census/howourcensusworks/aboutcensuses/censushistory/censustakingintheancientworld.

Pinchot, A. (1931, January). We Met Mr. Hoover. *Nation Associates*, 43-44.

Preis, T., H. S. Moat, and H. E. Stanley (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports* 1684(3), 1–6.

Richardson, G., Komai, A., Gou, M., Park, D. (2013, November 22). *Stock Market Crash of 1929*. Federal Reserve History. https://www.federalreservehistory.org/essays/stock-market-crash-of-1929.

Rosenthal, Caitlin. (Guest). (2018, November 13). *Why Management History Needs to Reckon with Slavery.* [Audio podcast]. hbr.org/podcast/2018/11/why-management-history-needs-to-reckon-with-slavery.

Sullivan, A. (1995, March 11). Retiring Baby Boomers Dread the End of the Boom Times. *The New York Times.* Retrieved from www.nytimes.com/1995/03/11/your-money/IHT-retiring-baby-boomers-dread-the-end-of-the-boom-times.html?searchResultPosition=9

Tanter, Raymond J. (1966). Dimensions of conflict behavior within and between nations, 1958-60. Journal of Conflict Resolution, 41-64.

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (June 2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, 1437-1467.

University of Michigan Survey of Consumers. (2001, January). December 2000 Monthly Survey. www.data.sca.isr.umich.edu/fetchdoc.php?docid=23607.

U.S. Bureau of Labor Statistics. (2015). How the Government Measures Unemployment. www.bls.gov/cps/cps_htgm.htm#where.

U.S. Census Bureau. (2000, May). *Factfinder for the Nation: History and Organization*. U.S. Department of Commerce, Economics and Statistics Administration. www.census.gov/history/www/census_then_now/.

U.S. Census Bureau. (2015, April). Current Population Survey Interviewing Manual. www.census.gov/programs-surveys/cps/technical-documentation/complete.html.

U.S. Bureau of Labor and Statistics (2021, October). Current Population Survey Non-Response Rates. www.census.gov/programs-surveys/cps/technical-documentation/methodology/non-response-rates.html.

Wan, X., Yang, J., Marinov, S. et al. (2021). "Sentiment correlation in financial news networks and associated market movements." Sci Rep 11, 3062 https://doi.org/10.1038/s41598-021-82338-6.