JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12370

ECONOMETRICS MEETS SENTIMENT: AN OVERVIEW OF METHODOLOGY AND APPLICATIONS

Andres Algaba

Faculty of Social Sciences and Solvay Business School
Vrije Universiteit Brussel
and Department of Economics
Universiteit Gent

David Ardia

Department of Decision Sciences
HEC Montréal

Keven Bluteau

Department of Economics
Universiteit Gent
and Department of Decision Sciences
HEC Montréal

Samuel Borms*

Faculty of Social Sciences and Solvay Business School
Vrije Universiteit Brussel
and Institute of Financial Analysis
University of Neuchâtel

Kris Boudt

Faculty of Social Sciences and Solvay Business School
Vrije Universiteit Brussel
Department of Economics
Universiteit Gent
and School of Business and Economics
Vrije Universiteit Amsterdam

*Corresponding author contact email: samuel.borms@unine.ch; Tel: +32 496 95 53 22.

Journal of Economic Surveys (2020) Vol. 34, No. 3, pp. 512–547

Abstract. The advent of massive amounts of textual, audio, and visual data has spurred the development of econometric methodology to transform qualitative sentiment data into quantitative sentiment variables, and to use those variables in an econometric analysis of the relationships between sentiment and other variables. We survey this emerging research field and refer to it as *sentometrics*, which is a portmanteau of sentiment and econometrics. We provide a synthesis of the relevant methodological approaches, illustrate with empirical results, and discuss useful software.

Keywords. Qualitative data; Sentiment analysis; Sentometrics; Survey; Textual analysis

1. Introduction

There is a long-standing tradition of using sentiment as either a parameter or a variable in econometric modeling. Historically, the use of questionnaires and proxies to quantify sentiment variables has been predominant. In recent years, it has become popular to analyze the sentiment embedded in textual, audio, and visual data. Such data are becoming increasingly available in large amounts due to the digitization of communication media. These media are carriers of potentially interesting information useful for economic analysis. This has spurred a new strand of econometric research that investigates the transformation of large volumes of qualitative sentiment data into quantitative sentiment variables, and their subsequent application in an econometric analysis of the relationships between sentiment and other variables. We refer to this emerging field as *sentometrics*, which is a portmanteau of sentiment and econometrics.

In this survey, we overview the methodology and applications related to an econometric analysis of sentiment extracted from qualitative data. We first define sentiment as the disposition of an entity toward an entity, expressed via a certain medium. Examples of entities include individuals, news media, companies, government associations, industries, and markets. This disposition can be conveyed numerically but is primarily expressed qualitatively through text, audio, and visual media. Sentometrics studies the computation of sentiment from any type of qualitative data, the evolution of sentiment, and the application of sentiment in an economic analysis using econometric methods. Many approaches already exist for using econometrics with textual data, as recently overviewed by Gentzkow *et al*. (2019a), but our focus on qualitative sentiment data is unique. The overview by Lewis and Young (2019) is limited to the most important analytical approaches for analyzing textual content in accounting and finance.

The goal of our survey is to provide a synthesis of the relevant work that serves as a gateway for researchers in econom(etr)ics, finance, and machine learning interested in the analysis of qualitative sentiment data. The survey takes a hands-on approach by synthesizing the research around the common challenges. The first critical step is the clarification of the problem that one is trying to solve. In function of the question, one collects, prepares, and selects the different data. The filtered qualitative data are then transformed into numbers using domain-specific sentiment quantification techniques. These numbers are next aggregated into meaningful sentiment variables. The different intermediate aggregation steps involve combinations of sentiment calculation methods, as well as the use of various within-text, across-text, and across-time aggregation methods. These variables are used as main input in an econometric model that is set up to solve the question at hand. An important part of the econometric analysis of qualitative sentiment data is a continuous validation activity. For each of the above topics, we discuss the relevant methodological approaches and illustrate with empirical results. We also have a section that sums up some of the available software to perform each step.

2. Definition of Sentiment

The term "sentiment" is used in many different contexts and research areas, but there is no established definition. We propose a working definition that encapsulates the most important characteristics of sentiment from the perspective of a researcher wishing to transform textual, audio, and visual data into

sentiment variables and to apply them in an economic analysis. We also summarize the literature that highlights other characteristics and alternative definitions of sentiment.

2.1 *Working Definition*

We propose the following generic definition of sentiment:

> Sentiment is the disposition of an entity toward an entity, expressed via a certain medium.

This working definition of sentiment embeds three components. First, the expression by an entity of a disposition in the form of verbal or nonverbal communication. To observe the private state of mind of any entity, one has to look at their subjective expressions through both qualitative sentiment data such as textual, audio, and visual data, as well as numeric data such as quantitative survey data and stock market data. The combined use of different sources of qualitative sentiment data is called multimodal sentiment, compared to the use of one data source, which is referred to as unimodal sentiment. Not many studies have assessed the added value of multimodal sentiment, but, in general, findings confirm an increased level of accuracy over unimodal sentiment (see, e.g., Soleymani *et al*., 2017). Despite the fact that there are differences between, for example, the expression of emotion and sentiment (or other human subjective terms), it is not necessarily in the interest of a researcher interested in sentiment to make this distinction explicit. Second, the disposition has a measurable polarity or semantic orientation that shows through the medium of expression. It reveals the direction and intensity of the subjective expression, on a discrete or a continuous scale. Many definitions simply use positive or negative to indicate semantic orientation, but application-specific terminologies, such as bullish and bearish (Antweiler and Frank, 2004), dovish and hawkish (Picault and Renault, 2017), or Democrat and Republican (Gentzkow and Shapiro, 2010), can be helpful. Sentiment is usually asserted at different levels of granularity (e.g., a sentence, an entire article, a sequence of sounds, or an image). Third, the sentiment is oriented toward (an aspect of) another entity, or exceptionally the expressing entity itself.

This general synthesis encompasses a broad range of different sentiment definitions that are currently used in the field of economics. Casey and Owen (2013) describe consumer confidence as the consumer expectations about the future state of the economy. Ludvigson (2004) notes that surveys are most often used to measure consumer confidence. De Long *et al*. (1990) define investor sentiment as a belief about future cash flows and investment risks that is not justified by the fundamentals. Baker and Wurgler (2007) list several mediums through which this investor disposition is expressed. Notable examples include the use of investor surveys, proxies for investor mood changing variables such as the number of hours of daylight, and the analysis of market data such as trading volume, implied volatility, mutual fund flows, the premium on dividend-paying firms, and the closed-fund discount. In a survey on sentiment in finance, Kearney and Liu (2014) argue that there are two types of sentiment—namely, investor sentiment, which includes only the subjective judgments and behavioral characteristics of investors, and text-based or textual sentiment, which may also contain a more objective reflection of the conditions of a certain entity. Kräussl and Mirgorodskaya (2017) hypothesize that media sentiment translates into investor sentiment. Moreover, Chang *et al*. (2015) say that sentiment affects the formation of investors' beliefs and thereby their reactions to information shocks.

In the remainder of the paper, we focus on the use of qualitative sentiment data as the medium through which the sentiment is expressed.

2.2 *Other Definitions*

From a psychological viewpoint, Munezero *et al*. (2014) state that sentiment is one of the so-called human subjectivity terms that reflects a person's desires, beliefs, and feelings. These human subjectivity

terms are features of a person's private state of mind that can only be observed through textual, audio, or visual communication. For instance, the tonality of one's voice, or the frequency and length of pauses are informative about underlying sentiment. The same holds for pitch, facial and bodily gestures, the word use in a written article, and the colors present in a picture. Other human subjectivity terms include affect, feeling, emotion, and opinion. While Munezero *et al*. (2014) argue that there are some notable differences, these terms are mostly used interchangeably in different strands of the literature. For instance, the distinction that sentiment involves enduring emotional dispositions toward an object, whereas emotions are briefer, is not of direct interest for most economic applications. Based on social theory, Evans and Aceves (2016) classify sentiment as a human state that reflects its condition at a given time and place. Other states include preference, uncertainty, and ideology, among others.

Taboada (2016) details linguistic sentiment as the expression of subjectivity either as a positive or as a negative opinion through language. Soleymani *et al*. (2017) define sentiment as a long-term disposition with a certain polarity toward an entity. From a text mining perspective, Taboada *et al*. (2011) treat sentiment as equivalent to semantic orientation, containing an evaluative factor (i.e., positive or negative) and a corresponding strength. In a survey on sentiment analysis, Liu (2015) makes no distinction between sentiment and opinion and defines an opinion as a quintuple of (1) the expressed sentiment, (2) the entity toward which it is expressed, (3) the particular (aspect of the) entity that is mentioned, (4) the opinion holder, and (5) a time stamp. This definition is closest to ours. Van de Kauter *et al*. (2015) distinguish between explicit sentiment (conveying subjective private states) and implicit sentiment (conveying factual information). Shapiro *et al*. (2018) deploy a characterization of emotions along the two dimensions valence (how positive) and arousal (how charged). The valence of word-of-mouth in marketing research is referred to by Gelper *et al*. (2018) as a discrete or continuous metric that captures the attitudes toward a brand. Additionally, sentiment can be looked at from the perspective of the sender (e.g., the sentiment attached by the author of a text), and from the perspective of the receiver (e.g., the sentiment perceived by the average reader).

In their analysis of political sentiment, Grimmer and Stewart (2013) use both the terms "sentiment" and "tone." They determine tone based on whether information is conveyed positively or negatively. Tone is more often used in the accounting and finance literature. Bajo and Raimondo (2017) characterize tone of news as a combination of the degree of positiveness, negativeness, and uncertainty. Feldman *et al*. (2010) define tone as the optimism or pessimism of the information embedded in qualitative verbal disclosures. Henry (2008) defines tone in earnings press releases as the effect of a communication. In the construction of a news-based coincident index of business cycles, Thorsrud (2018) uses tone as a synonym for sentiment and identifies it by determining whether news articles are positive or negative. In this survey, we also treat tone as a synonym for sentiment.

3. Problem Definition

Sentiment data have the potential to help solve or understand many problems involving the use of econometrics, across the fields of economics, finance, accounting, marketing, psychology, and computer science, among others. The adequate choice of methods to analyze sentiment data depends on the goal of the analysis. A common ground for econometrics applied to sentiment analysis is that one first needs to measure sentiment. Below, we discuss the use of qualitative sentiment data in applied economic theory and as an information source in nowcasting and forecasting economic variables.

3.1 *The Role of Qualitative Sentiment Data in Applied Economic Theory*

At least since the work of Keynes (1936), economists have been wondering what role sentiment plays in influencing economic decision-making. Understanding the relationship between sentiment data and

decision-making at the micro and macro level is still important in economic theory. Sentiment can be considered either to contain fundamental information in the news sense or to capture irrationality up to "animal spirits" in the noise sense. Both types of shocks move economic expectations and market outcomes at different horizons (Angeletos *et al*., 2018). The challenge according to Angeletos and La'O (2013) is that economists first have to model and then to quantify the sentiment forces behind the formation of market expectations. They develop an economic theory that does not depart from rationality but rather connects market expectations with market outcomes through external shocks they call sentiments. Barsky and Sims (2012) create a dynamic stochastic general equilibrium (DSGE) model that accommodates both the information and animal spirits view of confidence. They find most empirical evidence for the perspective that innovations in confidence reflect information about future economic prospects. One explanation for the importance of economic sentiment is that it acts frequently as a self-fulfilling prophecy (see Petropoulos Petalas *et al*., 2017, and the references therein). When there is consumer or business pessimism about economic growth, actual negative growth can be a direct consequence of it. A more specific example is bank runs. When too many depositors' sentiment about other depositors is negative, the unwanted outcome, a bank run, is more likely to materialize (Diamond and Dybvig, 1983).

In this regard, sentiment indices based on qualitative data can provide a more direct data-driven instrument to assess various types of economic shocks, or proxy for matters such as confidence or expectations. Larsen and Thorsrud (2019) use structural vector autoregression (VAR) to identify news and noise shocks in a panel of text-based measures and other economic variables.

Sentiment proxies provide a way to test behavioral hypotheses on the aggregate level or on the individual level. In general, the key questions pondered in a behavioral analysis are "What drives sentiment?" and "What is the behavioral impact of the sentiment transmitted?" An example of a behavioral hypothesis is whether entities inflate the tone in their written communication to influence market reactions. Both Picault and Renault (2017) (for the European Central Bank) and Arslan-Ayaydin *et al*. (2016) (for firms) validate this hypothesis. Sentiment in texts can be argued to be driven by a self-interest to generate particular external outcomes. In Garz (2014), the evidence shows a strong bias in terms of the number of negative and positive reports related to unemployment that is not the consequence of an asymmetric interpretation of the official numbers but rather associated with noneconomic information and the process of news production itself. Along these lines, the degree of sentiment involved in images appended to advertisements can also have clear intentions to impact customer behavior. Kalogeropoulos (2018) studies the impact of various media outlets on individual economic expectations, not finding tone to be a good predictor. In the finance literature, behavioral theory predicts that short-horizon returns are reversed in the long run (Tetlock, 2007).

There is a growing concern about the severity and impact of media bias. Closely related, Flaxman *et al*. (2016) explain that there are two strands of thought about the impact of improved production, distribution, and discovery of news articles (or generally, any multimedia). Some defend it increases exposure to diverse perspectives; others argue that it increases ideological segregation. They find empirical support for both camps, and thus, a further investigation of the impact of a biased media production and consumption behavior would be worthwhile.

Boudt and Thewissen (2019) base their analysis of CEO letters on psychological phenomena such as framing (Tversky and Kahneman, 1981) and the serial position effect (Glanzer and Cunitz, 1966). These and similar phenomena can also be used to better understand the sentiment conclusions. Purely as an illustration, stronger weights of negative sentiment words in comparison to those of positive sentiment words could be supported by the negativity bias, claiming that negative things have a greater impact. It could also be related to the fact that news media tend to emphasize negative news (e.g., Lowry, 2008). The priming, agenda setting, and framing communication theories described in Scheufele and Tewksbury (2007) can also be subject to more precise testing using multimodal sentiment measures.

Many different entities can have sentiment attributed from different data sources. Sentiment across these different entities tends to interact in particular ways, possibly in a contagious manner. The assessment of these sentiment flows, the feedback effects, and the associated information dispersion over time concern

a network analysis of sentiment. The Global Database of Events, Language, and Tone (GDELT) project[1] is the most comprehensive effort to date of a global network analysis of events and related sentiment. These data and social media data are useful to analyze what network structures would mitigate behavioral misperceptions, a question brought up by Teoh (2018).

Larsen and Thorsrud (2018) use a graphical Granger causality modeling framework to gain insights into the network of economically relevant news topics. Every node in the graph represents a sentiment/topic time series. The graph can be used to detect which narratives dominate and what the degree of news spillover is – that is, what news stories from which countries Granger cause the occurrence of any other news stories.

Eshbaugh-Soha (2010) emphasizes the role news coverage and tone can have on government trust and how it is central to explaining effective leadership. News provides country leaders a means to communicate their messages, but the local perception can differ significantly. Therefore, an interesting study would be to assess the spread between the sentiment of news reported in one region and the sentiment of similar news reported in another region.

3.2 *Measuring, Nowcasting, and Forecasting of and with Sentiment*

Sentiment time series indices aim to reflect the evolution of sentiment over time. A well-known text-based example is the economic policy uncertainty (EPU) index of Baker *et al*. (2016).[2] This index measures uncertainty, a specific type of sentiment. Manela and Moreira (2017) create a news-based measure of option-implied uncertainty, arguing that it incorporates disaster concerns expressed via the media. In many other applications, sentiment is also considered an explicit or implicit proxy for a certain desired output, such as for the visualization of company reputation (Saleiro *et al*., 2017).

We refer to quantifying already observed sentiment as sentiment measurement, while the prediction of the unobserved current and future sentiment is called sentiment nowcasting and forecasting, respectively. Sentiment is a latent variable, meaning that it is not readily observable. Measuring sentiment is a key task in any sentiment-based analysis.

In the now and forecasting literature, sentiment measures are considered a timely driver of other variables. There are three approaches to the type of sentiment that is used. Sentiment is proxied using available (questionnaire-based) indices, sentiment is constructed itself from a qualitative data source commonly using relatively simple methods, or sentiment is bought from a data provider such as Reuters (e.g., Thomson Reuters MarketPsych Indices) who in general uses a more complex methodology for the computation. The obtained sentiment is then transformed for usage in prediction models, to obtain the best possible prediction at any time. Sentiment variables are rarely used alone as explanatory variables but are usually added to a set of standard explanatory variables to see whether its integration improves or deteriorates forecasting performance.

The integration of sentiment has indeed already shown its capacity to improve forecasting performance. A significant impact of sentiment expressed through diverse media on stock returns and trading volume is found by Heston and Sinha (2017), Jegadeesh and Wu (2013), Tetlock *et al*. (2008), Tetlock (2007), and Antweiler and Frank (2004). Ardia *et al*. (2019b) incorporate textual sentiment time series into the long-term forecasting of the U.S. industrial production growth rate using sparse regression techniques.

Beyond improved predictions, using sentiment data is very flexible and timely, especially compared to traditional sentiment extraction methods such as surveys. Changes in sentiment methodology can easily be backtested using the available data. Modifying the structure of a survey, however, necessitates the survey to be sent out again to obtain new results. Information derived from sentiment data hardly suffers from release lags, making timely sentiment an ideal variable to enhance nowcasting models and consequently to craft timelier policy responses.

As in Hamilton *et al*. (2016), sentiment analysis on word level can be used to measure the time-varying perception surrounding certain words. For instance, "terrific" had a negative connotation up to 1960, but then became more positive. Lukeš and Søgaard (2018) find that words predictive of sentiment at one point in time remain not necessarily equally predictive at a later point, and that models trained on old data perform worse than models trained on recent data. They suggest a predictive feature selection approach to deal with temporal polarity shifts. The implication of changes in language over time is that the methods of sentiment quantification should evolve with it.

It is becoming well established in economics and finance that adding soft (qualitative) information on top of hard (quantitative) information results in predictive information gains. However, the soft information is usually explored through textual content. Audio and visual content have been explored less so but may deliver additional information value, according to Mayew and Venkatachalam (2012), who find that vocal cues of managers during conference calls predict a firm's future performance.

4. Qualitative Sentiment Data

The various ways in which economic agents express their sentiment leads to textual, audio, and visual sentiment data. Sentiment data are short for "sentiment-bearing" data. Most of the examples and methods in the remainder of this survey focus on textual data, because audio, and visual sentiment analysis is still in its infancy (Soleymani *et al*., 2017). Teoh (2018) does so similarly but acknowledges the rising relevance of audio and visual data. Currently, the main focus of current research in sentometrics effectively lies with textual data due to their wide availability in the digital form of news media articles, company filings, or social media posts (see, e.g., Loughran and McDonald, 2016).

The choice for textual data thus comes from the fact that most research and applications have been developed for this type of data. An advantage of focusing on texts is that many other forms of unstructured data can be transformed into textual data and then analyzed as if they were textual. For instance, audio data are often transcribed into textual data and can thus be analyzed using tools from this domain.

Multimodal sentiment analysis techniques are expected to gain importance due to the Internet, which has become more of a widespread multimedia platform. Where possible, we highlight close relations between the analysis of textual data and the analysis of audio and visual data, covering potential similarities from one data-type approach to the other. Doing so, we outline a uniform high-level framework that is applicable to all these data sources. The concepts of feature extraction, quantification, aggregation, modeling, and validation are very much transferable, though almost never presented as such.

4.1 *Information Sources*

The information sources for sentiment analysis in econometrics can be grouped in two ways. First, it can represent where the data were published. This includes news outlets (a journal, a social media channel, YouTube, a vlog, or a blog), companies and governments (regarding the publication of an official press release or an official report), or publication venues (an academic journal or a book publisher), among others. The source in this context should not be confused with the actual expresser of the sentiment; for instance, the source can be a journal, and the expresser a company or one of its top managers.

Second, it can represent from where the data are retrieved. The largest worldwide textual data providers are LexisNexis, Dow Jones' Factiva, and Reuters. Access to these databases is paid. A cheaper alternative, if allowed, is to scrape textual data from the web. A specific scraping procedure needs to be set up, which is a cumbersome activity, and in general goes with a considerable degree of hit-and-miss in terms of texts successfully collected. There also circulate some freely available data sets – for instance, the eight text

data sets analyzed by Zhang *et al*. (2015) or the list of freely available text data sets provided by Ravi and Ravi (2015).[3]

The acquisition of the data requires a good data management system, able to structurally store many gigabytes, such as MySQL. The database should also have fast query functionalities, for example, delivered by technologies such as Solr or Elasticsearch.

4.2 *Alternative Sentiment Variables*

Instead of the algorithmic extraction of sentiment from data, sentiment is also often proxied by asking people through surveys. The U.S. Consumer Confidence Index or the European Economic Sentiment Indicator is actively monitored examples of indices based on surveys (see, e.g., the analyses of Ludvigson, 2004, and Gelper and Croux, 2010). However, surveys have the downside of being costly, are hard to replicate, have a publication lag, and cannot be backtested. Both survey-based measures and data-based measures have their value, and are in many cases complementary. Ardia *et al*. (2019b) show that the specification that includes both time series measures generates the best out-of-sample predictive power. Baker and Wurgler (2006) derive a sentiment index through a principal component procedure from six sentiment proxies proposed in the literature, without going through any sentiment quantification process themselves.

Similar to textual data providers, there exist a number of textual sentiment data providers. Two often-used solutions are the series from RavenPack, and the Refinitiv (formerly Thomson Reuters) MarketPsych Indices.

4.3 *Data Limitations*

A first limitation concerns data availability and the disagreement between textual databases. Ridout *et al*. (2012) find preliminary evidence that there are stories (in their study mostly international coverage) from printed newspapers that are systematically missing in electronic databases. Thus, not only do texts need to be collected, but the content from multiple sources also needs to be aggregated neatly. Chiou and Tucker (2017) cover some of the likely issues of content aggregation. The problem of data availability and data disagreement is small for open government databases, such as accounting textual data (e.g., EDGAR), court decisions (e.g., PACER), or patents (e.g., Espacenet). Much in the same way Riffe *et al*. (2019) note that the universe of online posts is "unlimited and unknowable and inherently unstable over time," the problem becomes more persistent for data coming from corporate resources, news media, or social media. Any sample drawn from that data might not be representative due to nonrandom sampling; true probability sampling is hard in the context of big data. Lacy *et al*. (2015) mention convenience sampling (a sample primarily defined by availability) and purposive sampling (a sample primarily defined by the nature of a research undertaking) as common practices.

An important aspect in data collection is the notion of data vintages. In a real-time setting, a researcher uses the data available at a given time, called a vintage or a snapshot. Yet, many data are subject to revisions. For instance, most data used in macroeconomics are updated one or more times until final numbers are reached (Croushore and Stark, 2003). The compilation of the FRED-MD historical vintage database of macroeconomic indicators was a response to this problematic (McCracken and Ng, 2016). This same difficulty persists in textual data, particularly with online publication and social media, with the data frequently updated, revised, or even removed from the information outlets. Saltzis (2012) reveals in a sample of breaking news stories on six major U.K. online news sites that the stories were updated on average 5.7 times. As such, the traditional process of scraping websites for historical news may lead to a forward-looking bias since the retrieved news will typically be the latest version of the news articles and not the one at the time of first publication. This phenomenon is crucial to deal with in intraday studies.

The problems described above lead to issues of reproducibility and limitations to generalizability of results.

5. Preprocessing, Enrichment, and Selection of Qualitative Sentiment Data

Textual, audio, or visual data rarely arrive in a format that is ready for input into an algorithm. The data typically start off being very unstructured, and through a sequence of steps, structure is imposed to make the data ready for further analysis. We define restructuring as doing two things: preprocessing and enriching the data. Both the preprocessing and metadata enrichment ideally come before the actual data selection to have the most information available to do an optimal filtering.

5.1 *Restructuring Textual Data*

In this subsection, we describe the preprocessing and metadata generation concerning textual data. Bholat *et al*. (2015) provide a useful summary of many relevant text mining techniques for preprocessing and data enhancement.

5.1.1 *Preprocessing*

Raw textual data often come in a JSON or an XML file from which the actual text needs to be extracted first. This process is called parsing. Depending on the type of data available, this can be a relatively straightforward or tedious task. As part of this process, remaining garbage such as HTML tags, addresses, or other formatting is removed, or simply not selected through the parsing algorithm.

Furthermore, textual data are inherently (ultra)high-dimensional (Kelly *et al*., 2019). Gentzkow *et al*. (2019a) highlight that to structure a text with a length of w words, each of which is drawn from a vocabulary of q possible words, the unique representations of this text has dimension q^w. Moreover, all characters in the text are probably not equally informative in assessing the sentiment of a particular document. For example, stop words such as "the" are seldom indicative and are usually removed to reduce the noise and the dimensionality. Some type of further cleaning is often required to deal with issues such as spelling mistakes or (nonstandard) abbreviations (Nowak and Smith, 2017). Denny and Spirling (2018) outline several common preprocessing steps. The output is a corpus of cleaned texts.

Textual data come in various granularities: words, sentences, paragraphs, and whole articles. Sentiment is the output of a function applied to specific components extracted from texts, also called terms. The most common kind of components are n-grams, a sequence of n words. Breaking up text into n-grams is called tokenization. If $n = 1$, tokens are referred to as unigrams. A bag-of-words approach presumes that the relative order of unigrams is irrelevant, but words are not necessarily independent of each other. More generally, a bag-of-words can be denoted by bag-of-tokens, where tokens can be any sequence of words. Further cleaning is needed to drop, for instance, punctuation marks, or transform all terms into lowercase, stemmed, or lemmatized form.

Terms are summarized into a document-term matrix, with the rows as the documents, the columns as the terms, and the cells as the values that measure the (weighted) frequency of occurrence of the terms. A document-term matrix is usually of high dimension and consequently very sparse, meaning, with a lot of zero entries. In a document-term matrix, the sparsest features are typically removed.

5.1.2 *Metadata Enrichment*

A corpus as is consisting of only documents can be enriched by adding all sorts of metadata. Metadata either already exist or are objective, such as a time stamp, the author, the news outlet, the language, or the geography. A good case for having metadata is that textual information is expressed across many different venues, including newspaper articles, newswires, and social media, all with a possibly differing degree of information value. Heston and Sinha (2017) emphasize the importance of studying news types to understand how financial markets process information and when underreaction and overreaction in returns occur. Aggregation across a metadata marker gives information about the sentiment concerning that particular metadata – for example, the sentiment about a given economic topic.

The available qualitative metadata need to be quantified for further use in the analysis. This can be done using binary or relevance variables. In the first case, one enumerates all qualitative metadata across the corpus for a given article and assigns a value of 1 if the metadata are of importance to that article, and 0 if not. A relevance variable follows the same principle but assigns a continuous score based on the connectedness of the metadata to the article. Some metadata lend better to be modeled as a dummy variable (e.g., language or geography) and others as a relevance variable (e.g., predefined topics). If there are too many individual instances of the metadata, one can consider to group them. Other metadata can be generated using text mining models. The first type of metadata that can be generated are entities, using named entity recognition extraction techniques. The second type of useful metadata are topics and related keywords based on a supervised or an unsupervised topic model. The features can be valued as the probability score coming out of the topic model. Readability of a text (e.g., Loughran and McDonald, 2014) or the tense of a text are two other potentially useful metadata indicators.

5.2 *Restructuring Audio and Visual Data*

The underlying raw format of audio and visual data is less comprehensible and vaster than textual data. One second of a video is size-wise equivalent to at least hundreds of pages of text; the maxim "An image is worth a thousand words" is no exaggeration. We emphasize some important aspects about the restructuring of audio and visual data.

5.2.1 *Preprocessing*

For sentiment classification, visual and audio data are processed into emotional clues handy to discriminate between different sentiment categories. A major focus of sentiment extraction in visual data is on facial expressions. Secondary are other nonverbal expressions (e.g., hand gestures) and environmental factors such as what is happening in the background. There are seven basic emotion classes (danger, sadness, surprise, fear, disgust, joy, and contempt) that can be inferred using a facial expression coding system originally proposed by Ekman and Friesen (1976). One can then construct variables that express the distance between several of these positional facial characteristics.

Visual data can be boiled down to image data. A video in that respect is a collection of segments, and every segment is a collection of images. Audio data can be boiled down to textual data using speech-to-text technology complemented with specific audio features (such as pause duration).

5.2.2 *Metadata Enrichment*

The principles of metadata enhancement for textual, audio, and visual data are similar, but the content of the metadata is different. Qualitative metadata such as author or time of publication are the same. Examples of useful audio features are pitch, pause, laughter, overlaps, and voice intensity; examples of

visual features are color and motion. Videos have both visual and audio features. Wang *et al*. (2003) describe features and extraction techniques across four categories (spatial visual, motion, coding, and audio).

A downside of features retrieved from nontextual data is that many of them require a large-dimensional representation. For a video, a manner to construct a feature called "smiling" could be to take the number of seconds a person smiles in the video. The decision of whether a person smiles is a function of various facial characteristic points that should be mapped deterministically to the binary outcome "smiling" or "not smiling."

5.3 *Selection of the Relevant Data*

Following a general-to-specific approach, the original and vast corpus of documents needs to be trimmed to a subselection of relevant texts. If the selection procedure is too restrictive, important data may be omitted; however, if irrelevant data are included, it may drastically lower the signal-to-noise ratio. This can be considered as "querying" the corpus database to extract the right selection of texts. Querying can be based on a search of keywords in the database using, for example, a *regular expression*. It should be made clear beforehand which texts are necessary to include in the analysis, or the selection can be approached as an optimization problem itself. The latter strategy would require defining different sets of keywords and finding out which keywords give the best outcome in terms of an objective (e.g., forecast accuracy).

To model an outcome variable, it is not always needed to focus exclusively on (sentiment) measures directly related to that variable. On the contrary, Larsen *et al*. (2020) argue that many news topics are, in their case, of interest to form inflation expectations, and thus, it would be limiting to only target media mentioning terms related to inflation. Kelly *et al*. (2019) tackle the problem of selection simultaneously with modeling a set of observed covariates. They propose a method that only includes phrases of interest when useful, conditioning on the observed covariates, building on the model of Taddy (2015a).

6. Quantification of Sentiment

Any sentiment measure is a proxy for the actual prevailing sentiment and needs to be *estimated*. This can be done by human annotators or by a statistical function. A wide variety of techniques exist to infer the sentiment embedded in qualitative data, but measuring sentiment is inherently application- and data-specific. Therefore, it is neither possible nor recommended to consider sentiment computation in a single manner.

Sentiment is quantified for a given observational data unit – for instance, a text or a video. Quantification of sentiment is either on a discrete scale (classification into two or more classes, such as negative, positive, and neutral) or on a continuous scale. Based on decision rules, one can go from continuous to discrete output. Some methods produce a tuple of a positive and a negative sentiment (probability) score. Multiple sentiment scores from one computation method can be considered as separate methods and can turn out to be more informative.[4] Sentiment scores might benefit from a normalization for interpretation purposes and possible outlier elimination. Sentiment analysis can also take up a more fine-grained externalization, called aspect-based sentiment analysis (De Clercq *et al*., 2017). This type of sentiment analysis separately measures the sentiment for different aspects and entities mentioned in the data unit. This is a combined problem that requires the extraction of entities and their aspect terms, classifying the aspect terms, before doing the sentiment calculation for each of the extracted combinations. One could draw an analogy with "feature-based opinion summarization" (Hu and Liu, 2004), which is less specific.

6.1 *Textual Data*

Textual sentiment quantification uses tools from the broad field of natural language processing (NLP) to quantify the sentiment of a given text. It consists of many NLP-related subtasks, such as identifying entities and extracting relevant features. We briefly discuss the lexicon-based and machine learning approaches as the two main types of methods for sentiment computation. The data unit is usually a document, a paragraph, or a sentence. Some fields prefer one over the other. Sentiment can be detected more precisely at sentence level, but in political science, for instance, most often the analysis remains at document level since it requires less heavy NLP (Grimmer and Stewart, 2013). For a broader treatment of textual sentiment computation and associated subtasks, we refer to Liu (2015) and Ravi and Ravi (2015).

The classification accuracy of the various sentiment approaches varies. Typically, machine learning algorithms outperform lexicon-based methods out-of-sample at the expense of computational efficiency and model transparency. The difference in performance is a function of the type of texts and the domain specificity of the lexicon employed. Ribeiro *et al*. (2016) provide an extensive overview of the accuracy of both lexicon-based and machine learning–based sentence-level sentiment analysis. They compare 24 popular sentiment methods over 18 labeled data sets. Their experiments convey first of all a rather low average level of accuracy. More importantly, there are large differences in the accuracy across sentiment methods and across data sets. Their results also reveal no outstanding method at the sentence level. The conclusion is that a sentiment quantification method needs to be selected carefully depending on the purpose.

6.1.1 *Lexicon-Based Approaches*

A lexicon-based computation of sentiment is the most straightforward, efficient, and parsimonious method. Turney (2002) defines lexicon-based sentiment analysis as "calculating sentiment for a document from the sentiment of words or phrases in the document." Mechanically, this requires the use of a sentiment lexicon with sentiment information about important (combinations of) words, which is then matched with a text. A lexicon is thus a collection of pairs of words (or a sequence of words) and associated sentiment scores. In most cases, lexicons stick to unigrams, but for some applications, it is more effective to use n-grams. Picault and Renault (2017) construct a lexicon specific to European Central Bank communication and explicitly consider n-grams, such as the positive bigram "lower unemployment." The size of a lexicon ranges on average from in the hundreds to in the thousands. There is no preferred lexicon size; too large can mean inaccuracy due to noise, and too small might mean not enough coverage or a lack of important words. Comparing lexicons is not always easy, given the often varying sizes but also because there is no universal polarity grading system (Ravi and Ravi, 2015).

There is a distinction between general lexicons and domain-specific lexicons. Both the Henry lexicon (Henry, 2008) and the Loughran and McDonald lexicon (Loughran and McDonald, 2011) were developed as a response to the suboptimal applicability of generic lexicons to texts in the finance domain – for example, earnings press releases. The Lexicoder Sentiment Dictionary (Young and Soroka, 2012) is tailored to news content about politics. Lexicons are simple and the least black-box solution, and usable at any text level. However, lexicons can be brittle when facing domain shift and complex syntactic constructions (Täckström and McDonald, 2011). Very few lexicons are domain-portable, meaning applicable across several domains and text structures. It is difficult to achieve, if it is at all, and therefore hardly desirable.

Liu (2015) sees three broad ways of generating lexicons – namely, manually, dictionary based, and corpus based. An additional approach to building lexicons involves a combination of manual labor and a statistical methodology, which may arise from the machine learning literature. It is important to differentiate between machine learning algorithms for lexicon construction and those algorithms to measure sentiment but with no explicit intention to obtain a sentiment lexicon. We cover the latter

in the next subsection. Apart from the manual approach, all methods entail automatic processes to varying degrees.

The manual approach to building lexicons has annotators assigning a sentiment score to selected words. Notable fully hand-curated lexicons are the Stone *et al*. (1963) General Inquirer and the Bradley and Lang (1999) ANEW word lists. Crowdsourcing platforms such as Amazon Mechanical Turk have made the task of developing high-quality manual lexicons more accessible nowadays. To our knowledge, the NRC lexicon of Mohammad and Turney (2013) was the first to be built using crowdsourcing services.

A dictionary-based approach allows producing lexicons more cheaply while keeping a good level of accuracy. This method starts from a list of seed sentiment words with known polarity (often found using the manual approach) and then expands this list by using synonyms and antonyms coming from a large base dictionary. A suitable base dictionary is the WordNet database (Miller, 1995). This lexicon in conjunction with sentiment seed words was used to produce WordNet-Affect (Strapparava and Valitutti, 2004) and SentiWordNet Baccianella *et al*. (2010).

The corpus-based method adapts an existing lexicon using information from a domain-specific corpus. The researcher first needs to adjust the sentiment orientation of the words to the new domain. Second, it may use linguistic rules to include new words in the lexicon. In this regard, Hatzivassiloglou and McKeown (1997) introduce the notion of sentiment consistency. For instance, adjectives with a similar sentiment orientation are often used in groups. Kanayama and Nasukawa (2006) propose the idea of sentiment coherency; the same sentiment orientation tends to be expressed in consecutive sentences, while sentiment change is expressed by an adversarial expression (e.g., "but" or "however").

Statistical methodologies are the fastest and cheapest but the most prone to error. They typically start from a set of words from a previously built lexicon or a corpus, and then, a statistical methodology is used to find the sentiment orientation of those words. Jegadeesh and Wu (2013) use a regression framework to measure the sensitivity of words ("word power") to stock returns; this could then be used to form a finance-specific sentiment lexicon. Lexicons can also be derived from (Bayesian) regularized methods, such as the Ridge, the LASSO, or the elastic net regression (see, e.g., Pröllochs *et al*., 2015; Nowak and Smith, 2017). Pröllochs *et al*. (2015) argue that a shrinkage approach (Ridge regression) is superior over a variable selection approach (LASSO regression) because multicollinearity among the token predictors tends to be strong. In the corpus-based category, Engle *et al*. (2020) create a climate change vocabulary based on a collection of climate change of white papers and glossaries. Their final lexicon is composed of the unique stemmed unigrams and bigrams, weighted by their respective term frequency–inverse document frequency (tf-idf) scores. To create a daily climate change index, instead of term matching, they use the cosine similarity between the tf-idf scores of a given article and the scores in the lexicon.

Lexicons do not cope with the linguistic context around which the sentiment words appear. To this end, advanced lexicon-based methods integrate so-called valence shifters in the sentiment computation. Common types of valence shifters are amplifiers (e.g., very), downtoners (e.g., barely), negators (e.g., not), and adversative conjunctions (e.g., but). These valence shifters act on polarized words in the lexicon in particular ways depending on how close they appear to these polarized words. Taking the example of negation, one way to apply it to lexical entries is termed shift negation (Taboada *et al*., 2011) as opposed to switch negation.[5] Having lexicons consisting of *n*-grams would also allow disambiguating of word use in different contexts. According to Young and Soroka (2012), even a modest integration of contextual (preprocessing) routines is fruitful. Taboada (2016) enumerates multiple linguistic insights to account for in sentiment analysis.

Not only domain specificity, but also language specificity is important. Most resources are still in English (Ravi and Ravi, 2015). In practice, one often sticks to translation. Either one translates the focused text from a resource-poor language into a resource-rich language (usually English) for which a robust sentiment method (e.g., lexicon) is available, or one translates an existing word list into the focused language. A third option is to translate annotated corpus resources from a resource-rich language to the focused language and use these to develop (or improve) another sentiment method. In many circumstances,

however, the performance of translation results in a loss of accuracy. Mohammad *et al*. (2016) surprisingly find that, with Arabic social media as the focused texts, sentiment analysis of automatic English translations is competitive to existing Arabic sentiment analysis systems. On the other hand, translation made the human annotations become worse than sentiment analysis, and adding Arabic translations of sentiment-labeled English tweets data to Arabic training data resulted in a drop in accuracy, due to bad translations. Translation invariably comes with additional problems to solve. Bannier *et al*. (2019) start from the English Loughran and McDonald lexicon by doing word-by-word translation to German. On top of that, they deal with distinct grammatical features of the German language related to inflectional and lexical morphology, as well as compound wording. They claim to have described a comprehensive framework for future adaptations of dictionaries into other languages. To test the equivalence between their lexicon and the Loughran and McDonald one across positive, negative, and neutral categories, they rely on the two-sided equivalence test of Blair and Cole (2002). The test checks for accordance in terms of the mean number of detected polarity categories, given a confidence interval.

6.1.2 *Machine Learning Approaches*

The extraction of sentiment as a stand-alone problem is studied by machine learning and computational linguistics scientists. The purpose is to optimize the measurement of sentiment based on a learning algorithm typically benchmarked against an annotated data set of text with corresponding sentiment values. The objective, in this case, is well defined and dependent on the type of data source (e.g., product reviews or images) and the type of sentiment output (e.g., classification into positive or negative). The learning algorithm identifies the characteristics among the preprocessed smaller pieces of textual characteristics (i.e., words, *n*-grams, phrases, counts, and other information) that are most important in measuring sentiment. A survey of different machine learning algorithms applicable to text is given in Evans and Aceves (2016). Machine learning can be branched into supervised and unsupervised learning, both used on many occasions for sentiment analysis.

Supervised machine learning requires an annotated data set – meaning, a set of documents with, for every document, a sentiment value, leading to what is often called the gold standard. Annotation can already exist from the data (e.g., product rating stars), but, in most cases, is constructed manually. Building such a data set from scratch can be expensive and time-consuming while also prone to bias. Especially for domain corpora, annotation can be hard due to possibly complex specific sociolinguistic contexts (Hamilton *et al*., 2016). The annotation cost also depends on the type of text. Van de Kauter *et al*. (2015) review some of the complexities of doing annotation. Taddy (2013a) outlines a procedure to select from a large corpus the texts that are most useful to annotate. Determining the best data examples to be labeled is referred to as (pool-based) active learning. Once the tagged data set is obtained, a specific machine learning algorithm is trained with it. Pang *et al*. (2002) clarify the sentiment classification problem, and experiment with the Naive Bayes, maximum entropy classification, and support vector machines (SVMs) learning techniques. Naive Bayes and SVM are essentially text regressions of the sentiment target variable on a large-dimensional space of textual elements, such as words, which get assigned a weight. More recently, neural networks, primarily due to the emergence of deep learning, have become more prominent. One can also combine several learning algorithms. For instance, Das and Chen (2007) employ a majority voting scheme across five classifiers, claiming that it minimizes false positives.

An unsupervised learning approach lets the data decide the categories or representation by themselves. Any unsupervised method is typically hybrid or semisupervised, as there is need for specific minimal inputs from the modeler. A classic example is the suggested approach by Turney (2002), which ranks phrases based on their pointwise mutual information (PMI) with respect to two seed words, one negative (“poor”) and one positive (“excellent”). It infers the semantic orientation from the semantic association with respect to a manual set of seed words. Remus *et al*. (2010) develop the German SentiWortschatz

dictionary using the PMI approach. A vector space model (VSM) is a more complex undertaking. These models generate word embeddings, which are latent quantitative vector representations of textual information, such as documents, paragraphs, words, phrases, and even letters. A VSM learns distributed vector representations that capture many precise syntactic and semantic word relationships. Words closer to each other in terms of linguistic context receive a more similar quantitative representation because they are assumed to share the same semantic meaning.[6] Global matrix factorization methods (co-occurrence counts based) and local context window methods (prediction based) are the two main families for learning word vectors.

Latent semantic analysis (LSA) is a notable example of a global matrix factorization method. It reduces high-dimensional count vectors to a lower dimensional latent semantic vector space. Hofmann (2001) introduces a probabilistic version of LSA, defining the semantic space over a set of latent variables referred to as "aspects" based on a generative model for word-to-document co-occurrences. His model allows figuring out, for instance, which latent aspects are most likely to generate a word, or what the latent class posterior probabilities are given a certain document and word. Liu *et al*. (2009) refactor the model to capture a multidimensional measure of blog sentiment, considering sentiment as a joint contribution of a few hidden factors. They call their work sentiment probabilistic LSA (S-PLSA). In a subsequent time series regression, they form sentiment variables as the average sentiment mass attributed to each of the hidden sentiment factors.

Most of the recent research on word embeddings has gravitated toward the prediction-based method using neural network architectures. The Word2Vec approach of Mikolov *et al*. (2013) is one of the earliest and best-known techniques in this category. Word2Vec uses the continuous bag-of-words (CBOW) or the continuous skip-gram model architecture. In CBOW, one tries to predict the current word in a text from a window of surrounding context words. In contrast, in the skip-gram model, one tries to predict the surrounding context words using the current word. Mikolov *et al*. (2013) also formalized the idea of using vector operation, such as $vec(\text{“}Madrid\text{”}) - vec(\text{“}Spain\text{”}) + vec(\text{“}France\text{”}) \approx vec(\text{“}Paris\text{”})$. GloVe (Pennington *et al*., 2014) aims at taking the best of the count-based and prediction-based methods, with a first attempt to integrate both global and local statistics. Pennington *et al*. (2014) find that the quality of GloVe's learned representations is slightly better than Word2Vec's vectors, but it depends on the task at hand. A more recent method is fastText (Bojanowski *et al*., 2017). It incorporates subword information into the learning process such that words not observed in the training corpus (out-of-vocabulary) can still be assigned a word vector. The current state of the art in word embeddings is the deep neural network Bidirectional Encoder Representations from Transformers (BERT) models and its variants (Devlin *et al*., 2018). These models most explicitly integrate global and local context. For example, the word vector for "right" in "I am right" and "I take a right turn" will be different.

Estimated word embeddings are used as an input to more traditional sentiment classification methods (e.g., logistic regression), or to probabilistic methods such as the one proposed by Taddy (2015b). Alternatively, by selecting several known positive and negative seed words, the vector space can be used to pinpoint words adjacent to those seed words and consider them as carrying the same polarity. The SENTPROP method from Hamilton *et al*. (2016) first constructs a lexical graph from a VSM with the words connected according to their embedding using cosine similarity, and then, performs label propagation to define the polarity. The sentiment score of a word is proportional to the probability of a random walk hitting that word, as propagated starting from a seed set. To obtain confidence bands around the scores, they bootstrap over random subsets of seed words.

In the same vein as for lexicons, learning algorithms are ideally adapted for specific domains and languages to optimize the sentiment quantification. Thus, for optimal accuracy, the analysis for a specific domain needs a separate annotated data set, as opposed to using an annotated broad corpus and the resulting generic trained algorithm. Transfer learning is the strand that investigates the optimal conversion of methods in one domain or one language to another. Good transfer learning minimizes the burden on the researcher to acquire equally informative domain-specific annotated corpora for all domains of interest.

An application of transfer learning is to deduce sentence-level sentiment from document-level sentiment labels. Täckström and McDonald (2011) use hidden conditional random fields as a latent variable structure model to deduce the latent sentence-level sentiment.

6.2 *Audio and Visual Data*

Some of the tools discussed for textual sentiment computation are also of value for the extraction of sentiment from audio and visual data. A lexicon can be constructed with entries such as "light smile," "big smile," "eye contact," "crying," "shouting," "high pitch," or "low pitch," all with a certain calibrated polarity, and the number of seconds the action is held as a measure of polarity strength.

Domain specificity can be thought of as speaker specificity in the context of audio data. Speaker-dependent approaches give (much) better results than speaker-independent approaches (Poria *et al*., 2016). The number of possible speakers is almost always larger than the number of possible languages or domains, making it infeasible to develop a specific algorithm for every individual speaker. However, making algorithms for types of speakers (e.g., political speakers) makes sense and is achievable.

Rousseeuw *et al*. (2018) define a measure of directional outlyingness that is applied on image data to detect (sudden) changes in how a video frame appears relative to another frame. A transformed aggregation of the various outlyingness measures would make a good candidate as a proxy for sentiment.

7. Aggregation of Sentiment Variables

Most researchers are not interested in an entity's or a data unit's sentiment at one specific point in time but in the average value on several moments, or across many entities, methods, and data sources. Therefore, appropriate aggregation is required.

7.1 *Within-Unit*

An essential aspect of the sentiment quantification as discussed in Section 6 is within-unit aggregation. For textual data, this becomes within-document or intratextual aggregation. Within-document aggregation is the weighting of the document-level sentiment information (e.g., the sentiment of a word or of a sentence) into a score that represents sentiment for that document. For visual data, this becomes, for instance, within-video aggregation, which consists of the aggregation of sentiment of the different segments of the video into a whole video sentiment score.

A widely used weighting scheme, in preprocessing and for text aggregation, is the tf-idf statistic. This scheme weighs terms based on their frequency of occurrence ("tf"), but revalues upward the words appearing across few documents ("idf"), under the idea that less frequent terms can be of greater value to detect the specificity of a document. This weighting approach makes document specificity a function of term use rather than term meaning. Another option is to weight based on reader's attention, which could be assumed higher in the beginning and end of a text. Allee and DeAngelis (2015) find an important degree of dispersion of sentiment in financial disclosures. Documents have typically one dominant sentiment class but no uniform sentiment across paragraphs or sentences. Boudt and Thewissen (2019), for example, show a clearly U-shaped pattern of sentiment within CEO letters.

Poria *et al*. (2016) outline two approaches to aggregating, or *fusing*, textual, audio, and visual signals, which happens when dealing with video material. A first strategy is to combine characteristics from every type of data into a joint vector and use this vector as input in a classification algorithm. The second strategy is to model sentiment individually per data stream, and then combine the unimodal results based on suitable metrics and weighting. The dynamic weighting of the unimodal results is an interesting research issue to explore. Pham *et al*. (2018) propose a third strategy, closest related to the first strategy,

aiming at a joint multimodal representation. They use an unsupervised encoder-decoder framework but admit that a unimodal textual approach led to the best overall results in their empirical video analysis.

7.2 *Cross-Sectional*

Cross-sectional aggregation can occur at multiple levels. A first level is across documents at a given frequency, which results in a time series. This across-document aggregation is the natural next step after within-document aggregation. For example, to obtain a weekly time series, all sentiment scores need to be aggregated at a weekly frequency. An interesting possibility for the aggregation is to let the weights depend on the articles' reach (e.g., the number of reads). One can then decide to further adjust the weights based on some empirical knowledge – for example, to cope with the underrepresentation of far-right voters on social media, as suggested in Ceron *et al*. (2014).

A second level is across documents for a given metadata marker. For instance, one could aggregate sentiment scores for all documents coming from a given source, or discussing a certain entity. Only considering a limited number of sources to measure sentiment for a given period risks to give a biased estimate due to an unrepresentative sample. Typically, the first and the second levels are combined to obtain a time series for a given metadata occurrence. Many of such combinations capture different dynamics of the corpus and its metadata. Borovkova *et al*. (2017) obtain weekly sentiment values by a weighting that takes into account the relevance and novelty scores supplied by the Thomson Reuters News Analytics database.

A third possible level of cross-sectional aggregation is across sentiment methods. The order of when to do this aggregation depends on the goal. In the simplest scenario, only one method is used, or multiple methods are kept side by side – meaning no across-method aggregation at all. Another simple scenario is to average the sentiment scores from any given number of methods to obtain an averaged sentiment score. Boudt *et al*. (2018) take the centered average of the scores coming from the lexicons they apply. A more statistical approach is commonality extraction, using principal component analysis or latent factor modeling. Rogers *et al*. (2011) define sentiment as the first principal component over a range of sentiment measures. Lastly, an objective-based approach optimally weighs different methods based on their relationship with a target variable or based on another quantifiable objective. We further develop the techniques, problems, and open questions regarding the last two approaches in Section 8.

7.3 *Across-Time*

Across-time aggregation aims to smooth obtained sentiment time series or, more generally, to infuse a certain time dependency pattern. There are various valid reasons for smoothing. One of those is to remove outliers. This especially holds for short-term sentiment series, for example, at a daily frequency. Thorsrud (2018) applies a 60-day moving average to his daily tone-adjusted textual topic time series to filter out the noise. Another motivation for smoothing is related to the belief that sentiment at a certain time usually also partly reflects earlier sentiment. Sentiment needs to be updated when new information arrives but remains affected by previous information. Ardia *et al*. (2019b), for example, use beta weighting schemes covering a large number of possible time dynamics. They use a data-driven calibration to deal with the problem of not knowing in advance which time pattern has the most value for forecasting. The Kalman filter is also an appropriate technique to smooth out sentiment time series. It can be used to retrieve the unobserved sentiment state from the observed (already aggregated) sentiment variable. Borovkova *et al*. (2017) employ a simple local-level state-space model, leading to significantly less noisy sentiment variables. Shapiro *et al*. (2018) use a monthly fixed effect as time series sentiment indicator, controlling for newspaper and article-type fixed effects.

7.4 *Across Variables or Proxies*

The combination of the likely heterogeneity in the input data, the number of variables that can be associated with the data, and the number of sentiment implementations and aggregations may give rise to many constructed sentiment time series. For instance, Gelper and Croux (2010) use a one-factor model, estimated either as the first principal component or using partial least squares, to form an aggregate sentiment indicator from 160 sentiment proxies. In Ardia *et al*. (2019b), the different sentiment variables are weighted and assembled into a sentiment-based index using the elastic net regression. The obtained sentiment index is specific to the dependent variable used in the regression. Aggregation here is thus across metadata as well, which is usually not done at the across-document level. For example, to measure the sentiment around the economy, one may want to obtain this sentiment as a weighted average of several components such as employment, production, and the business cycle. Borovkova *et al*. (2017) obtain a final aggregated weekly sentiment index as an average of sentiment indices about important financial institutions, weighted by a bank-related measure such as net debt.

In a multivariate setting, one can repeat this process of creating separate sentiment indices for a series of proxies and then aggregate across these sentiment time series to obtain a final sentiment measure. That measure ought to be the optimized representation of the latent variable that is assumed to be represented by the collection of proxies. An example of a latent variable is the reputation of a company, which depends on observable variables such as profitability, market share, stock returns, and sustainability. Simplicity in weighting might be desired (e.g., equal weighting), but more complex (aggregation) schemes deserve to be studied. Larsen and Thorsrud (2019) use the marginal likelihoods across predictive regression models to form weights aggregating text-based time series into an index that best captures the variable to predict. Going forward, the idea of forecast combination could be useful for across-proxy aggregation.

Nimark and Pitschner (2019) define two interesting aggregated measures based on topic probabilities coming from a probabilistic topic model. The first is topic-specific deviation of a certain news topic ("specialization"); the second measures the news homogeneity in terms of agreement which topic is deemed most important. Empirically, they use the measures to show that different news sources emphasize different topics, but major events make news coverage more homogeneous. Similar measures could be constructed to test for the sentiment agreement across various sources.

Creating interactions of sentiment time series with other variables allows testing their interplay in explaining a dependent variable. The joint assessment of sentiment and topics is most prevalent in the literature (see, e.g., the sentiment-adjusted topic measures of Thorsrud, 2018, or the context-specific sentiment time series in Calomiris and Mamaysky, 2019). Calomiris and Mamaysky (2019) and Glasserman and Mamaysky (2019) use an entropy-based measure to characterize a collection of news during a given time frame in terms of "unusualness" and create simple interaction terms with sentiment variables aggregated at the same frequency. These interaction terms add information, allowing one, for example, to uncover that negative unusual news leads to an increase in U.S. stock market volatility (Glasserman and Mamaysky, 2019). Boudt *et al*. (2018) assess the interaction of sentiment with various company variables (finding that the informativeness of sentiment depends on the level of information asymmetry), while Arslan-Ayaydin *et al*. (2016) interact sentiment with managerial compensation (finding that the informativeness of sentiment depends on the incentives to manipulate the sentiment). García (2013) interacts a measure based on the *New York Times* news with a dummy variable to indicate a recession and concludes that daily stock returns are better predicted during recessions.

8. Modeling

This section is mainly approached as the problem of modeling an outcome variable Y as a function of the sentiment variables stored in a matrix $\mathbf{S}$, and possibly a number of control variables in another matrix $\mathbf{X}$. It can very generally be seen as modeling the joint density function $f(Y, \mathbf{S}, \mathbf{X})$.

8.1 *Time Series Models*

A very simple setup exists in modeling the output variable with a small number of sentiment variables and possibly other explanatory variables through a linear regression. Simple means it can be solved with ordinary least squares (OLS) regression. Penalized, or regularized, regression is required when OLS regression cannot be applied – that is, when the number of explanatory variables is too high relative to the sample size, or when there is a severe problem of multicollinearity. Regularization of a high-dimensional variables set shrinks the coefficients of the least informative variables toward zero. The Ridge (Hoerl and Kennard, 1970) and the LASSO (Tibshirani, 1996) approaches are the most common ways to specify the penalized regression. The elastic net regularization of Zou and Hastie (2005) embeds both the Ridge and the LASSO.

Factor models extract one or more latent common patterns among a set of time series. Thorsrud (2018) develops a mixed-frequency time-varying dynamic factor model from which he extracts a daily-news-based coincident index of business cycles. Both the mixed-frequency and (dynamic) factor aspects are useful approaches. For the first, for example, sentiment aggregated at both weekly and quarterly frequency could be fed through a mixed-frequency factor model to obtain a short term, a long term, and an overall trend. Similarly, grouped data settings can be used to extract common sentiment in groups of time series – for example, a common factor for every industry group consisting of all firms' sentiment measures. Andreou *et al.* (2019) derive asymptotics to identify common and group-specific factors in such a setting. Specifically, they introduce a test to assess which factors are common across a set of group-specific vectors.

The news-based measure from Manela and Moreira (2017) is an estimate from an SVM regression using the VIX index as dependent variable and normalized n-gram counts from texts as independent variables. This is a valid way to create a final optimized index – that is, to let an index be constructed from how well it captures a target variable. However, using such sentiment proxies in a second-stage regression usually has an impact on the uncertainty surrounding the then estimated coefficients. Manela and Moreira (2017) adjust the standard errors around the eventual point estimates to account for the uncertainty that is introduced by the first-stage regression.

Many target variables of interest could be discrete – for instance, an indicator variable whether a month lies in a recession period or not. Regularization is also perfectly applicable in a nonlinear context. Pure machine learning algorithms, such as SVM, neural networks, or Random Forest, are more relevant in a nonlinear setup, also applicable in case of time series variables.

Multiple sentiment variables and target variables can be jointly modeled in a multivariate regression framework, such as VAR models (see Qin, 2011, for a historical development of VAR models, and Lütkepohl, 2017, for a survey on structural VAR models). These frameworks are in general less prone to identification issues, since the variables are treated as endogenous, unless when explicitly considered exogenous or not modeled.

8.2 *Generative Models*

One can distinguish between two key econometric approaches to measuring sentiment (Gentzkow *et al.*, 2019a). Sentiment is either seen as a function of the written text (sentiment $= f(\text{text})$), or the written text is seen as a function of the underlying sentiment (text $= f(\text{sentiment})$). In the latter case, sentiment can be considered as a parameter of a stochastic process that generates texts as realizations. A seminal research paper in this is field is by Blei *et al.* (2003) proposing the latent Dirichlet allocation (LDA) model. Under this model, documents are assumed to be random mixtures over a predefined number of latent topics, where each topic is characterized by a distribution over words. Fitting such a model on a corpus of texts allows studying topic prevalence (the proportion of a document devoted to a topic) and topic content (the words used to discuss a topic).

Blei and Lafferty (2006) come up with a dynamic topic model that allows the content of the topics to change over time. Blei and Lafferty (2007) extend the LDA model by making correlation across topic proportions possible. Roberts *et al*. (2016) develop a structural topic model that lets the discovery of topics be a function of both word counts and observable covariates. These covariates can consist of sentiment variables, or metadata such as author and time of publication. The generative paradigm in a sentiment context thus starts from a statistical model that should be viewed as the source of all statements generated. For example, a model can be set up in which tokens are hypothesized to follow a generative model conditioned on a sentiment variable.

Taddy (2013b) introduces a framework to obtain low-dimensional document representations rich in sentiment information, called multinomial inverse regression (MNIR). He defines sentiment as the observable variables (e.g., product rating or whether a text is positive or not) impacting the composition of text data. Hence, his approach clearly follows the "text $= f$(sentiment)" assumption. The most probable sentiment output can be associated with any unseen text using forward regression. Taddy (2015a) extends the MNIR framework to also account for potentially larger dimensions of the sentiment variables, referred to as distributed multinomial regression (DMR).

8.3 *Combining Time Series Models and Joint Generative Models*

Given the natural role that topics play as metadata features, the joint generative modeling of topics and sentiment is very useful, especially when a time series perspective is included. The dynamic topic model framework of Blei and Lafferty (2006) can be deemed a time series generalization of the topic models proposed earlier. Eguchi and Lavrenko (2006) address both the topic and sentiment of a text unit using probabilistic generative modeling. Every statement is considered to have a set of topic-bearing and a set of sentiment-bearing words, each coming from respectively an underlying topic and sentiment language model. The dependence between both models is explicitly taken into account, under the assumption that sentiment depends on the topic. This assumption is, for example, supported by the importance of domain-specific sentiment lexicons.

Lin and He (2009) jointly extract document-level sentiment and the mixture of topics using an unsupervised procedure. They go from the three-layered LDA (topics associated with documents, and words associated with topics) to their joint sentiment/topic (JST) model, having four layers (sentiment labels associated to documents, topics associated with sentiment labels, and words associated with sentiment labels and topics). The joint sentiment and topic modeling answers to the need for domain specificity of sentiment analysis. It generally is approached as a two-stage process: first the detection of topics, and then the assignment of sentiment labels.

He *et al*. (2013) and Fu *et al*. (2015) further develop two related joint sentiment-topic models that allow selected dynamic parameters to account for the time variation in topics and sentiment. The inclusion of external variables makes it easier to interpret the driving processes behind discourse and content of qualitative material. In the approach of Gentzkow *et al*. (2019b) to measure trends in the degree of polarization in political speech, one can, for instance, include observed and unobserved speaker-specific characteristics.

There does not seem to be any longitudinal approach that uses the current state of a set of external variables (e.g., representing the economic and financial markets) as drivers for the time variation of the used sentiment and topics in written media articles. Such a holistic parametric model has, however, clear advantages in terms of econometric inference about the relationship between the observed news coverage, the features of the news sources, and the time variation in the variables system.

8.4 *Normal and Abnormal Sentiment*

There are several modeling approaches to decomposing sentiment into a normal and an abnormal component. Huang *et al*. (2014) distinguish between normal tone and abnormal tone, defining abnormal tone as the residual of a regression of tone on firm-specific characteristics. Ardia *et al*. (2019a) make the same distinction. They similarly consider a regression approach but use a static observable factor model, more precisely a market-cap-weighted sentiment index, with abnormal tone also the residual. Other alternatives could be to use the residuals of a simple mean model or of a latent factor model.

Hubert and Labondance (2018) identify sentiment as the unpredictable component of lexicon-based textual tone, orthogonal to a series of variables representing economic fundamentals. In other words, they define sentiment as the soft information conveyed through the tone of a communication beyond traditional quantitative and qualitative information conveyed through the content. Sentiment is obtained as the residual, with its first-order autoregressive component removed, from a regression on the variables representing the fundamental content.

8.5 *Attribution Analysis for Model Interpretation*

Interpretation is strongly tied to the problem definition and generally qualitative. On the statistical side, we point out attribution analysis to interpret measured, nowcasted, and forecasted sentiment.

Sentiment aggregation and modeling condenses a lot of information into a few quantitative sentiment representations of interest. A natural question is then how much of the final value is explained by the input data. Obtaining such a decomposition of the final value into the contributions of the component input data is the purpose of a top-down attribution analysis. These constituents are weighted based on their relationship with a target variable, and thus, allows studying the relative importance of every constituent or of groups of constituents. This, in fact, is a more fine-grained approach to doing sentiment decomposition, though typically not model-based. Aggregation based on the metadata features allows obtaining a predefined decomposition of the relevant sentiment and may help with identifying the underlying sentiment drivers in relation to a target variable. Because of the linearity of the aggregation performed in Ardia *et al*. (2019b), the attribution to any of the aggregation dimensions could be easily obtained. For example, they attribute the full sentiment-based forecast of the U.S. industrial production growth to six clusters of separate economic topics. The aggregate news index from Thorsrud (2018) can also be decomposed in terms of topic contribution. An interesting avenue to explore is to do the same attribution to various news sources and bring this into relation to how readers are exposed to these sources and their potential media biases. Larsen *et al*. (2020) analyze the variation in attribution by looking at the proportion of attribution that is unchanged for model updates up to 60 months in the future. During the global financial crisis, the predictive attribution relationship turned out to be much less stable, with only a small proportion of the explanatory news variables remaining important. This speaks in favor of doing regular model reestimations when times are troubling to incorporate the relevant news. Calomiris and Mamaysky (2019) also detect strongly time-varying coefficient estimates for news measures when forecasting the stock market. This is due to both the changing mix of the news sources as well as the actual impact of the news. Interestingly, Larsen and Thorsrud (2018) find that narratives mostly go viral during downs in the business cycle, albeit for a duration of only a few months.

In case of multivariate economic systems, impulse response functions in the VAR framework are usually used for interpretation. An impulse response function describes a variable's evolution along a specified time horizon after a shock in the regression system. When a meaningful sentiment shock is infused, its impact on all other variables can be quantified and understood across time. Borovkova *et al*. (2017) analyze the impact of a one-standard-deviation change in sentiment on various macroeconomic variables and find it to last significantly up to two months later.

9. Validation

The entire workflow is about extracting sentiment variables from qualitative data and using those variables in an economic analysis. Validation takes place at the end of every step but can be broken down into four categories: (1) evaluation of the quality and selection of the data, (2) evaluation of the sentiment quantification and aggregation, (3) model estimation and hypothesis testing, and (4) evaluation of the out-of-sample statistical and economic performance of the model-based predictions.

Many choices in the econometric analysis of textual, audio, and visual sentiment remain *ad hoc*. To adequately gauge the presence and impact of sentiment, the entire analysis should be frequently validated in a problem-specific way, both quantitatively and qualitatively. Comprehensive validation combines tools from econometrics with tools from machine learning. Machine learning is mostly about accuracy of prediction; econometrics is about uncovering (causal) relationships between economic variables.[7] Validation essentially jointly tests the current step and all previous steps as to whether they satisfy the assumptions for correct further (econometric) analysis.

When a sentiment variable does not seem to have a significant effect on the variable of interest, it may be due to two things. Either there is no significant effect of sentiment, or there is a significant effect, but the sentiment variables used are a weak proxy for real sentiment and do not capture the significant relationship. This can be conceived as a "joint hypothesis" problem. To mitigate this problem, the validation in the field of sentometrics is largely twofold. First, one should validate the sentiment variables created and then the model. When a model is deemed adequate in a statistical sense, further validation includes the interpretation of the results. A sentiment-based model that cannot be interpreted is not useful to convincingly answer the question outlined.

9.1 *Data Quality and Data Selection*

Since textual, audio, and visual data arrive in raw formats, the quality can vary substantially. Chances that are not all data units are fully cleaned even after preprocessing. Data quality checking is an iterative process. It is natural to go back to the cleaning and selection when some errors are found a few steps further in the workflow.

A basic quality check asks whether everything necessary for analysis is present. For instance, to be able to do a time series analysis, time stamps are inevitable. Any preprocessing of data involves a clear trade-off between simplifying the data and information loss. Denny and Spirling (2018) document the sensitivity of textual preprocessing choices on the outcome of an unsupervised analysis. They devise a scoring and regression approach to quantify this sensitivity.

Validation of the data quality and its selection exists in minimizing the exposure to the limitations described in Section 4.3 or acknowledging them going forward. Ideally, the selected data are maximally spread out across relevant data sources. If there are several major broadcasters but data for only one are available, there is a severe risk of bias when generalizing any obtained results from this restricted data set, as opposed to being only interested in and sticking with the conclusions of the particular data source studied.

Directing the analysis of audio data via speech-to-text to a textual analysis brings up the question of how trustworthy the conversion was. It is important to treat every transformation step and its possible errors as such, not confusing the textual data for the actual source audio data.

The data should be controlled for duplicates or near duplicates. If the duplicated data entries come from a different source, the content has likely been consumed more widely. A way to omit duplication but still maintain the implications it has is to add a metadata component that counts the number of duplicated occurrences. Wang *et al*. (2014) provide a (technical) overview with different techniques useful for duplicate detection.

9.2 *Sentiment Quantification and Aggregation*

The quantification of sentiment is highly important because it provides the numbers that any further step and interpretation is based on.

Relying on machine learning to train sentiment classifiers works under the assumption that the annotated data set is a faithful representation of the actual sentiment. Not every annotation procedure leads to a reliable annotation set. The quality of the gold standard can be measured by the level of interannotator agreement using, for instance, Cohen's kappa. To measure the effectiveness of a sentiment classifier or a lexicon, one has to compare the model-generated scores with the gold standard. More precisely, the trade-off between precision (the proportion of positives that is correct) and recall (the proportion of positives that is found) is at stake.[8] Recall and precision extend easily from a two-class problem (e.g., positive sentiment versus negative sentiment) to a multiclass setting doing micro or macro averaging (see, e.g., Zhang and Zhou, 2014).

Every lexicon tends to undergo one or more rounds of expert-based checks, to explicitly classify words into positive or negative, delete irrelevant words, and correct obvious mistakes. The validity of individual entries of lexicons are thus still mainly evaluated by humans. Overall, lexicons should undergo the same level of scrutiny as any other sentiment computation method in terms of validation. It should be tested if the accuracy of domain-specific lexicons is higher than generic lexicons. Loughran and McDonald (2011) use careful inspection of frequently occurring words as the only basis to create their alternative word lists. To validate this procedure, they relate tone computed from their negative lexicon to filing period excess stock returns, finding this sentiment measure to be in general more significant than tone based on the generic Harvard dictionary negative lexicon. The approach of Labille *et al*. (2017) compares a set domain lexicons on other domain texts. If the domain-specific lexicon is well constructed, it should rank first in terms of accuracy for the domain it is designed for. Apart from accuracy levels, another simple comparison procedure is an ANOVA analysis to see which lexicon's score variability is best captured by human coders. When the lexicon is generated with a regression, one looks at fit or information criteria statistics to validate the overall power of a lexicon (e.g., Pröllochs *et al*., 2015). An imbalance between positive and negative entries might make sense from a domain-specific perspective but should be defendable, since bias in the sentiment quantification algorithm can also be due to a biased training set. The Loughran and McDonald dictionary, for instance, is left with the large proportion of 78% negative words as a result of the domain adaptation to financial disclosures.

When creating sentiment measures, a first and simple analysis is to determine the correlation with existing related sentiment time series. All EPU indices of Baker *et al*. (2016) were validated using a very diligent human audit process, showing that the computer-generated indices are highly correlated with the human-generated ones. Soo's (2018) media sentiment housing index correlates strongly with the University of Michigan Survey of Consumers, albeit lagging. He further validates his index by confirming a reasonably strong lagged correlation with a multifactor index that combines multiple proxies, constructed based on the methodology of Baker and Wurgler (2006). The most difficult task can end up to find related proxies, as sometimes, 0 they are rare or do not exist at all. Another simple time series validation procedure is what is referred to as event validation. This entails visualizing a sentiment measure and confirming whether sharp increases or drops coincide with the incidence of important events that intuitively would result in a strong increase or decrease in sentiment, respectively.

9.3 *Econometric Modeling and Interpretation*

Many models are evaluated by measuring the accuracy in an out-of-sample prediction exercise. However, prediction is not always of interest; measuring which words and how they convey sentiment can be a more important objective that is not always related to prediction accuracy. As mentioned in Justin Grimmer's comment on Taddy (2013b), how to do trustworthy task-specific sentiment evaluation still needs to be

formalized. How is one to know with a high degree of confidence whether a token can be attributed to a particular sentiment feature? This problem can be particularly apparent when doing a large-dimensional regression of a sentiment variable on unigrams, for instance, with the resulting coefficients of the unigrams not always easy to interpret and sensitive to change across different specifications.

The problem of extensive and problem-specific validation is brought up in detail in Grimmer and Stewart (2013). For supervised methods, validation is fairly straightforward; it boils down to minimizing the prediction error in replicating a set of annotated outputs or maximizing the classification accuracy (typically making use of confusion matrices). A data set is best divided into a training, a validation, and a testing set to avoid a biased view on accuracy due to overfitting (Varian, 2014). Alternatively, one can do k-fold cross-validation or rolling forecasting origin cross-validation when dealing with time series. Unsupervised methods require combining experimental, substantive, and statistical evidence to show the conceptual validity of a model output. Proper validation of unsupervised models is especially important when used for inference or measurement rather than prediction or exploration (Roberts *et al*., 2016).

9.3.1 *Model Estimation and Hypothesis Testing*

It is common to evaluate the in-sample goodness of fit of a sentiment-based regression model with the (adjusted) R^2 statistic. When adding their word flow measures, Calomiris and Mamaysky (2019) find a substantial increase in the R^2 for predicting returns, volatility, and drawdown risk. The main concerned parameters are those associated with the sentiment variables. Their significance should be assessed statistically and economically. Statistical significance shows whether an effect exists, but its applicability is mainly limited to low-dimensional models. Gandomi and Haider (2015) review various issues of doing econometrics in a big data environment, pointing out the "irrelevance of statistical significance." Economic significance inspects the sign and strength of the association. Economic meaning can be given through, for instance, an attribution analysis.

In general, textual, audio, and visual data bring the known endogeneity challenges to econometricians. The creation and publication of texts, videos, and speeches is correlated with many factors, so positing a cause and effect remains dangerous when no further insights into the (many) underlying factors of the data are available. Is it the sentiment of the alternative data set that is at the heart of a certain correlation or causality, or is the sentiment a reflection of associated underlying factors? Does the sentiment impact the outcome variable directly or indirectly, and through what mechanism? Larsen and Thorsrud (2018) partition their network of sentiment/topic variables into more and less exogenous variables. Variables are considered exogenous if they have predictive power for other topics but are not (often) predicted themselves. The most exogenous variables seem to be associated with economic fundamentals. Hubert and Labondance (2018) correct for endogeneity in their central bank tone measure by stripping away fundamentals, expectations of future fundamentals, standard monetary shocks, investor sentiment, and past sentiment shocks. Benhabib and Spiegel (2019) deal with endogeneity and, more specifically, reverse causality using instrumental variables. They use political data to instrument for differences in (survey-based) sentiment levels by state. When testing the effect of sentiment on the target variable and finding significant results, it is also recommended to test the effect of the target variable on sentiment via a reverse (lagged) regression specification. Few research papers on sentiment have carried forward this robustness step.

Model uncertainty is assessed through analyzing the impact of sentiment parameter estimates across various model specifications. This has to do with both a good and exhaustive definition of the control variables $\mathbf{X}$ and with testing for enough different model structures. Soo (2018) creates robustness variables from the qualitative data themselves. He computes indices from those news articles that convey fundamental market information rather than sentiment, adds those to his regression specifications, and finds that his major sentiment index remains significant. Varian (2014) states that it is important "to be

explicit about examining how parameter estimates vary with respect to choices of control variables and instruments." Validation is rarely a black-and-white matter. The researcher should identify when and how sentiment is informative and when it is not.

9.3.2 *Out-of-Sample Evaluation*

An out-of-sample version of the R^2 statistic can be used to measure the relative reduction or increase in the mean square out-of-sample prediction error of a sentiment-based forecasting strategy with respect to a baseline strategy. Using the out-of-sample R^2, Adämmer and Schüssler (2019) document statistically significant increased predictive power of the monthly U.S. equity premium when using news-aggregated variables combined with a model-switching strategy. Caporin and Poli (2017) use five metrics (mean absolute error, mean square error, heteroskedasticity adjusted mean square error, QLIKE loss function, and R^2 of Mincer–Zarnowitz forecasting regressions) to compare the forecasting performance of a news-based realized volatility model versus a baseline.

The magnitude of the impact of sentiment variables on economic and financial variables is highly subject to time variation. Stability needs to be tested by performing the analysis, and measuring the performance, on various subsamples, or by doing rolling forward regressions.

A simple way to sidestep the issue of endogeneity is by comparing existing linear models to models enhanced with the quantified alternative data sources and to simply focus on whether predictive power improves or existing (significant) relationships hold. This could be formulated as testing models "controlling for sentiment." The model confidence set procedure from Hansen *et al*. (2011) allows testing whether different model specifications are truly different according to some significance level.

10. Software

This section points to a selection of useful software tools to carry out a detailed sentometrics analysis from textual data.[9] The selection is by no means exhaustive – meaning, there exist plenty of other software tools equally of use to perform (parts of) a sentometrics analysis. We limit ourselves to the open-source R and Python programming environments, due to their large popularity, strong communities of developers, and relatively gradual learning curves. For instance, MATLAB rarely comes to mind for doing textual analysis, but its Text Analytics Toolbox has many capabilities for doing powerful preprocessing, vectorization, sentiment analysis, and topic modeling. One does not necessarily have to choose one programming environment. Like Glasserman and Mamaysky (2019), a common workflow includes doing the handling of textual, audio, and visual data in Python, and the statistical analysis in R. The available software is linked to specific tasks involved in an econometric analysis of qualitative sentiment data and is summarized in Table 1.

The **quanteda** package (Benoit *et al*., 2018) is a general text mining toolkit in R. Its development has been actively supported by the European Commission. The package **tidytext** (Silge and Robinson, 2016) can also be used to do many text processing tasks, following "tidy" data principles. The **tm** package (Feinerer *et al*., 2008) is an older textual analysis framework, but is still used as a backend in many text-related R packages.[10]

The **NLTK** library (Bird *et al*., 2009), short for Natural Language Toolkit, is the text mining toolkit counterpart in Python, albeit even more exhaustive.[11] In Python, the **spaCy** library (Honnibal and Montani, 2017) is the most complete alternative. It is faster, but more black box. The **TextBlob** library (Loria, 2019) is built on the **NLTK** library and is therefore more specialized as to what concerns several textual extractions, such as sentiment analysis through machine learning classification.

The R package **sentometrics** (Ardia *et al*., 2020) provides a collection of functions to do sentiment computation, sentiment aggregation, and (high-dimensional) sentiment-based regression. Wischnewsky

Table 1. Nonexhaustive Overview of Textual Data Analysis Tools in R and Python.

Tasks	Restructuring			Sentiment quantification		Time series		Econometric analysis	
Software	Cleaning	Metadata	Tokens	Lexicon-Based	ML	Aggregation	Visualization	Regression	Validation
R									
caret					√			√	√
glmnet					√			√	√
quanteda	√	√	√	√	√				
rJST		√			√				
SentimentAnalysis				√	√				
sentometrics		√		√		√	√	√	√
textir		√						√	√
tidytext	√	√	√				√		
tm	√	√	√	√					
Python									
NLTK	√	√	√	√	√			√	
scikit-learn	√	√	√		√			√	√
spaCy	√	√	√						
TensorFlow					√			√	√
TextBlob	√	√	√		√				

Note: The abbreviation ML stands for machine learning. A tick indicates that the software can be directly or indirectly (i.e., by minimally chaining with other available tools) used to perform a particular workflow step. The packages included for R are **caret** (Kuhn, 2018), **glmnet** (Friedman *et al.*, 2010), **quanteda** (Benoit *et al.*, 2018), **rJST** (Boiten, 2019), **SentimentAnalysis** (Feuerriegel and Pröllochs, 2019), **sentometrics** (Ardia *et al.*, 2020), **textir** (Taddy, 2018), **tidytext** (Silge and Robinson, 2016), and **tm** (Feinerer *et al.*, 2008). For **Python**, the libraries tabulated are **NLTK** (Bird *et al.*, 2009), **scikit-learn** (Pedregosa *et al.*, 2011), **spaCy** (Honnibal and Montani, 2017), **TensorFlow** (Abadi *et al.*, 2016), and **TextBlob** (Loria, 2019).

Journal of Economic Surveys (2020) Vol. 34, No. 3, pp. 512–547

et al. (2019) use the package to create a "Sentoindex" that represents financial stability sentiment as expressed during testimonies at U.S. Congressional hearings. The sentiment computation in **sentometrics** is lexicon-based, but other sentiment scores can be used as input for further aggregation. The **sentometrics** package also provides a simple keywords-based approach to generating metadata features. The **SentimentAnalysis** package (Feuerriegel and Pröllochs, 2019) can be used to create lexicons and compute sentiment according to the method of Pröllochs *et al*. (2015).

The regression framework in **sentometrics** relies on both the **caret** (Kuhn, 2018) and **glmnet** (Friedman *et al*., 2010) packages but is specific to the sentiment time series generated within the package. The **glmnet** package implements various penalized regressions; the **caret** package provides more generic classification and regression modeling. The inverse text regression methods developed by Taddy (2013b) and Taddy (2015a) are available in the R package **textir** (Taddy, 2018).[12] The **rJST** package (Boiten, 2019) implements the joint sentiment/topic model of Lin and He (2009).

Python's **scikit-learn** (Pedregosa *et al*., 2011) is one if its most established machine learning libraries. It supports the majority of the common learning algorithms used in sentiment analysis and is easy to use with respect to feature engineering. To do the same, but also particularly deep learning, Google's **TensorFlow** library (Abadi *et al*., 2016) is the standard, albeit imposing more on the user in terms of setting up the individual components of a chosen model. Since recently, there also exists a comprehensive R interface to the **TensorFlow** framework.

These days, research papers also go increasingly accompanied with standalone open-source replication code (see, e.g., the MATLAB code used in Thorsrud, 2018).[13] Another example, to do sentiment analysis benchmarking, is the online tool iFeel 2.0 (Araújo *et al*., 2016) based on Ribeiro *et al*. (2016).

A shortcoming of the current software landscape is that there are no libraries that propose a full and easy integration of the required data handling, machine learning, and econometric tools. The preprocessing and sentiment quantification packages have very little in common with the packages used for modeling. Having to combine too many packages or even multiple programming languages is prone to error, for instance, due to the usage of different types of object classes that need to be converted.

11. Concluding Remarks

Sentiment analysis allows us to accurately and automatically map alternative data into quantitative statistics as a support for decision-making across many business applications. Economists and investors, and also politicians and journalists, have started to embrace the utilization of econometric methods in the analysis and application of textual, audio, and visual data, to understand historical evolutions and better forecast future evolutions.

We overview the emerging field of sentometrics that investigates the transformation of qualitative data into quantitative sentiment variables, and their subsequent application in an econometric analysis of the relationships between sentiment and other variables. This survey is organized around the different steps of a typicalanalysis. The most important terminology is collected in the appendix.

Textual, audio, and visual data will continue to become more cheaply and widely available, together with becoming more easily accessible. The interest of public and private institutions to monetize these data and their proprietary data will grow as well. We recommend further research on multimodal sentiment analysis in econometrics. The future will be exceedingly multimedia in terms of content generated, hence the analysis indispensably multimodal. A major challenge is the development of appropriate technology for unified multimodal sentiment analysis systems.

Progress toward better integrated and more reproducible sentiment data research will require collaborative cross-disciplinary efforts. We end this paper with a call for more efforts toward reproducibility in the econometric study of sentiment from qualitative data. It would benefit greatly from reference data and associated state-of-the-art performance, for different sentiment quantification

techniques, data, and econometric approaches. In the field of computer science, such practices are more widespread. Other researchers can evaluate any new approach on the reference data and as such provide a consistent picture of reproducibility or improved performance. Even though the sharing of code and data has gained adoption, there are yet no standard practices on how to do so. The reference data and results should be made available through an open database with easy access and well-documented formats. This matches with the proposition of Lacy *et al*. (2015) to set up a standard scholarly repository to share research-related materials. As a companion to this survey paper, we have therefore set up a collaborative econometrics and sentiment GitHub project to gather such resources.[14]

Acknowledgments

We thank the Associate Editors (Les Oxley and Stelios Bekiros) and two anonymous Referees, seminar participants at Ca' Foscari University of Venice, the European Commission JRC Ispra "Big Data and Forecasting Workshop" (Ispra, 2019), Ghent University, HEC Montral, the International Conference on Computational and Financial Econometrics (London, 2019), Skema Business School, University of Delaware, and Vrije Universiteit Brussel for their useful comments. We are also grateful to Francesco Audrino, Leopoldo Catania, Maxime De Bruyn, William Doehler, Nitish Sinha, and Leif Anders Thorsrud for stimulating discussions and feedback. This project benefited from financial support from Innoviris (https://innoviris.brussels), IVADO (https://ivado.ca), swissuniversities (https://www.swissuniversities.ch), and the Swiss National Science Foundation (http://www.snf.ch, grants #179281 and #191730).

Notes

1. See https://www.gdeltproject.org.
2. The EPU index for various countries can be retrieved from: http://www.policyuncertainty.com. The online publication of text-based indices is becoming prevalent, see also: https://www.retriever-info.com/fni.
3. A collection of open-source textual, audio, and visual data can be found at https://pathmind.com/wiki/open-datasets.
4. For instance, a net textual sentiment value of two can be obtained from both two positive and zero negative words, or 20 positive and 18 negative words. The number of polarized words can be retained as separate sentiment scores, else its information can be used for within-unit aggregation.
5. The importance and application of valence shifters is also a function of the document type. Hutto and Gilbert (2014) created the VADER sentiment analysis system for social media texts, letting word shape (e.g., capitalization), slang (e.g., "kinda"), and emoticons, among others, act as valence shifters.
6. Word embeddings are an advanced way of doing text vectorization, compared to, for instance, the simpler construction of a document-term matrix.
7. Advancements in machine learning and econometrics have been going more hand in hand. An interesting example is "double" or "orthogonal" machine learning, a development that aims to deal with the invalidity of inference infused by many machine learning methods (see mainly Chernozhukov *et al*., 2017 and related work).
8. The precision and recall metrics can be combined in the F_β-score, with $F_\beta \equiv (1+\beta^2)\frac{\text{precision}\times\text{recall}}{\beta^2\times\text{precision}+\text{recall}}$. The β factor defines the relative level of importance put on recall. If $\beta = 1$, both metrics are weighted equally in a harmonic mean sense.
9. A well-known open-source software tool for audio data processing is openSMILE (Eyben *et al*., 2013). The LVA software (https://lva650.com) can be used for preprocessing, deconstruction, and immediate emotion analysis of audio data (see Mayew and Venkatachalam, 2012, for an application in finance). For visual data processing, alternatives are the commercial softwares OKAO Vision System from OMRON (https://plus-sensing.omron.com/technology) or Luxand FaceSDK

(https://www.luxand.com/facesdk), both mainly for facial features extraction. A good commercial speech-to-text technology is Vocapia (https://www.vocapia.com). In the open-source sphere, the DeepSpeech project (Hannun *et al*., 2014) and associated software packages are very useful. Generally, the outputs returned by the above tools can be easily loaded into any programming environment to perform the remaining steps in the analysis.

10. A helpful starting point to explore the plethora of textual analysis tools in R is CRAN's Task View "NaturalLanguageProcessing" (https://CRAN.R-project.org/view=NaturalLanguageProcessing).
11. The **quanteda** package website gives an overview of the actual functions across the packages referred to perform several specific tasks (see https://quanteda.io/articles/pkgdown/comparison.html).
12. The DMR from Taddy (2015a) is also implemented with the programming language Julia, available at https://github.com/AsafManela/HurdleDMR.jl, which mainly includes the Hurdle Distributed Multiple Regression algorithm from Kelly *et al*. (2019).
13. The code is available at https://github.com/leifandersthorsrud/NCI.
14. See https://sborms.github.io/econometrics-meets-sentiment.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. and Zheng, X. (2016) TensorFlow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, (pp. 265–283). USENIX Association.

Adämmer, P. and Schüssler, R.A. (2019) Forecasting the equity premium: mind the news! Forthcoming in Review of Finance.

Allee, K.D. and DeAngelis, M.D. (2015) The structure of voluntary disclosure narratives: evidence from tone dispersion. *Journal of Accounting Research* 53(2): 241–274.

Andreou, E., Gagliardini, P., Ghysels, E. and Rubin, M. (2019) Inference in group factor models with an application to mixed frequency data. *Econometrica* 87(4): 1267–1305.

Angeletos, G.-M., Collard, F. and Dellas, H. (2018) Quantifying confidence. *Econometrica* 86(5): 1689–1726.

Angeletos, G.-M. and La'O, J. (2013) Sentiments. *Econometrica* 81(2): 739–779.

Antweiler, W. and Frank, M.Z. (2004) Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance* 59(3): 1259–1294.

Araújo, M., Diniz, J.P., Bastos, L., Soares, E., Júnior, M., Ferreira, M., Riberio, F. and Benevenuto, F. (2016) iFeel 2.0: a multilingual benchmarking system for sentence-level sentiment analysis. In *Proceedings of the 10th International AAAI Conference on Web and Social Media* (pp. 758–759).

Ardia, D., Bluteau, K., Borms, S. and Boudt, K. (2020) The R package sentometrics to compute, aggregate and predict with textual sentiment. Forthcoming in Journal of Statistical Software.

Ardia, D., Bluteau, K. and Boudt, K. (2019a) Media and the stock market: their relationship and abnormal dynamics around earnings announcements. Working Paper.

Ardia, D., Bluteau, K. and Boudt, K. (2019b) Questioning the news about economic growth: sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting* 35(4): 1370–1386.

Arslan-Ayaydin, Ö., Boudt, K. and Thewissen, J. (2016) Managers set the tone: equity incentives and the tone of earnings press releases. *Journal of Banking and Finance* 72: 132–147.

Baccianella, S., Esuli, A. and Sebastiani, F. (2010) SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*.

Bajo, E. and Raimondo, C. (2017) Media sentiment and IPO underpricing. *Journal of Corporate Finance* 46: 139–153.

Baker, M. and Wurgler, J. (2006) Investor sentiment and the cross-section of stock returns. *Journal of Finance* 61(4): 1645–1680.

Baker, M. and Wurgler, J. (2007) Investor sentiment in the stock market. *Journal of Economic Perspectives* 21(2): 129–152.

Baker, S.R., Bloom, N. and Davis, S.J. (2016) Measuring economic policy uncertainty. *The Quarterly Journal of Economics* 131(4): 1593–1636.
Bannier, C., Pauls, T. and Walter, A. (2019) Content analysis of business communication: introducing a German dictionary. *Journal of Business Economics* 89(1): 79–123.
Barsky, R.B. and Sims, E.R. (2012) Information, animal spirits, and the meaning of innovations in consumer confidence. *American Economic Review* 102(4): 1343–1377.
Benhabib, J. and Spiegel, M.M. (2019) Sentiments and economic activity: evidence from US states. *Economic Journal* 129(618): 715–733.
Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. and Matsuo, A. (2018) quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3(30): 774.
Bholat, D., Hans, S., Santos, P. and Schonhardt-Bailey, C. (2015) Text mining for central banks. Technical report, Centre for Central Banking Studies, Bank of England.
Bird, S., Klein, E. and Loper, E. (2009) *Natural Language Processing with Python*. Beijing: O'Reilly Media.
Blair, R.C. and Cole, S.R. (2002) Two-sided equivalence testing of the difference between two means. *Journal of Modern Applied Statistical Methods* 1(1): 139–142.
Blei, D.M. and Lafferty, J.D. (2006) Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (ICML '06), pp. 113–120. ACM.
Blei, D.M. and Lafferty, J.D. (2007) A correlated topic model of Science. *Annals of Applied Statistics* 1(1): 17–35.
Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.
Boiten, M. (2019) *rJST: Joint Sentiment Topic Modelling*, R package.
Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5: 135–146.
Borovkova, S., Garmaev, E., Lammers, P. and Rustige, J. (2017) SenSR: a sentiment-based systemic risk indicator. Technical Report 553, De Nederlandsche Bank.
Boudt, K. and Thewissen, J. (2019) Jockeying for position in CEO letters: impression management and sentiment analytics. *Financial Management* 48(1): 77–115.
Boudt, K., Thewissen, J. and Torsin, W. (2018) When does the tone of earnings press releases matter?*International Review of Financial Analysis* 57: 231–245.
Bradley, M.M. and Lang, P.J. (1999) Affective norms for English words (ANEW): instruction manual and affective ratings. Technical report.
Calomiris, C.W. and Mamaysky, H. (2019) How news and its context drive risk and returns around the world. *Journal of Financial Economics* 133(2): 299–336.
Caporin, M. and Poli, F. (2017) Building news measures from textual data and an application to volatility forecasting. *Econometrics* 5(3): 1–46.
Casey, G.P. and Owen, A.L. (2013) Good news, bad news, and consumer confidence. *Social Science Quarterly* 94(1): 292–315.
Ceron, A., Curini, L., Iacus, S.M. and Porro, G. (2014) Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16(2): 340–358.
Chang, C.-C., Hsieh, P.-F. and Wang, Y.-H. (2015) Sophistication, sentiment, and misreaction. *Journal of Financial and Quantitative Analysis* 50(4): 903–928.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017) Double/debiased/Neyman machine learning of treatment effects. *American Economic Review* 107(5): 261–265.
Chiou, L. and Tucker, C. (2017) Content aggregation by platforms: the case of the news media. *Journal of Economics & Management Strategy* 26(4): 782–805.
Croushore, D. and Stark, T. (2003) A real-time data set for macroeconomists: does the data vintage matter?*Review of Economics and Statistics* 85(3): 605–617.
Das, S.R. and Chen, M.Y. (2007) Yahoo! for Amazon: sentiment extraction from small talk on the web. *Management Science* 53(9): 1375–1388.

De Clercq, O., Lefever, E., Jacobs, G., Carpels, T. and Hoste, V. (2017) Towards an integrated pipeline for aspect-based sentiment analysis in various domains. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 136–142. ACM.
De Long, J.B., Shleifer, A., Summers, L.H. and Waldmann, R.J. (1990) Noise trader risk in financial markets. *Journal of Political Economy* 98(4): 703–738.
Denny, M.J. and Spirling, A. (2018) Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis* 26(2): 168–189.
Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018) BERT: pre-training of deep bidirectional transformers for language understanding. Working Paper.
Diamond, D.W. and Dybvig, P.H. (1983) Bank runs, deposit insurance, and liquidity. *Journal of Political Economy* 91(3): 401–419.
Eguchi, K. and Lavrenko, V. (2006) Sentiment retrieval using generative models. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 345–354. ACM.
Ekman, P. and Friesen, W.V. (1976) Measuring facial movement. *Environmental Psychology and Nonverbal Behavior* 1(1): 56–75.
Engle, R., Giglio, S., Kelly, B., Lee, H. and Stroebel, J. (2020) Hedging climate change news. *Review of Financial Studies* 33(3): 1184–1216.
Eshbaugh-Soha, M. (2010) The tone of local presidential news coverage. *Political Communication* 27(2): 121–140.
Evans, J.A. and Aceves, P. (2016) Machine translation: mining text for social theory. *Annual Review of Sociology* 42: 21–50.
Eyben, F., Weninger, F., Gross, F. and Schuller, B. (2013) Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pp. 835–838. ACM.
Feinerer, I., Hornik, K. and Meyer, D. (2008) Text mining infrastructure in R. *Journal of Statistical Software* 25(5): 1–54.
Feldman, R., Govindaraj, S., Livnat, J. and Segal, B. (2010) Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies* 15(4): 915–953.
Feuerriegel, S. and Pröllochs, N. (2019) *SentimentAnalysis: Dictionary-Based Sentiment Analysis*, R package.
Flaxman, S., Goel, S. and Rao, J. (2016) Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80: 298–320.
Friedman, J., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1): 1–22.
Fu, X., Yang, K., Huang, J.Z. and Cui, L. (2015) Dynamic non-parametric joint sentiment topic mixture model. *Knowledge-Based Systems* 82: 102–114.
Gandomi, A. and Haider, M. (2015) Beyond the hype: big data concepts, methods, and analytics. *International Journal of Information Management* 35(2): 137–144.
García, D. (2013) Sentiment during recessions. *Journal of Finance* 68(3): 1267–1300.
Garz, M. (2014) Good news and bad news: evidence of media bias in unemployment reports. *Public Choice* 161(3–4): 499–515.
Gelper, S. and Croux, C. (2010) On the construction of the European economic sentiment indicator. *Oxford Bulletin of Economics and Statistics* 72(1): 47–62.
Gelper, S., Peres, R. and Eliashberg, J. (2018) Talk bursts: the role of spikes in pre-release word-of-mouth dynamics. *Journal of Marketing Research* 55(6): 801–817.
Gentzkow, M., Kelly, B. and Taddy, M. (2019a) Text as data. *Journal of Economic Literature* 57(3): 535–574.
Gentzkow, M. and Shapiro, J.M. (2010) What drives media slant? Evidence from U.S. daily newspapers. *Econometrica* 78(1): 35–71.
Gentzkow, M., Shapiro, J.M. and Taddy, M. (2019b) Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica* 87: 1307–1340.
Glanzer, M. and Cunitz, A.R. (1966) Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior* 5(4): 351–360.
Glasserman, P. and Mamaysky, H. (2019) Does unusual news forecast market stress?*Journal of Financial and Quantitative Analysis* 54(5): 1937–1974.

Grimmer, J. and Stewart, B.M. (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3): 267–297.
Hamilton, W.L., Clark, K., Leskovec, J. and Jurafsky, D. (2016) Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 595–605. ACM.
Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A. and Ng, A.Y. (2014) Deep speech: scaling up end-to-end speech recognition. Working Paper.
Hansen, P., Lunde, A. and Nason, J. (2011) The model confidence set. *Econometrica* 79: 453–497.
Hatzivassiloglou, V. and McKeown, K.R. (1997) Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 174–181.
He, Y., Lin, C., Gao, W. and Wong, K.-F. (2013) Dynamic joint sentiment-topic model. *ACM Transactions on Intelligent Systems and Technology* 5(1): 1–21.
Henry, E. (2008) Are investors influenced by how earnings press releases are written?*Journal of Business Communication* 45(4): 363–407.
Heston, S. and Sinha, N. (2017) News vs. sentiment: predicting stock returns from news stories. *Financial Analysts Journal* 73(3): 67–83.
Hoerl, A. and Kennard, R. (1970) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12(1): 55–67.
Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42(1–2): 177–196.
Honnibal, M. and Montani, I. (2017) *spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing*, Python library.
Hu, M. and Liu, B. (2004) Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artificial Intelligence*, AAAI'04, pp. 755–760. AAAI Press.
Huang, X., Teoh, S.H. and Zhang, Y. (2014) Tone management. *Accounting Review* 89(3): 1083–1113.
Hubert, P. and Labondance, F. (2018) Central bank sentiment. Working Paper.
Hutto, C.J. and Gilbert, E. (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, Vol. 23, pp. 216–225.
Jegadeesh, N. and Wu, D. (2013) Word power: a new approach for content analysis. *Journal of Financial Economics* 110(3): 712–729.
Kalogeropoulos, A. (2018) Economic news and personal economic expectations. *Mass Communication and Society* 21(2): 248–265.
Kanayama, H. and Nasukawa, T. (2006) Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 355–363.
Kearney, C. and Liu, S. (2014) Textual sentiment in finance: a survey of methods and models. *International Review of Financial Analysis* 33: 171–185.
Kelly, B.T., Manela, A. and Moreira, A. (2019) Text selection. Working Paper.
Keynes, J.M. (1936) *The General Theory of Employment, Interest, and Money*. London, UK: Palgrave Macmillan.
Kräussl, R. and Mirgorodskaya, E. (2017) Media, sentiment and market performance in the long run. *European Journal of Finance* 23(11): 1059–1082.
Kuhn, M. (2018) *caret: Classification and Regression Training*, R package.
Labille, K., Gauch, S. and Alfarhood, S. (2017) Creating domain-specific sentiment lexicons via text mining. In *Proceedings of the 6th KDD Workshop on Issues of Sentiment Discovery and Opinion Mining*, pp. 1–9.
Lacy, S., Watson, B.R., Riffe, D. and Lovejoy, J. (2015) Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly* 92(4): 791–811.
Larsen, V.H. and Thorsrud, L.A. (2018) Business cycle narratives. Technical Report 7468, CESifo.
Larsen, V.H. and Thorsrud, L.A. (2019) The value of news for economic developments. *Journal of Econometrics* 210(1): 203–218.

Larsen, V.H., Thorsrud, L.A. and Zhulanova, J. (2020) News-driven inflation expectations and information rigidities. Forthcoming in Journal of Monetary Economics.
Lewis, C. and Young, S. (2019) Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research* 49(5): 587–615.
Lin, C. and He, Y. (2009) Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pp. 375–384. ACM.
Liu, B. (2015) *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press.
Liu, Y., Yu, X., Huang, X. and An, A. (2009) Blog data mining: the predictive power of sentiments. In P.S. Yu and C. Zhang (eds), *Data Mining for Business Applications* (pp. 183–195). Dordrecht: Springer.
Loria, S. (2019) *TextBlob: Simplified Text Processing*, Python library.
Loughran, T. and McDonald, B. (2011) When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance* 66(1): 35–65.
Loughran, T. and McDonald, B. (2014) Measuring readability in financial disclosures. *Journal of Finance* 69(4): 1643–1671.
Loughran, T. and McDonald, B. (2016) Textual analysis in accounting and finance: a survey. *Journal of Accounting Research* 54(4): 1187–1230.
Lowry, D. (2008) Network TV news framing of good vs. bad economic news under Democrat and Republican presidents: a lexical analysis of political bias. *Journalism & Mass Communication Quarterly* 85(3): 483–498.
Ludvigson, S.C. (2004) Consumer confidence and consumer spending. *Journal of Economic Perspectives* 18(2): 29–50.
Lukeš, J. and Søgaard, A. (2018) Sentiment analysis under temporal shift. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 65–71. ACM.
Lütkepohl, H. (2017) Estimation of structural vector autoregressive models. *Communications for Statistical Applications and Methods* 24: 421–441.
Manela, A. and Moreira, A. (2017) News implied volatility and disaster concerns. *Journal of Financial Economics* 123(1): 137–162.
Mayew, W.J. and Venkatachalam, M. (2012) The power of voice: managerial affective states and future firm performance. *Journal of Finance* 67(1): 1–43.
McCracken, M.W. and Ng, S. (2016) FRED-MD: a monthly database for macroeconomic research. *Journal of Business & Economic Statistics* 34(4): 574–589.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Vol. 2 (pp. 3111–3119).
Miller, G.A. (1995) WordNet: a lexical database for English. *Communications of the ACM* 38(11): 39–41.
Mohammad, S., Salameh, M. and Kiritchenko, S. (2016) How translation alters sentiment. *Journal of Artificial Intelligence Research* 55: 95–130.
Mohammad, S.M. and Turney, P.D. (2013) Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29(3): 436–465.
Munezero, M.D., Montero, C.S., Sutinen, E. and Pajunen, J. (2014) Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing* 5(2): 101–111.
Nimark, K.P. and Pitschner, S. (2019) News media and delegated information choice. *Journal of Economic Theory* 181: 160–196.
Nowak, A. and Smith, P. (2017) Textual analysis in real estate. *Journal of Applied Econometrics* 32(4): 896–918.
Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, Volume 10 of *EMNLP '02*, pp. 79–86.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Pennington, J., Socher, R. and Manning, C. (2014) GloVe: global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. ACM.
Petalas, D.P., vanSchie, H. and Vettehen, P.H. (2017) Forecasted economic change and the self-fulfilling prophecy in economic decision-making. *PLoS One* 12(3): e0174353.
Pham, H., Manzini, T., Liang, P.P. and Poczos, B. (2018) Seq2Seq2Sentiment: multimodal sequence to sequence models for sentiment analysis. In *Proceedings of the Grand Challenge and Workshop on Human Multimodal Language*, pp. 53–63. ACM.
Picault, M. and Renault, T. (2017) Words are not all created equal: a new measure of ECB communication. *Journal of International Money and Finance* 79: 136–156.
Poria, S., Cambria, E., Howard, N., Huang, G.-B. and Hussain, A. (2016) Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174(A): 50–59.
Pröllochs, N., Feuerriegel, S. and Neumann, D. (2015) Generating domain-specific dictionaries using Bayesian learning. In *Proceedings of the European Conference on Information Systems*, pp. 1–14.
Qin, D. (2011) Rise of VAR modelling approach. *Journal of Economic Surveys* 25(1): 156–174.
Ravi, K. and Ravi, V. (2015) A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems* 89: 14–46.
Remus, R., Quasthoff, U. and Heyer, G. (2010) SentiWS - a publicly available German-language resource for sentiment analysis. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, pp. 1168–1171. European Languages Resources Association (ELRA).
Ribeiro, F.N., Araújo, M., Gonçalves, P., Gonçalves, M.A. and Benevenuto, F. (2016) SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science* 5: 1–23.
Ridout, T.N., Fowler, E.F. and Searles, K. (2012) Exploring the validity of electronic newspaper databases. *International Journal of Social Research Methodology* 15(6): 451–466.
Riffe, D., Lacy, S., Watson, B.R. and Fico, F. (2019) *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. New York: Routledge.
Roberts, M.E., Stewart, B.M. and Airoldi, E.M. (2016) A model of text for experimentation in the social sciences. *Journal of the American Statistical Association* 111(515): 988–1003.
Rogers, J.L., Van Buskirk, A. and Zechman, S.L.C. (2011) Disclosure tone and shareholder litigation. *Accounting Review* 86(6): 2155–2183.
Rousseeuw, P., Raymaekers, J. and Hubert, M. (2018) A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics* 27: 345–359.
Saleiro, P., Rodrigues, E., Soares, C. and Oliveira, E. (2017) TexRep: a text mining framework for online reputation monitoring. *New Generation Computation* 35(4): 365–389.
Saltzis, K. (2012) Breaking news online. *Journalism Practice* 6(5–6): 702–710.
Scheufele, D.A. and Tewksbury, D. (2007) Framing, agenda setting, and priming: the evolution of three media effects models. *Journal of Communication* 57(1): 9–20.
Shapiro, A.H., Südhof, M. and Wilson, D. (2018) Measuring news sentiment. Technical Report 2017-01, Federal Reserve Bank of San Francisco.
Silge, J. and Robinson, D. (2016) tidytext: text mining and analysis using tidy data principles in R. *Journal of Open Source Software* 1(3): 37.
Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F. and Pantic, M. (2017) A survey of multimodal sentiment analysis. *Image and Vision Computing* 65: 3–14.
Soo, C.K. (2018) Quantifying sentiment with news media across local housing markets. *The Review of Financial Studies* 31(10): 3689–3719.
Stone, P.J., Dunphy, D.C. and Smith, M.S. (1963) The general inquirer: a computer approach to content analysis. In *Proceedings of the American Federation of Information Processing Societies spring joint computer conference*, pp. 241–256.
Strapparava, C. and Valitutti, A. (2004) WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*.
Taboada, M. (2016) Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2: 325–347.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37(2): 267–307.

Täckström, O. and McDonald, R. (2011) Discovering fine-grained sentiment with latent variable structured prediction models. In P. Clough, C. Foley, C. Gurrin, G.J.F. Jones, W. Kraaij, H. Lee and V. Mudoch (eds), *Proceedings of the Advances in Information Retrieval*, pp. 368–374. Springer.
Taddy, M. (2013a) Measuring political sentiment on Twitter: Factor optimal design for multinomial inverse regression. *Technometrics* 55(4): 415–425.
Taddy, M. (2013b) Multinomial inverse regression for text analysis. *Journal of the American Statistical Association* 108(503): 755–770.
Taddy, M. (2015a) Distributed multinomial regression. *Annals of Applied Statistics* 9(3): 1394–1414.
Taddy, M. (2015b) Document classification by inversion of distributed language representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 45–49.
Taddy, M. (2018) *textir: Inverse Regression for Text Analysis*, R package.
Teoh, S.H. (2018) The promise and challenges of new datasets for accounting research. *Accounting, Organizations and Society* 68–69: 109–117.
Tetlock, P.C. (2007) Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance* 62(3): 1139–1168.
Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008) More than words: quantifying language to measure firms' fundamentals. *Journal of Finance* 63(3): 1437–1467.
Thorsrud, L.A. (2018) Words are the new numbers: a newsy coincident index of the business cycle. *Journal of Business & Economic Statistics* 38(2): 393–409.
Tibshirani, R.J. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B* 58(1): 267–288.
Turney, P. (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 417–424. ACM.
Tversky, A. and Kahneman, D. (1981) The framing of decisions and the psychology of choice. *Science* 211(4481): 453–458.
Van de Kauter, M., Desmet, B. and Hoste, V. (2015) The good, the bad and the implicit: a comprehensive approach to annotating explicit and implicit sentiment. *Language Resources & Evaluation* 49(3): 685–720.
Varian, H.R. (2014) Big data: new tricks for econometrics. *Journal of Economic Perspectives* 28(2): 3–28.
Wang, H., Divakaran, A., Vetro, A., Chang, S.-F. and Sun, H. (2003) Survey of compressed-domain features used in audio-visual indexing and analysis. *Journal of Visual Communication and Image Representation* 14(2): 150–183.
Wang, J., Shen, H.T., Song, J. and Ji, J. (2014) Hashing for similarity search: a survey. Working Paper.
Wischnewsky, A., Jansen, D.-J. and Neuenkirch, M. (2019) Financial stability and the Fed: evidence from congressional hearings. Technical Report 633, De Nederlandsche Bank.
Young, L. and Soroka, S. (2012) Affective news: the automated coding of sentiment in political texts. *Political Communication* 29(2): 205–231.
Zhang, M.-L. and Zhou, Z.-H. (2014) A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8): 1819–1837.
Zhang, X., Zhao, J. and LeCun, Y. (2015) Character-level convolutional networks for text classification. In C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett (eds), *Advances in Neural Information Processing Systems* (pp. 649–657). Red Hook, NY: Curran Associates, Inc.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* 67(2): 301–320.

Appendix: Glossary

Corpus. A corpus in linguistics jargon designates the collection of textual data units (e.g., documents) to be analyzed. It can be generalized to indicate the collection of data units from textual, audio, or visual data.

Features. A feature is a broad term to represent any type of metadata attached to the original textual, audio, or visual data as stored in a corpus. Examples are source, expresser, entity, location, topic, and so on. This definition is slightly different but in line with how features are used in a machine learning context, where they refer to the set of explanatory variables. In video and audio data, (low-level) features are compact, mathematical representations of the physical properties of the data (Wang *et al*., 2003).

Lexicon. A lexicon is a list of tokens (e.g., words, a sequence of words, a facial expression, or a sound) with, for each token, an associated score that represents its average sentiment. Also interchangeably called a sentiment lexicon, a sentiment word list, or a sentiment dictionary.

Natural language processing (NLP). The broad subfield within artificial intelligence occupied with the understanding, interpretation, and manipulation of human language. It draws from computer science, computational linguistics, and machine learning.

Polarity. The polarity (or semantic orientation) of an expression (whether it is a text, a sound, or something else) represents its degree of positivity. Polarity categories go from very positive to very negative, discrete, or continuous.

Sentiment. Sentiment equals the disposition of an entity toward an entity, expressed via a certain medium. This working definition consists of (1) the expression by an entity of its disposition, in the form of verbal or non-verbal communication, (2) the expression has a polarity or a semantic orientation measurable on a discrete or a continuous scale, and (3) the expression is oriented toward (an aspect of) an entity.

Sentiment analysis. Sentiment analysis is about the extraction of sentiment from the medium it is expressed through. Multimodal sentiment analysis covers textual, audio, and visual media.

Sentometrics. The term "sentometrics" is a portmanteau of sentiment and econometrics. It deals with the computation of sentiment from any type of qualitative data, the evolution of sentiment, and the application of sentiment in an economic analysis using econometric methods.

Supervised learning. Supervised learning is a branch of machine learning that requires an annotated data set (i.e., a set of input data with associated output values) to train a model.

Unsupervised learning. Unsupervised learning is a branch of machine learning where the input data decide the output categories or representation by themselves. Any unsupervised method is typically hybrid or semisupervised, as there is often need for certain minimal inputs from the modeler.