ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Linking words in economic discourse: Implications for macroeconomic forecasts

J. Daniel Aromi

IIEP-Baires UBA-Conicet FCE, Universidad Catolica Argentina

ARTICLE INFO

Keywords:
Macroeconomic forecasting
Text-based data
Combining forecasts
Natural language processing
Uncertainty

ABSTRACT

This paper develops indicators of unstructured press information by exploiting word vector representations. A model is trained using a corpus covering 90 years of Wall Street Journal content. The information content of the indicators is assessed through business cycle forecast exercises. The vector representations can learn meaningful word associations that are exploited to construct indicators of uncertainty. In-sample and out-of-sample forecast exercises show that the indicators contain valuable information regarding future economic activity. The combination of indices associated with different subjective states (e.g., uncertainty, fear, pessimism) results in further gains in information content. The documented performance is unmatched by previous dictionary-based word counting techniques proposed in the literature.

1. Introduction

A large quantity of unstructured economic data is generated and disseminated everyday through multiple channels. For example, unstructured data is found in corporate and government documents, expert reports, mass media, and social media. Improvements in data availability and processing capacity have allowed for methods of summarizing and evaluating the information provided by unstructured data. Multiple works have demonstrated the relevance of unstructured data in macroeconomic and financial contexts (Alexopoulos & Cohen, 2015; Aromí, 2017, 2018; Baker, Bloom, & Davis, 2016; Balke, Fulmer, & Zhang, 2017; Hansen, McMahon, & Prat, 2017; Loughran & McDonald, 2011; Stekler & Symington, 2016; Tetlock, 2007). These contributions typically compute interpretable indicators that are based on a small set of keywords or predefined dictionaries. The resulting indicators are interpreted as metrics of uncertainty, pessimism, or the level of attention allocated to a topic of interest (e.g., recession).

One relevant question is whether natural language processing tools can interpret information in a more efficient manner. For example, can valuable indicators of uncertainty be built using these tools? How does the performance of these indices compare to the performance observed with more traditional methods? The gains in accuracy are a function of the efficiency of learning algorithm and the informativeness of the training corpus. Positive results would be relevant to macroeconomic analysis. More precise metrics can lead to the discovery of empirical regularities or the revision of previously estimated associations. In this work, the performance of a specific natural language processing tool is evaluated in the context of business cycle forecast exercises.

More specifically, the use of word vector representations (WVRs) is considered with an unsupervised learning model that has been successfully tested for natural language processing (Collobert, Weston, Bottou, Karlen, Kavukcuoglu, & Kuksa, 2011; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014). The algorithms for such models can be understood as dimensionality-reduction techniques that summarize word semantic and syntactic information. Following Pennington et al. (2014), word vector representations are trained using word co-occurrence statistics from a large

E-mail address: josedanielaromi@uca.edu.ar.

https://doi.org/10.1016/j.ijforecast.2019.12.001
0169-2070/

collection of text from The Wall Street Journal. Closely related words are identified by inspecting the similarity between their numeric representations. These associations allow for indicators with a straightforward interpretation.

Preliminary evaluations show that the resulting vectors are able to learn meanings in an economic context. Vectors are shown to resolve word ambiguities and recognize the relationships between economic entities. For example, the word "vice" can refer to immoral behavior or to rank. In the case of "vice president" for instance, trained vectors resolve this ambiguity in favor of the second option: according to the distance between vectors, the closest words are "executive" and "president". More relevantly, WVRs are shown to identify sets of related words (e.g., words related to uncertainty). In this way, WVRs allow for indicators whereby, instead of using predefined dictionaries or subjective judgments, relevant words are identified through quantitative information generated by unsupervised learning algorithms.

In the first set of exercises, a metric of uncertainty is evaluated. This choice reflects the prominent role assigned to the concept of uncertainty in the analysis of business cycles (Baker et al., 2016; Jurado, Ludvigson, & Ng, 2015; Rossi & Sekhposyan, 2015). In-sample, the indicators are shown to provide information on future levels of employment, industrial production, investment, and the gross domestic product (GDP). A one standard deviation increment in the uncertainty index anticipates, on average, a 0.40 standard deviation drop in GDP growth over the next year. Out-of-sample exercises are implemented using Bayesian model averaging (BMA). This forecast combination tool allows for data-driven discovery of efficient specifications. These exercises show that the indicators that exploit WVRs allow for significant gains in accuracy for one-quarter-ahead forecasts through eight-quarter-ahead forecast horizons.

Beyond indicators approximating uncertainty, complementary explorations consider indices that capture manifestations of different subjective states that are conjectured to be relevant. These additional subjective states are identified by inspecting previous literature and by recurrence to subjective judgment. The resulting indices summarize manifestations linked to pessimism, fear, and anxiety. Out-of-sample forecast exercises show that these alternative indices contain additional information that can be combined to attain higher accuracy.

The extent to which indices based on natural language processing techniques are more precise than traditional methods is unknown. Traditional methods are based on knowledge in the form of dictionaries and subjective judgments. As a result, they incorporate information that might allow for precise metrics. In the final set of exercises, the information content that results from traditional methods is compared to information content that results from the use of WVRs. A new set of business cycle forecast exercises is implemented with that purpose.

Four text analysis methodologies are considered: Baker et al. (2016), Economist (2001), Loughran and McDonald (2011), and Tetlock (2007). The indices proposed in Baker et al. (2016) and The Economist (2001) are based on the presence of a small sets of words. In Loughran and McDonald (2011) and Tetlock (2007), the indices are based on large lists of words built using predefined dictionaries and expert judgment. Forecasting exercises show that the performance of indices based on WVRs compares favorably with dictionary-based word counting techniques proposed in the literature. For example, in the case of one-year-ahead GDP forecasts, the forecasts based on traditional indices are unable to improve upon baseline forecasts. By contrast, uncertainty indices based on WVRs are able to generate highly significant improvements in forecast accuracy.

The rest of the paper is organized as follows. The next section presents the methodology and the data. The properties of trained word vectors are preliminarily explored in Section 3. Forecast exercises associated with the metric of uncertainty are presented in Section 4. Section 5 evaluates the information provided by multiple indices approximating alternative subjective states. In Section 6, comparisons with traditional methodologies are presented. Finally, Section 7 offers our conclusions.

2. Methodology and data

The construction of the indicators proposed in this work involves two steps. In the first step, GloVe (Pennington et al., 2014), a WVR model, is trained using a corpus covering 90 years of Wall Street Journal content (1900–1989). This unsupervised learning model can be understood as a dimensionality-reduction technique that summarizes semantic and syntactic information provided by word co-occurrence statistics. In the second step, indicators that summarize relevant aspects of press content are defined. This step involves identifying a relevant keyword and exploiting associations in trained WVRs. More specifically, having selected a relevant keyword or set of keywords (e.g., uncertainty), closely associated words are identified by computing the distance between their respective WVRs. The indicator is given by the frequency of these closely associated words. In this section, the methodology is outlined in more detail and the training corpus is presented. In Section 4, this type of index is computed to carry out in-sample and out-of-sample forecasting exercises.

2.1. Word vector representations

As previously indicated, the first step involves learning to represent words as vectors. The objective is to generate quantitative representations that summarize word semantic content. While there are multiple methods proposed in research on natural language processing (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Mikolov et al., 2013; Pennington et al., 2014), the common elements are the joint use of word occurrence statistics and some form of dimensionality-reduction technique.

The GloVe model (Pennington et al., 2014) implemented in the current study is an unsupervised learning model that summarizes word co-occurrence statistics—that is, information on the number of times a word appears in the context of other words. This information can be thought as a large, sparse matrix. The GloVe model associates each word in the dictionary with a vector whose

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.

dimensionality is typically between 100 and 600. In this way, with dictionaries containing more than 10,000 words, dimensionality is reduced by two or three orders of magnitude. This type of representation has been shown to extract word semantic (and syntactic) information efficiently (Collobert et al., 2011; Mikolov et al., 2013; Pennington et al., 2014). In particular, this quantitative representation can be used to assess the semantic similarity between different words. Relatedness is established by computing the distance between the respective vectors. Further, these models are able to establish analogies such as "queen is to king as woman is to man" in the form of the vector equation $queen - king = woman - man$. This type of relationship has been described as linear substructures of meaning. Although GloVe is not the only method that computes vector representations of words, it has been shown to perform better than alternative methods in multiple natural language processing tasks (Pennington et al., 2014). The authors argue that this is due to its ability to combine the benefits of global matrix factorization (such as latent semantic analysis (Deerwester et al., 1990)) and local-context window methods (such as the skip-gram model in Mikolov et al. (2013)).

The method is global in the sense that all vectors are computed in a single optimization exercise. Under this model, the objective is to generate word vectors such that their scalar product approximates as much as possible the number of times a word in the dictionary occurs in the context of any other word in the dictionary. Let W denote the dictionary and let X_{ij} denote the number of times word i occurs in the context of (i.e., is close to) word j. Also, let v_i denote the WVR of word i, and let $\tilde{v}_j$ denote the context WVR of word j. Then, under the GloVe model, the objective is to minimize the average distance between X_{ij} and $v_i \cdot \tilde{v}_j$.[1] The objective function allows a weighted average of this distance.

Formally, the loss function of the model is given by

$$\sum_i \sum_j f(X_{ij}) \left[v_i \cdot \tilde{v}_j + b_i + \tilde{b}_j - log(X_{ij}) \right]^2$$

where the minimization with respect to $\{v_i, \tilde{v}_i, b_i, \tilde{b}_i\}_{i \in W}$, $\{v_i\}_{i \in W}$ is the sequence of word vectors, $\{\tilde{v}_i\}_{i \in W}$ is the sequence of context word vectors, and the sequences of scalars $\{b_i\}_{i \in W}$ and $\{\tilde{b}_i\}_{i \in W}$ are the respective word biases. The word biases are used to account for differences in the frequency of words. The word vectors are the parameters of interest that summarize valuable word information. In the exercises below, each word i is represented by a vector of the sum of the two vectors: $v_i + \tilde{v}_i$. The weighting function $f(X_{ij})$ is increasing but concave, to limit the influence of frequent word co-occurrences.[2] This is a log-bilinear regression model. That is, the log of X_{ij} is projected linearly with respect to v_i and $\tilde{v}_j$. The model is fitted using stochastic gradient descent (Duchi, Hazan, & Singer, 2011). More details can be found in Pennington et al. (2014).

Following parameter values that are in line with those used in research on natural language processing, the vector dimensionality is 100 and the window size used to compute the term co-occurrence is 5. The vocabulary used in the implementation is given by words with a frequency of 100 or higher in the training corpus. An analysis of the robustness of this implementation indicates that the results are not sensitive to variations in the values of these parameters. Vector representations of words are computed using the package text2vec (Selivanov & Wang, 2016) in platform R. The same package was used in other related tasks (e.g., tokenization, computing a term co-occurrence matrix).

2.2. Quantitative indicators

In the second step, WVRs are used to construct quantitative indicators of information in the press.[3] The intention is to generate indicators that exploit knowledge captured by word vectors and can be interpreted in a straightforward manner. The procedure involves, first, identifying a keyword representing a relevant aspect of the content (e.g., "uncertainty"). Next, the set of K most closely related terms are found based on the cosine distance between the respective vectors. Finally, the indicator is given by the frequency of the selected words.

More formally, given dictionary W and keyword $k \in W$, the set of K closest words is identified by computing the cosine distance: $\frac{v_w \cdot v_j}{\|v_w\| \|v_j\|}$. Given this set of words, $K \subset W$, the index for a selected set of text C is given by

$$I_C^k = \frac{\sum_{w \in K} c_w}{\sum_{w \in W} c_w}$$

where c_w indicates the number of times word w is observed in the selected set of text C. The set of selected text in the exercises below is given by economic press content over a specific time window.

Considering the high level of attention placed on the concept of uncertainty (Baker et al., 2016; Jurado et al., 2015; Rossi & Sekhposyan, 2015), in the first set of exercises, an indicator for "uncertainty" is computed and evaluated. Beyond "uncertainty", other indices that approximate related but different manifestations in press content are constructed and evaluated. More specifically, these manifestations are "pessimism", "fear", and "anxiety". As explained below, these choices reflect previous literature and subjective judgments regarding relevance in macroeconomic contexts.

[1] The model generates two vectors for each word in the dictionary: a vector representation and a context vector representation. In practice, the information captured by these two vectors is combined by computing a simple average.

[2] More specifically, following Pennington et al. (2014), the weighting function $f(x) = (x/100)^{3/4}$ if $x < 100$, and otherwise $f(x) = 1$.

[3] Initially, these indicators were computed for the period of 1990–2016. Subsequently, in the implementation of out-of-sample forecast exercises, the indicators were computed for older periods.

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.

Table 1
Description of training corpus and test corpus.

Corpus	Number of articles	Number of tokens
Full Corpus (1900–2017)	4.475.187	233.776.933
GloVe Training Corpus (1900–1989)	3,233,481	134,797,611

2.3. Data

The corpus used to train the vectors was given by text published in the Wall Street Journal between 1900 and 1989. The selected text corresponded to the content made available by a public webpage.[4] For each article published in the newspaper, this website provided access to the headline, the lead, and a fraction of the body. To avoid concerns regarding forward-looking biases, the training corpus was constructed using a time period that predates the period of the forecasting exercises that are presented in the next section. Table 1 shows information on the corpus used to train the WVRs and the corpus used to compute the indicators.

In the initial text processing stage, numbers and punctuation marks are deleted from the text. Moreover, all text is converted to lower case and stop words are filtered.[5] After applying the minimum frequency filter, the dictionary of the training dataset comprised 28,296 words. This is the number of 100-dimensional vectors computed in the GloVe model implementation.

The training corpus is relatively small compared to some databases used in the field of natural language processing.[6] However, the training corpus is focused on economic discussions and can be considered to follow a relatively stable natural language. Additionally, the corpus used to compute the indices (i.e., the test corpus) shares the theme and style of the training corpus. As a result, there are reasons to remain optimistic regarding the implementation's ability to learn word meanings in an economic context.

Beyond text, a second set of data used in this study was given by real-time economic activity indicators for the years 1966 through 2017. Four variables were selected: employment (Nonfarm Payroll Employment), industrial production (Industrial Production Index: Manufacturing), investment (Real Gross Private Domestic Investment: Nonresidential), and GDP.[7] The information was obtained from the Federal Reserve Bank of Philadelphia's Real-Time Data Research Center.[8] The database built for the exercises below preserves the real-time nature of the original data. More specifically, for each sample quarter t, the values of the economic activity indicators, current values, and lagged values are given by information available at the time that the data corresponding to quarter t was first released. For example, real GDP data for the third quarter of 1999 is given by information released on 28 October 1999. Table 2 provides descriptive statistics.

Table 2
Descriptive statistics: Quarterly growth rates.

Activity indicator	Mean	St. Dev.	Min	Max
Employment	0.0039	0.0051	−0.0202	0.0166
Industrial production	0.0060	0.0186	−0.1098	0.0598
Investment	0.0085	0.0228	−0.1190	0.0531
GDP	0.0061	0.0074	−0.0274	0.0264

Note: Figures correspond to first releases. Sample period is from 1966–2017.

3. Preliminary analysis of trained word vectors

Before proceeding to the forecasting exercise, preliminary evaluations of the information captured by word vectors are presented. First, some selected associations between word vectors are used to demonstrate that trained vectors are able to learn word meanings in an economic context. Second, an indicator designed to capture manifestations of "uncertainty" is evaluated in terms of its contemporaneous association with relevant macroeconomic events.

3.1. Vectors and meaning in economic context

The extent to which WVRs capture meaning is an empirical matter that depends on the informativeness of the training corpus and the efficiency of the learning model. One reason for concern is that, as previously indicated, the training corpus is relatively small compared to corpora typically used in the field of natural language processing. A collection of tasks was used to evaluate the latent information embodied in trained vectors. The evaluations implemented below are based on associations between trained vectors. Three tasks are considered below: resolution of ambiguity in word meaning, entity identification through vector composition, and identification of words indicative of tone or topic. The last task is the most relevant for the construction of indicators that reflect information in the press.

Ambiguous words are a common challenge in natural language processing applications. In particular, they are a problem for indicators based on predefined dictionaries. For example, Aromí (2018), Garcia (2013), and Tetlock (2007) have shown that words in the negative category of the Harvard IV dictionary can be used to anticipate financial and macroeconomic dynamics. Nevertheless, this category includes ambiguous words such as "capital", "tire", and "vice". In economic contexts, these words are not likely to transmit negative information. The presence of

[4] The text was extracted from: http://pqasb.pqarchiver.com/djreprints/. While this webpage is no longer available, similar content can be found in the WSJ archive (https://www.wsj.com/news/archive/).

[5] The list of stop words can be found in Appendix.

[6] For example, in Pennington et al. (2014) WVRs are trained using corpora ranging from 1 billion tokens to 42 billion tokens.

[7] For National Income and Product Accounts, the information reported in the real-time dataset is the quarterly advance release.

[8] The data can be downloaded from https://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/.

the word "capital" typically reflects discussions regarding financial issues, rather than discussions regarding the death penalty. Similarly, the word "tire" typically refers to the manufacturing of rings of rubber, rather than the need for rest or sleep. In the case of "vice" the most likely use is linked to the title of a corporate or bureaucratic position (e.g., vice chairman). The use of this word to refer to immoral or wicked behavior will be less common.

Table 3 shows that for each of the above-mentioned ambiguous terms, there is a set of words with the closest vectors. The selected words suggest that word vectors are able to identify the most likely meaning. For example, in the case of "tire" the closest terms are related to the manufacturing of rings of rubber: "goodyear", "firestone", and "akron". A final example of an ambiguous word is given by "default". The set of closest terms ("payment", "debt", and "obligations") suggests that the identified meaning points to failures to fulfill an obligation, rather than to preselected options. These examples are suggestive of potential efficiency gains associated with the use of unsupervised learning algorithms instead of predefined dictionaries. In addition, it is observed that ambiguous, context-dependent natural language requires the acquisition of knowledge through field-specific collections of text.

WVRs have been shown to learn relationships between words (Mikolov et al., 2013). For example, in natural language processing tasks, computed vectors have been shown to learn associations such as "king"-"male"+ "female" → "queen". This type of association can be used to identify related entities in economic settings. A couple of examples are shown in Table 3. The results suggest that vectors trained using economic press content learn to identify relationships between government entities and manufacturing corporations.

Finally, and more relevant to the current analysis, vectors are shown to identify groups of words related to the tone or the topic in a collection of texts. Valuable associations were indeed learned. For example, the word "uncertainty" was identified as close to other words that manifest negative, forward-looking, emotional, and cognitive states. This evidence indicates that these associations between words can be used to construct indices that approximate "uncertainty" as communicated by economic press content. As an additional example that shows that WVRs can be used to identify topics, vectors also learned to identify words related to "debt".

3.2. Indicator of uncertainty

In this subsection, an indicator of uncertainty is presented. As previously indicated, the concept of uncertainty has received substantial attention in the analysis of business cycles (Baker et al., 2016; Jurado et al., 2015; Rossi & Sekhposyan, 2015). Choosing "uncertainty" as a keyword is a natural choice given this related literature. In the first step of the computation, the vectors associated with the words "uncertainty", "uncertain", and "uncertainties" are added to compute a new vector w_U. The distance between this new vector and all words in the vocabulary W is computed. As previously described, the set of K words whose vectors are closest to w_U are then selected. Finally, the index is given by the frequency of these K words. Indices were computed to compile the second corpus, covering material published between January 1990 and February 2017. This second dataset contains approximately 98 million tokens.

Fig. 1 shows the values of three specifications of the uncertainty indicator. Each index was computed using a different number of words related to uncertainty. Increments in the indices can be observed around the three recessions that took place during the sample period. This increment is particularly conspicuous in the case of the recession linked to the 2008 global financial crisis. Interestingly, in the case of the 2007–2009 recession, the indices show increments several months before the start of the recession in December 2007. Additionally, spikes in the indices are observed around three well-known crisis episodes: the Asian crisis of 1997, the Russian crisis of 1998, and the 2011 debt-ceiling crisis. These associations suggest that meaningful information is captured by the index. Its ability to anticipate economic activity is evaluated in the following sections.

4. Macroeconomic forecasts

In this section, the information content of the indicators of uncertainty is assessed through business cycle forecast tasks. Beyond its intrinsic value, these exercises can serve as a general gauge of the relevance of these indicators for macroeconomic analysis. Positive results would suggest that academics and policymakers can benefit from the application of natural language processing techniques to large collections of unstructured data.

The first group of forecasting tasks involves in-sample exercises. In this case, the focus is placed on characterizing statistically and economically significant associations between indicators of information in the press and subsequent business cycle trajectories. A second group of exercises involves out-of-sample exercises. In this second case, gains in forecast accuracy are evaluated.

The forecasting models are given by autoregressive specifications that are complemented with an indicator of lagged press content. The growth rate of each economic activity indicator over the following h quarters is modeled as a function of lagged quarterly growth rates. The number of lags is selected by minimizing the Bayesian information criterion.

More formally, let a_t be the value of an economic indicator in quarter t. The growth rate computed for quarter t is given by $\Delta a_t = log(a_t) - log(a_{t-1})$. Let $\Delta^h a_t$ represent the growth rate computed for quarter t for a window of size h. That is, $\Delta^h a_t = log(a_t) - log(a_{t-h})$. The baseline autoregressive model satisfies

$$\Delta^h a_{t+h} = \alpha + \sum_{s=0}^{p} \beta_s \Delta a_{t-s} + u_t \quad (1)$$

To evaluate the predictive ability of indicators based on press content, this baseline model is modified by incorporating an indicator of press content as a predictor. Let I_t represent the value of an indicator of press content

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.

Table 3
Sample evidence on unsupervised learning of word meaning.

Selected vector	5 closest word vectors
Ambiguous words:	
Tire	Goodyear, firestone, akron, tires, rubber
Capital	Par, authorized, outstanding, shares, common
Vice	Executive, president, elected, director, manager
Default	Payment, debt, obligations, overdue, waiver
Vector compositions:	
Bundesbank-Germany+US	Fed, regulators, intervention, analysts, agency
Gm-cars+planes	Boeing, northrop, lockheed, aircraft, fighter
Tone/topic keywords:	
Uncertainty	Confusion, nervousness, apprehension, uneasiness, anxiety
Debt	Funding, longterm, financing, subordinated, restructure

Note: The distance is computed using cosine similarity. The closest words exclude words with the same root.

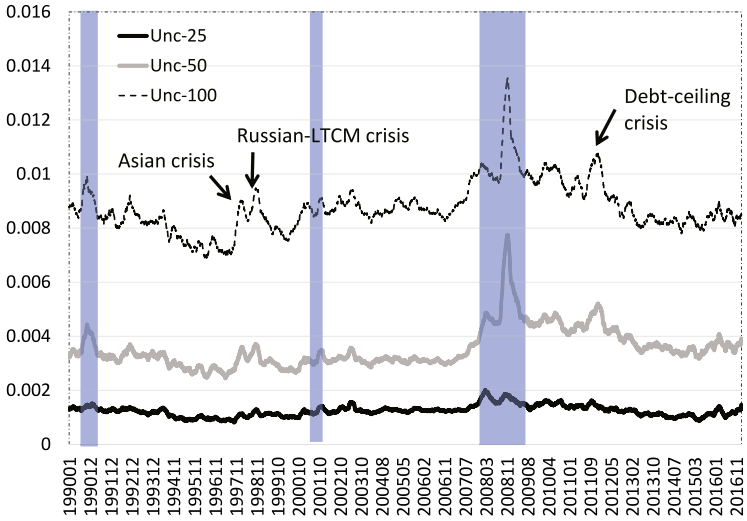

Fig. 1. Uncertainty indices. Note: The figure shows the average of the indices for 90-day moving windows. Horizontal bars indicate recessions.

corresponding to quarter t. Then, the forecasting model is given by the following equation:

$$\Delta^h a_{t+h} = \alpha + \sum_{s=0}^{p} \beta_s \Delta a_{t-s} + \beta_I I_t + u_t \quad (2)$$

The parameter of interest is β_I. Also, the relative metric of model fit, as indicated by increments in adjusted R^2s, is analyzed to assess the in-sample forecasting performance of the indicator. Models were estimated for the period from 1990–2017.[9] In this way, WVRs were trained with press content published before year 1990. As a result, they do not contain any forward-looking information.

In the first set of evaluations, the indicator of uncertainty is computed using the set of 100 most closely related words. The index is given by lagged frequency of words in this set during the previous 90 days. While the optimal specification of the indicator is unknown, this specification is used to provide a first evaluation of the information content. In the out-of-sample exercises developed below, convenient specifications were learned by implementing a flexible Bayesian model averaging framework.

When in-sample forecasting exercises are implemented, the estimated parameters reflect all information in the dataset, including future information. Beyond this feature, the exercise was designed to ensure that no other forward-looking elements are incorporated.[10] In particular, the forecasting exercise was carefully designed to take into account the schedule of economic data release. For each instance of the forecasting exercise, any information used to produce the forecast must have been available at the time the forecast was generated. Each forecast exercise is simulated to take place on the day in which a new quarterly figure is released. All information released on that day is incorporated into the information set. The indicator of press information I_t summarizes lagged press

[9] In the out-of-sample forecast exercises presented in the next subsection, pre-1990 data was used to train the forecasting model. In the current exercise, this data was not used, to avoid overlaps between the period used to train the GloVe model and the period of the in-sample forecast exercise.

[10] In the next subsection, out-of-sample exercises are described where forward-looking information in the estimated parameters has been eliminated.

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.

Table 4
Estimated forecast models.

	$h=1$	$h=2$	$h=4$	$h=8$
Employment				
$\hat{\beta}_I$	−0.263**	−0.342**	−0.408**	−0.389
t-stat.	[2.24]	[2.25]	[2.05]	[1.29]
Adj. R^2	0.710	0.659	0.523	0.315
Baseline adj. R^2	0.666	0.583	0.412	0.214
Industrial production				
$\hat{\beta}_I$	−0.280*	−0.341	−0.298	−0.180
t-stat.	[1.68]	[1.40]	[0.77]	[0.43]
Adj. R^2	0.414	0.290	0.155	0.029
Baseline adj. R^2	0.352	0.194	0.084	0.009
Investment				
$\hat{\beta}_I$	−0.342***	−0.396**	−0.354	−0.297
t-stat.	[2.96]	[2.00]	[1.23]	[0.94]
Adj. R^2	0.328	0.358	0.287	0.114
Baseline adj. R^2	0.230	0.225	0.180	0.041
GDP				
$\hat{\beta}_I$	−0.320***	−0.387***	−0.387***	−0.370**
t-stat	[3.85]	[3.01]	[2.44]	[2.06]
Adj. R^2	0.299	0.280	0.239	0.151
Baseline adj. R^2	0.217	0.156	0.113	0.035

*Significance levels: 0.10.
**Significance levels: 0.05.
***Significance levels: 0.01.
Standard errors are estimated following Newey and West (1987) and Newey and West (1994). Parameter estimates are standardized; absolute t-statistics are shown in brackets.

content up to 90 days before the release of the corresponding economic activity indicator. In other words, the forecasting exercise evaluates the predictive value of indicators of press content, immediately after incorporating the news proceeding from the first release of quarterly economic activity data.[11]

In the case of industrial production, the release dates are reported by the Board of Governors of the Federal Reserve System.[12] In the case of payrolls, the first release of information corresponding to any given month takes place on the first day of the following month. In the case of this variable, the index was cautiously constructed using information published no later than the last day of the month, when information was published for the first time. In the case of National Income and Product Accounts, starting in 1996, release dates are available from the Bureau of Economic Analysis webpage.[13] For earlier dates, release dates are not available. The release date was assumed to be the 28th day of the month of the release—that is, one day earlier than the average release day observed in the 1996–2017 period.

Table 4 shows evidence of the information content of the selected indicator. Four forecast horizons were considered: $h \in \{1, 2, 4, 8\}$. Adjusted R^2s show important gains in explanatory ability. This is especially clear in the case of longer forecasting horizons. In the case of one-year-ahead GDP forecast models, as the press content indicator is incorporated as a predictor, the adjusted R^2 increases from 0.113 to 0.239.

In all cases, the estimated coefficient is negative. Also, the fitted models point to economically significant associations. An increment of one standard deviation to the information metric anticipates a mean drop in economic activity growth that ranges from 0.18 to 0.40 standard deviations. For short-term forecast horizon models, statistically significant associations can be observed. As the forecast horizon grows, the number of statistically significant associations decreases. The indicator of press content is seen to be consistently informative in the case of GDP forecasts. By contrast, when industrial production forecasts are considered, the associated parameter is statistically significant only in the case of the shortest forecast horizon.

This preliminary evidence suggests that indices that exploit WVRs have information regarding future levels of economic activity. In particular, this can be inferred from increments in adjusted R^2s as these indices are incorporated in the forecasting models. At the same, the estimated associations are not always statistically significant. This could be the result of inefficient specification of the index reflecting information in the press. For example, the appropriate weight to assign to words characterized by different levels of associations is unknown with regard to the concept of uncertainty. In the current specification, zeros and ones are assigned based on an arbitrary threshold. It is moreover reasonable to assume that more recent information should be allocated heavier weights. In the out-of-sample forecast exercises presented below, these issues are dealt with by implementing a Bayesian model averaging framework.

4.1. Out-of-sample exercises

The previous evidence regarding in-sample predictive ability is here extended to an implementation of a set of out-of-sample forecast exercises. Forecasts were generated using models fitted with real-time data. Four different forecast horizons were considered: $h \in \{1, 2, 4, 8\}$. The test sample starts in the year 1990. Models were trained using expanding windows of historic data that begins in 1966 and ends h quarters before the date in which the corresponding forecast exercise was implemented.

The analysis implements forecast combinations to acknowledge uncertainty regarding optimal specifications of indicators that summarize information in the press. First, it must be noted that the choice of 100 words used in the previous forecasting exercises was arbitrary. A more efficient approach would allow for larger weights allocated to more closely related words. Also, optimal indicators would assign more weight to more recent information. Considering these concerns, indices associated with a different number of words and alternative lagged windows were incorporated in the exercise. The forecasts associated with each index were combined using BMA techniques. In this way, the weights assigned to different

[11] In the case of data that is published on a monthly basis (e.g., payrolls and industrial production), the exercises were carried out four times a year. More specifically, the exercises were carried out in January, April, July, and October.
[12] The list can be found at https://www.federalreserve.gov/releases/g17/.
[13] https://www.bea.gov/newsreleases/releasearchive.htm

Table 5
Out-of-sample predictive accuracy.

	$h=1$	$h=2$	$h=4$	$h=8$
Employment	0.921	0.907	0.967	0.996
	[0.00]	[0.00]	[0.08]	[0.44]
Industrial production	0.941	0.954	1.004	1.011
	[0.00]	[0.00]	[0.39]	[0.59]
Investment	0.943	0.939	0.911	0.945
	[0.00]	[0.00]	[0.00]	[0.07]
GDP	0.945	0.936	0.925	0.875
	[0.00]	[0.00]	[0.00]	[0.02]

Note: Relative RMSPEs; bootstrapped p-values are reported in square brackets for the test of the null hypothesis that the ratio of the RMSPEs is equal to one.

set of words and information with different lags were learned using historic regularities. In other words, the estimation of optimal forecast combination was used as a strategy to deal with risks associated with unknown models (Allan G., 2006).

Baseline forecasts correspond to those generated by the autoregressive model. As in the previous in-sample exercises, the number of lags was selected according to the Bayesian information criterion. Forecasts generated by the baseline model were compared to the combination of forecasts informed by different indicators of uncertainty in the press. To consider variation in the informativeness of more closely related words, indices with different number of related words were considered. The number of words used to construct the index was 100, 50, or 25. Keeping in mind that more recent news flow might be more informative, two window sizes for lagged information were considered. In addition to the previously proposed 90-day window specification, indices based on 30-day windows were considered. These alternative specifications resulted in six indicators of information in the press. Let $\{I_t^i\}_{i=1}^6$ represent the indices computed under the alternative specifications. Then, given a variable measuring economic activity a_t and a forecast horizon h, each indicator of uncertainty defines a forecasting model that satisfies

$$\Delta^h a_{t+h} = \alpha^i + \sum_{s=0}^{p} \beta_s^i \Delta^s a_t + \beta_I^i I_t^i + u_t^i \tag{3}$$

where u_t^i is normally distributed with mean 0 and variance σ_i^2. BMA exercises incorporate six models associated with different specifications of the indices and the baseline autoregressive model. Under BMA, forecast combinations involve computing weighted averages of the forecasts generated under each model. The weights are given by the posterior probability that the corresponding model is the true model. The current implementation follows the specification proposed in Faust, Gilchrist, Wright, and Zakrajšek (2013).

More formally, let $\{M_i\}_{i\in N}$ be a collection of models. Further, let θ_i represent the parameters $\{\alpha^i, \beta_1^i, \ldots, \beta_p^i, \beta_I^i, \sigma_i^2\}$, and let D be the observed data. Then, the posterior probability is given by

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_{j\in N} P(D|M_j)P(M_j)} \tag{4}$$

where

$$P(D|M_i) = \int P(D|\theta_i, M_i)P(\theta_i|M_i)d\theta_i \tag{5}$$

is the marginal likelihood of the ith model, $p(\theta_i|M_i)$ is the prior density of the parameter vector θ_i, and $P(D|\theta_i, M_I)$ is the likelihood function. Following the usual practice, it was originally assumed that the prior probabilities are the same for all models. It was also assumed that the prior density of the parameters $\{\alpha^i, \beta_1^i, \beta_p^i, \beta_I^i, \sigma_i^2\}$ is uninformative and proportional to $1/\sigma_i$. The prior for parameter β_I^i follows Zellner (1986), the g-prior specification: $\beta_I^i \sim N(0, \phi\sigma_i^2(I_i'I_i)^{-1})$. The parameter $\phi > 0$ controls the strength of the prior. Following previous literature, this parameter value was set to $\phi = 4$.[14] After computing the forecasts associated with each model, $\hat{a}_{t+h}^i$, and updating beliefs, the forecast combination is given by

$$\Delta^h \hat{a}_{t+h} = \sum_{i=0}^{N} P(M_i|D)\Delta^h \hat{a}_{t+h}^i \tag{6}$$

Table 5 shows information on the accuracy of forecasts that incorporate information from the press.[15] More specifically, the table shows the ratio between the root mean square error that results from the BMA approach and the root mean square error of the baseline model. The table also shows the p-values for the test of the null hypothesis that the ratio is equal to one. Given the presence of nested models, p-values are based on the bootstrap methods implemented in Faust et al. (2013). Gains in forecast accuracy are observed for most activity indicators and forecast horizons. For short-term forecast horizons ($h = 1$ and $h = 2$), gains in accuracy are statistically significant with p-values below 0.01. GDP forecasts show the most consistent gains associated with lagged information from the press. By contrast, in the case of industrial production, one-year-ahead and two-year-ahead forecasts are not seen to improve when compared to baseline forecasts.

The reported gains in forecast accuracy are consistent with the positive results observed in previously reported in-sample forecast exercises. Additionally, these results are indicative of gains associated with a data-driven selection of the specification of indicators of press content.

5. Indices measuring other subjective states

So far, the analysis has focused on indicators that capture expressions associated with uncertainty. The forecasting exercises above indicate that these indices contain information regarding the future evolution of the business cycle. The choice of uncertainty-related indices is a

[14] See for example Faust et al. (2013) and Fernandez, Ley, and Steel (2001). The results are not sensitive to changes in this parameter. These robustness exercises are available from the author upon request.

[15] The BMA implementation was estimated using the R package, BMS.

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.

Table 6
Words related to selected keywords.

Selected keyword	10 closest word vectors
Uncertainty	Uncertainties, confusion, nervousness, uncertain, apprehension, uneasiness, anxiety, feeling, fears, situation
Fear	Fears, worry, feared, causing, danger, worried, cause, trouble, talk, worries,
Pessimism	Optimism, feeling, prevalent, anxiety, uneasiness, apprehension, gloom, discouragement, prevails, persists
Anxiety	Uneasiness, apprehension, nervousness, causing, confusion, uncertainty, pessimism, disappointment, excitement, feeling

Note: The distance is computed using the cosine distance.

natural choice given the theoretical and empirical contributions that have focused on this concept in the context of business cycle studies (Baker et al., 2016; Jurado et al., 2015; Rossi & Sekhposyan, 2015).

On the other hand, the evaluation of indicators associated with alternative aspects communicated in the press is a logical extension to the previous exercise. It is likely that the uncertainty proxy does not capture all relevant factors in an appropriate manner. In this way, increases in information content can result from the consideration of additional indicators. In particular, proxies of alternative subjective states can be considered. In the exercises shown below, three types of related but different indicators are incorporated. The choice of these additional subjective states was guided by previous literature and subjective judgment. While it is acknowledged that it would be desirable to have a more systematic approach to feature selection, this is beyond the scope of the current exercise and is left for future explorations.

First, considering the current forecast task associated with business cycles, manifestations in the press related to "pessimism",—that is, an expectation of negative scenarios—are considered relevant. Suggesting potential complementarities, pessimism can be viewed as a first-moment feature of subjective states, whereas uncertainty can be linked to second-moment features. Moreover, pointing to a very prominent emotion, expressions related to "fear" are used as another potentially informative indicator. The perception of fear can be linked to the detection of threats that are likely to have behavioral correlates. Thus, it can be observed that the intensity of web searches related to "fear" has predictive value regarding investment decisions and stock market volatility (Da, Engelberg, & Gao, 2014). The fourth subjective state approximated through indicators is anxiety. In Nyman, Kapadia, Tuckett, Gregory, Ormerod, and Smith (2018), it is suggested that expressions related to "anxiety" capture important information regarding subjective states and associated behavior. The role of anxiety in the business cycle has also been stressed in Delis, Kouretas, and Tsoumas (2014).[16]

Table 7
Subjective states indices — Correlations.

	Uncertainty	Fear	Pessimism	Anxiety
Uncertainty	1	0.9181	0.795	0.844
Fear	–	1	0.780	0.809
Pessimism	–	–	1	0.885
Anxiety	–	–	–	1

Note: The indices are computed using the 100 most closely related words and 90-day windows.

The words most closely related to the selected keywords are shown in Table 6. These words are, for the most part, consistent with the expected associations. Most of these words point to negative emotional and cognitive states. It is also worth observing that some selected words do not contain a subjective element. For example, "cause", "situation", and "trouble" are words that, in principle, do not refer to emotions or other subjective states. Finally, it can be observed that "optimism" is the word most closely associated with "pessimism". This outcome suggests that the words "optimism" and "pessimism" are used in very similar contexts. After inspecting the other words associated with "pessimism", it can be conjectured that these contexts are predominantly negative.

Beyond specific observations, overall, the associations suggest that the selected keywords allow for the construction of indices that extract relevant information from unstructured data. This exploratory evidence also shows that these concepts are closely linked. For example, words such as "uncertainty", "uneasiness", and "anxiety" appear in Table 6 on multiple occasions. As a result, indices associated with these concepts are expected to have an important common component. At the same time, differences in these indices might allow for data-driven identification of the optimal weight for each indicator. Beyond rankings, complementarities between these closely related indicators can also be supposed. These possibilities are formally evaluated through a new set of business cycle forecast exercises.

The close association across different indicators confirms the correlation statistics in Table 7. The indices

[16] In related explorations, indices associated with positive words such as "optimism" and "excitement" were evaluated. No predictive value was observed in this case. This can be linked to the Pollyanna Hypothesis, according to which positive words are used more diversely and do not carry as much information as negative words. Consequently, negative words must be used in a more discriminatory manner (Boucher & Osgood, 1969; Garcia, Garas, & Schweitzer, 2012). Relatedly, Aromí (2017) and Tetlock (2007) observed that, in contrast to negative words, positive words do not provide any information regarding future stock market returns.

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.

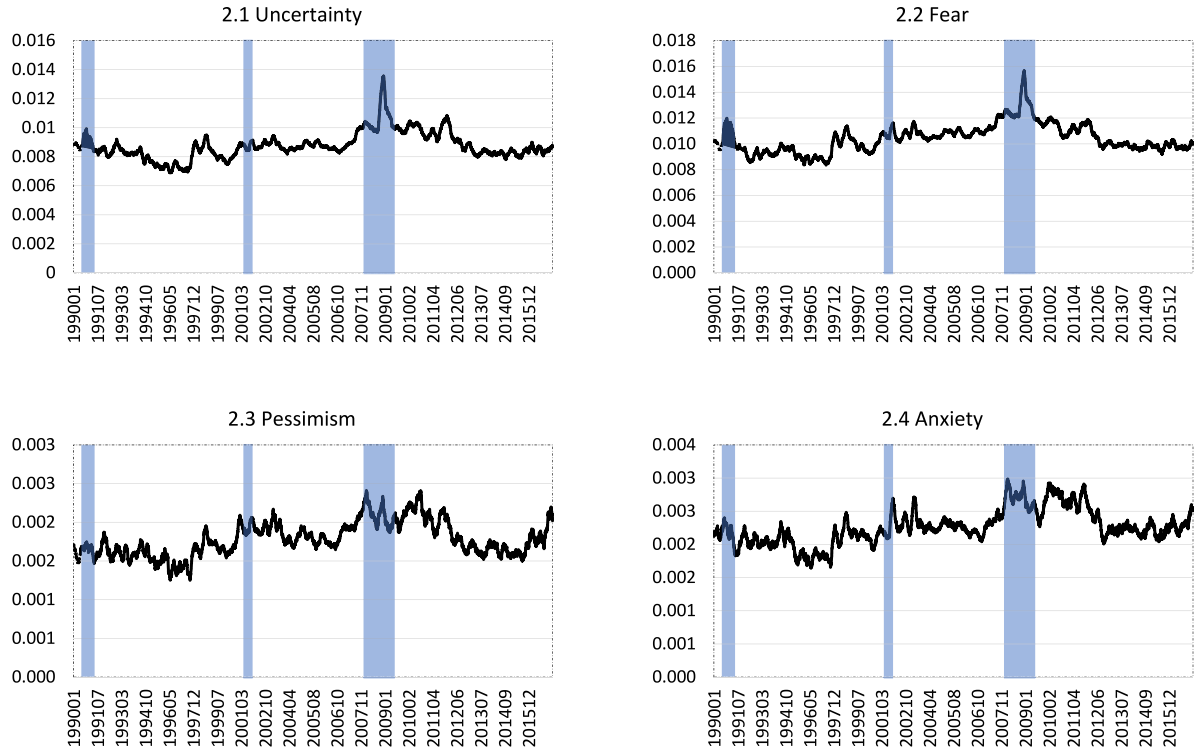

Fig. 2. Subjective state indices. Note: Indices are computed using the 100 most closely related words and 90-day windows.

were computed using 90-day windows. Correlation coefficients range from 0.78 to 0.92. The strongest correlation corresponds to the indicators approximating uncertainty and fear. The lowest (though still high) correlation corresponds to the indices measuring pessimism and fear. Additional preliminary evaluations result from inspecting Fig. 2. Consistent with the computed correlations, the indices are seen to co-move. In all cases, the lowest figures are observed during the mid-1990s; increments are detected around recessions; the 2008–2009 crisis is associated with important and persistent increments. However, some potentially meaningful differences can be distinguished. For example, the anxiety index has distinctive spikes around 9/11 and the 2003 invasion of Iraq. The increments around those events are not as prominent in the case of the other indices. In another example, the fear index and the uncertainty index have acute spikes around the most severe stage of the 2008–2009 crisis and these are not observed in the pessimism index or the anxiety index.

The first set of formal exercises involves the individual evaluation of the indices through dynamic regressions that incorporate one of the indicators as a predictor. In these evaluations, as in the previous case of the index approximating uncertainty, the indices reflect the frequency of the set of 100 most closely related words. In the case of pessimism and fear, multiple keywords are used to construct the respective indices. In the case of pessimism, the adjective "pessimistic" is also used as a keyword. In addition to the word "fear", the associated index was built using the words "fears" and "feared". In these cases, the vector representations of keywords were added and associated words were identified using this composite vector.[17]

Table 8 shows information for in-sample forecasting exercises. In all cases, the estimated coefficients are negative and the adjusted R^2s increase. Additionally, in most cases, the estimated coefficients differ significantly from zero. The indices approximating uncertainty and fear are seen to contain more information regarding future levels of economic activity. Moreover, the best performing index is not always the same, suggesting that forecast combinations might allow for more precise forecasts. While the uncertainty index seems to generate the most informative forecasts in the case of GDP and investment forecasts, the index associated with fear shows the strongest performance when employment and industrial activity are considered. The coefficients associated with indices approximating pessimism and anxiety are in most cases statistically significant. Nevertheless, the associations are clearly weaker, as indicated by the p-values and the absolute value of the standardized estimated coefficients.

As indicated in the previous section, this evidence might understate the information that can be inferred using indicators of press content based on WVRs. The failure to identify statistically significant associations might reflect inefficiencies in the selected specifications for the

[17] In the case of anxiety, the adjective "anxious" was not incorporated because it is an ambiguous word that can be linked to positive content. Consistent with this choice and according to the cosine distance, the computed vector representations for "anxiety" and "anxious" were dissimilar.

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.

Table 8
Estimated forecast models, $h = 1$.

	Uncertainty	Fear	Pessimism	Anxiety
Employment (baseline adj. $R^2 = 0.666$)				
$\hat{\beta}_I$	−0.263**	−0.296***	−0.109**	−0.128**
t-stat.	[2.24]	[3.16]	[2.02]	[2.09]
Adj. R^2	0.710	0.723	0.672	0.677
Ind. Prod. (baseline adj. $R^2 = 0.352$)				
$\hat{\beta}_I$	−0.280*	−0.314**	−0.139	−0.153
t-stat.	[1.68]	[2.18]	[1.65]	[1.57]
Adj. R^2	0.414	0.429	0.364	0.369
Investment (baseline adj. $R^2 = 0.230$)				
$\hat{\beta}_I$	−0.342***	−0.332***	−0.158*	−0.168*
t-stat.	[2.96]	[3.31]	[1.73]	[1.81]
Adj. R^2	0.328	0.312	0.240	0.243
GDP (baseline adj. $R^2 = 0.217$)				
$\hat{\beta}_I$	−0.320***	−0.292***	−0.139*	−0.175**
t-stat	[3.85]	[4.09]	[1.94]	[2.49]
Adj. R^2	0.299	0.286	0.229	0.239

*Significance levels: 0.10.
**Significance levels: 0.05.
***Significance levels: 0.01.
Standard errors are estimated following Newey and West (1987, 1994). Parameter estimates are standardized; absolute t-statistics are shown in brackets.

Table 9
Out-of-sample predictive accuracy.

	$h = 1$	$h = 2$	$h = 4$	$h = 8$
Employment	0.885	0.859	0.897	0.919
	[0.00]	[0.00]	[0.01]	[0.06]
Industrial production	0.906	0.916	0.938	0.967
	[0.00]	[0.00]	[0.02]	[0.13]
Investment	0.928	0.931	0.886	0.934
	[0.00]	[0.00]	[0.00]	[0.06]
GDP	0.935	0.924	0.910	0.860
	[0.00]	[0.00]	[0.00]	[0.01]

Note: Relative RMSPEs; bootstrapped p-values are reported in square brackets for the test of the null hypothesis that the ratio of the RMSPEs is equal to one.

indicators of information in the press. In the following subsection, out-of-sample exercises are described for a data-driven process designed to learn advantageous ways of summarizing information.

5.1. Out-of-sample forecasts

The methodology implemented in this subsection is the same as that implemented in the case of the indicators of uncertainty. To evaluate the information content of indices, forecast combinations based on BMA were evaluated against the baseline autoregressive model. Six indices were constructed, each one associated with a concept (e.g., fear). As in previously reported out-of-sample exercises, the alternative indices corresponded to different numbers of related words (100, 50, and 25) and different periods (30-day and 90-day windows). As a result, the BMA exercise involved 25 models. One model was the baseline model, and the remaining 24 models corresponded to the autoregressive models that incorporated one of the 24 indices reflecting information in the press.

The results are shown in Table 9. As can be seen, accuracy increased in all forecasting tasks, suggesting complementarities between indicators that proxy different subjective states. Compared to forecasts that only exploit indices summarizing manifestations of uncertainty, improvements were considerable in the case of payrolls and industrial production forecasts. A representative example is the case of one-year-ahead industrial production forecasts. In this case, the relative forecast accuracy improved from 1.004 to 0.938 when additional indicators of subjective states were incorporated. In other words, there was an improvement from a scenario with no information gain to a statistically significant 6% improvement in relative forecast accuracy.

Table 10 shows the posterior probabilities associated with the BMA exercise. Consistent with the evidence from in-sample forecasts, the BMA assigns a high posterior probability to models that incorporate indices related to fear and uncertainty. In the early stages of each forecasting exercise, optimal forecast combinations are not seen to focus on a particular type of index. The only exemption is with the employment forecasts. In this case, the sum of posterior probabilities assigned to models that incorporate an indicator of fear is 0.85. Interestingly, by the end of the sample (2016-III), almost all of the posterior probability mass is placed on models that incorporate indices that approximate manifestations of fear.

These exercises suggest that the combination of indicators capturing multiple subjective states is an advantageous strategy for business cycle forecasts. In particular, indicators of uncertainty and fear seem to provide the most valuable information. The selection of keywords was informed by previous literature and subjective judgments.

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.

Table 10
Posterior probabilities - $h = 4$.

	Employment				Industrial production		
	1990-I	2003-II	2016-III		1990-I	2003-II	2016-III
Uncertainty	0.076	0.055	0.006	Uncertainty	0.217	0.148	0.039
Fear	0.854	0.654	0.977	Fear	0.494	0.604	0.952
Pessimism	0.044	0.248	0.010	Pessimism	0.111	0.124	0.004
Anxiety	0.018	0.035	0.001	Anxiety	0.113	0.078	0.003
	Investment				GDP		
	1990-I	2003-II	2016-III		1990-I	2003-II	2016-III
Uncertainty	0.402	0.390	0.054	Uncertainty	0.254	0.250	0.223
Fear	0.421	0.533	0.945	Fear	0.446	0.519	0.730
Pessimism	0.081	0.013	0.000	Pessimism	0.113	0.094	0.009
Anxiety	0.061	0.003	0.000	Anxiety	0.125	0.084	0.013

Notes: The table shows, for each economic activity indicator and each set of subjective indicators, the sum of the posterior probabilities assigned to models that incorporate indicators of the respective subjective state. This information is shown for the beginning, the middle and the final period of the out-of-sample forecast exercises. Forecast tasks.

A systematic procedure for keyword selection would indeed be a desirable feature, though it is beyond the scope of the current study.

6. Comparison with alternative methods

This work focuses on methods that can automatically learn meaning in economic contexts using natural language processing tools that efficiently extract information from unstructured data. However, the indices are based on a small set of keywords or predefined dictionaries that contain a significant amount of information. These traditional methods reflect expert judgments regarding convenient categorizations or keywords. The relative performance of these alternative methodologies is unknown, and must be evaluated empirically.

The first evaluation of informational gains associated with the use of WVRs was based on a simple benchmark. Specifically, a simple indicator of uncertainty was used, based on the frequency of the words "uncertainty", "uncertain", and "uncertainties" (Unc-3). The information content of this indicator was compared to the information provided by the index based on the set of 100 most closely related words, as indicated by WVRs (Unc-WVR).

In addition, the information content of four traditional methods was compared to information captured by uncertainty indices that exploit WVRs. Following Tetlock (2007), multiple contributions have exploited the list of words categorized as negative in the Harvard IV dictionary.[18] To adapt this dictionary to the context, Loughran and McDonald (2011) proposed a list of words that transmit a negative tone in financial contexts.[19] In a simple yet potentially valuable approach, the monthly publication, The Economist, proposed the R-Index, a metric of the frequency with which the word "recession" is found in the economic press (Economist, 2001). Finally, an influential metric based on press content was proposed in Baker et al. (2016). The metric is known as the Economic Policy Uncertainty (EPU) index. This index computes the fraction of news articles that refer to economic policy and uncertainty.[20] These articles were identified using a small set of words.[21]

The performance of the uncertainty metric based on WVR was compared to the performance of indices associated with the previously described alternative methods. In the case of the first three alternative methods, the indices were computed using the test corpus of WSJ content used in this contribution. The EPU index was downloaded from the website created by the authors by searching for text content from a large collection of publications.

Table 11 shows the results of the in-sample forecast exercises with the five alternative indices and the index approximating uncertainty using the set of 100 words most closely related to uncertainty. The estimated parameters, p-values, and adjusted R^2s indicate that the index based on WVRs is the most informative indicator. This conclusion is valid independent of the economic activity metric under consideration.

Among the alternative indices, there was no clearly superior methodology. The index based on the Harvard-IV dictionary dominated in the case of GDP forecasts. The EPU was the most valuable indicator in the case of employment and industrial production forecasts. Nevertheless, in the case of investment forecasts, the R-word index had the strongest performance.

The absence of a clear ranking among the alternative methods implies that forecast combinations might be used to find an efficient way to incorporate the information provided by the respective indicators. Out-of-sample forecasts are generated through Bayesian model averaging. In the first exercise, the predictive ability associated with the four alternative methods was evaluated jointly through BMA. In the second exercise, alternative methods were considered jointly with uncertainty indices based on WVRs.

In the first set of exercises, nine models were considered. One model was associated with the baseline

[18] The list can be downloaded from http://www.wjh.harvard.edu/~inquirer/homecat.htm.

[19] The list can be downloaded from www3.nd.edu/~mcdonald/.

[20] More precisely, this metric of press content is one of three elements used to compute the EPU index.

[21] Details on the methodology and data can be found by visiting http://www.policyuncertainty.com/index.html.

Table 11
Estimated forecast models, $h = 1$.

	Unc-WVR	Unc-3	Harvard-IV	L&M(2011)	R-word	EPU
Employment						
$\hat{\beta}_I$	−0.263**	−0.098	−0.095	−0.132*	−0.114**	−0.182*
t-stat.	[2.24]	[1.54]	[1.34]	[1.77]	[2.26]	[1.97]
Adj. R^2	0.710	0.671	0.670	0.676	0.668	0.693
Ind. Prod.						
$\hat{\beta}_I$	−0.280*	−0.157*	−0.115	0.108*	0.076	−0.174*
t-stat.	[1.68]	[1.86]	[1.10]	[1.72]	[1.55]	[1.72]
Adj. R^2	0.414	0.369	0.357	0.356	0.350	0.373
Investment						
$\hat{\beta}_I$	−0.342***	−0.212	−0.243	−0.202	−0.326**	−0.237**
t-stat.	[2.96]	[1.59]	[1.59]	[1.42]	[2.57]	[2.11]
Adj. R^2	0.328	0.257	0.265	0.252	0.293	0.271
GDP						
$\hat{\beta}_I$	−0.320***	−0.228***	−0.302**	−0.263**	−0.263*	0.188**
t-stat	[3.85]	[2.95]	[2.24]	[2.51]	[1.96]	[2.18]
Adj. R^2	0.299	0.255	0.278	0.264	0.255	0.241

*Significance levels: 0.10.
**Significance levels: 0.05.
***Significance levels: 0.01.
Standard errors are estimated following Newey and West (1987, 1994). Parameter estimates are standardized; absolute t-statistics are shown in brackets.

Table 12
Out-of-sample predictive accuracy — Combination of alternative indicators.

[A] Alternative indicators				
	$h = 1$	$h = 2$	$h = 4$	$h = 8$
Employment	0.950 [0.00]	0.952 [0.01]	0.991 [0.20]	0.975 [0.18]
Industrial production	0.986 [0.01]	1.015 [0.68]	1.044 [0.47]	1.009 [0.54]
Investment	0.992 [0.31]	1.003 [0.88]	0.989 [0.42]	1.019 [0.54]
GDP	0.980 [0.01]	0.987 [0.14]	1.032 [0.19]	0.954 [0.11]
[B] WVR + Alternative indicators				
	$h = 1$	$h = 2$	$h = 4$	$h = 8$
Employment	0.924 [0.00]	0.913 [0.00]	0.982 [0.15]	0.990 [0.30]
Industrial production	0.953 [0.00]	0.982 [0.03]	1.037 [0.58]	1.009 [0.54]
Investment	0.946 [0.00]	0.944 [0.00]	0.914 [0.00]	0.953 [0.08]
GDP	0.944 [0.00]	0.951 [0.00]	0.960 [0.03]	0.902 [0.02]

Note: Relative RMSPEs; bootstrapped p-values are reported in square brackets for the test of the null hypothesis that the ratio of the RMSPEs is equal to one.

autoregressive specification. In addition, for each alternative method, indices were built using 30-day and 90-day lagged windows. In this way, eight additional models were added to the baseline specification. The results are shown in Panel A of Table 12. Compared to the baseline specification, gains in the forecasting accuracy were observed in the case of the shortest forecast horizon. For longer forecast horizons, no significant gain in accuracy was observed. Additionally, these alternative specifications did not match the forecasting performance observed in the case of the forecasts that exploited WVRs. For example, in the case of one-year-ahead GDP forecasts, forecasts based on the alternative indices generated a metric of accuracy that was 3.2% worse than that of the baseline model. By contrast, the uncertainty indices informed by WVRs generated forecasts that were significantly better than the baseline forecasts. In this case, the metric of accuracy improved by 7.5%.

Whereas the alternative indices did not perform especially well, they captured information that might be advantageously used in combination. In other words, the indices proposed in this work might be complemented by alternative text summarizing techniques. To evaluate this hypothesis, the indices that exploit WVRs to proxy manifestations of uncertainty were combined with alternative indices through BMA exercises. As in the previous sections, alternative specifications associated with number of words and time windows resulted in six indices of uncertainty. These forecast combination exercises involved identifying the weights assigned to 15 models.

However, the results, shown in Panel B of Table 12, suggest that there is no advantage to incorporating alternative indicators. Indeed, the accuracy, as indicated by relative RMSPEs, was worse than that observed when the uncertainty metrics based on WVRs were the only indicator of information in the press (see Table 5).

7. Conclusions

This study proposed a method of quantifying unstructured press information. The proposed method is based

on WVR, a tool developed in the field of natural language processing. The study showed that trained representations can learn meaningful relationships between words in economic contexts. These associations are exploited to build indicators of uncertainty and other subjective states in press content.

Using real-time data on economic activity, the indices were shown to capture valuable information. Implementations of in-sample and out-of-sample forecast exercises showed that indicators of uncertainty anticipate business cycle dynamics. BMA implementations illustrated the benefits associated with combining information from indices linked to different subjective states. Their information content compared favorably with that resulting from dictionary-based word counting techniques proposed in the literature. In this way, novel machine learning tools can generate interpretable and informative indicators that can be used for macroeconomic analysis.

There are several directions in which the current work can be extended. A natural path is associated with implementations based on larger training and testing corpora. Whereas larger collections of text do not necessarily include more information, a careful selection of additions to the corpus can result in more precise indicators. As previously indicated, automated methods of selecting relevant subjective features in unstructured data can also be explored. Finally, in the field of natural language processing, WVRs are used as inputs in neural network applications (Kim, 2014). Hence, while the property of straightforward interpretation would be lost, another possible extension involves exploring the performance of nonlinear forecasting models.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2019.12.001.

References

Alexopoulos, M., & Cohen, J. (2015). The power of print: Uncertainty shocks, markets, and the economy. *International Review of Economics & Finance*, *40*, 8–28.

Allan G., T. (2006). Forecast combinations. In G. Elliott, C. W.J.G.ranger, & A. G.T.immermann (Eds.), *Handbook of economic forecasting* (pp. 135–196). Amsterdam, Netherlands: Elsevier.

Aromí, J. D. (2017). Conventional views and asset prices: What to expect after times of extreme opinions. *Journal of Applied Economics*, *20*(1), 49–73.

Aromí, J. D. (2018). GDP Growth forecasts and information flows: Is there evidence of overreactions? *International Finance*, *21*(2), 122–139.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, *131*(4), 1593–1636.

Balke, N. S., Fulmer, M., & Zhang, R. (2017). Incorporating the beige book into a quantitative index of economic activity. *Journal of Forecasting*, *36*(5), 497–514.

Boucher, J., & Osgood, C. E. (1969). The pollyanna hypothesis. *Journal of Memory and Language*, *8*(1), 1–8.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, *12*(Aug), 2493–2537.

Da, Z., Engelberg, J., & Gao, P. (2014). The sum of all FEARS investor sentiment and asset prices. *Review of Financial Studies*, *28*(1), 1–32.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391–407.

Delis, M. D., Kouretas, G. P., & Tsoumas, C. (2014). Anxious periods and bank lending. *Journal of Banking & Finance*, *38*, 1–13.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, *12*(Jul), 2121–2159.

Economist (2001). The r-word. URL: https://www.economist.com/finance-and-economics/2001/04/05/the-r-word.

Faust, J., Gilchrist, S., Wright, J. H., & Zakrajšek, E. (2013). Credit spreads as predictors of real-time economic activity: a Bayesian model-averaging approach. *The Review of Economics and Statistics*, *95*(5), 1501–1519.

Fernandez, C., Ley, E., & Steel, M. F. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, *16*(5), 563–576.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance*, *68*(3), 1267–1300.

Garcia, D., Garas, A., & Schweitzer, F. (2012). Positive words carry less information than negative words. *EPJ Data Science*, *1*(1), 1–12.

Hansen, S., McMahon, M., & Prat, A. (2017). Transparency and deliberation within the FOMC: a computational linguistics approach. *Quarterly Journal of Economics*, *133*(2), 801–870.

Jurado, K., Ludvigson, S. C., & Ng, S. (2015). Measuring uncertainty. *American Economic Review*, *105*(3), 1177–1216.

Kim, Y. (2014). Convolutional neural networks for sentence classification. ArXiv preprint arXiv:1408.5882.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35–65.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–708.

Newey, W. K., & West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, *61*, 631–653.

Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., & Smith, R. (2018). *News and narratives in financial systems: exploiting big data for systemic risk assessment. Bank of England, Working paper no. 704.*

Pennington, J., Socher, R., & 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).

Rossi, B., & Sekhposyan, T. (2015). Macroeconomic uncertainty indices based on nowcast and forecast error distributions. *American Economic Review*, *105*(5), 650–655.

Selivanov, D., & Wang, Q. (2016). Text2vec: Modern text mining framework for r. In *Computer software manual (R package version 0.4. 0)*. Retrieved from https://CRAN.R-project.org/package=text2vec.

Stekler, H., & Symington, H. (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006–2010. *International Journal of Forecasting*, *32*(2), 559–570.

Tetlock, P. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel, & A. Zellner (Eds.), *Bayesian inference and decision techniques* (pp. 233–243). Amsterdam, Netherlands: North-Holland.

Please cite this article as: J.D. Aromi, Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2019.12.001.