DRAFT

Making text count:

economic forecasting using newspaper text*

Eleni Kalamara† Arthur Turrell‡ Chris Redl§ George Kapetanios¶ Sujit Kapadia‖

November 8, 2019

Abstract

We consider the best way to extract timely signals from newspaper text and use them to forecast macroeconomic variables, using three popular UK newspapers that span the political spectrum. We find that newspaper text can improve economic forecasts both in absolute and marginal terms. We introduce a powerful new method of making text count in forecasts that combines counts of terms with sophisticated supervised machine learning techniques. This method improves forecasts of macroeconomic variables including GDP, CPI, and unemployment, including compared to existing text-based methods. Forecast improvements occur when it matters most, during periods of stress. While we find that simple metrics go a long way in extracting signal, supervised machine learning methods go further and are likely to be transferable to other text analysis problems.

Keywords: text, forecasting, machine learning

JEL Codes: J6, J42, C55

*First version: September 10, 2019. This version: November 8, 2019. The views in this work are those of the authors and do not represent the views of the Bank of England or the European Central Bank. We are grateful to David Bholat, Paul Robinson, Misa Tanaka, and to conference participants at the Federal Reserve Board of Governors, the European Central Bank, the Bank of England, Danmarks Nationalbank, the 2019 European Economic Association meeting, the 2018 Royal Economic Society meeting, the 2019 Economic Statistics Centre of Excellence meeting, and the European Commission for suggestions. David Bradnum provided outstanding technical assistance throughout.

†King's College London. Email: eleni.kalamara@kcl.ac.uk

‡Bank of England. Email: arthur.turrell@bankofengland.co.uk

§Bank of England. Email: chrisredl@bankofengland.co.uk

¶King's College London. Email:george.kapetanios@kcl.ac.uk

‖European Central Bank. Email: Sujit.Kapadia@ecb.europa.eu

1 Introduction

As Arthur Miller wrote, "A good newspaper, I suppose, is a nation talking to itself". In this paper we show that newspapers say a lot about a nation's near-term economic future, especially during periods of stress. We also show that the best way to obtain that information is with highly non-linear text analysis methods that combine feature engineering and machine learning. The robustness and agility of the methods that do this are likely to have wide applicability. Our results demonstrate that newspapers offer policymakers a way to obtain high frequency signals about the real economy that can, potentially, improve decision making.

This paper addresses the question of how best to use contemporaneous newspaper text data to inform policymaking decisions. We do this by using a range of existing and novel methods of turning text into time series to both extract forward looking economic indicators from text and use newspaper text for economic forecasting. Our data are three UK daily newspapers with high circulation that span the political spectrum of UK mass media and cover a time period from 1990 to 2019.

We find that text significantly improves forecasts of macroeconomic variables, including GDP, CPI, and unemployment, relative to widely used benchmarks. This is especially true during periods of stress, suggesting that newspaper text could speak to recession prediction and is a strong complement to high frequency financial market data and to more expensive, and often less timely, survey data. In terms of what methods of turning text into time series work best, we show that even relatively simple transformed counts of words perform surprisingly well. Of existing methods, a dictionary of words associated with financial stability offers the best all-round performance. However, the method we introduce for forecasting, a combination of a large space of regressors based on text and supervised machine learning, far outperforms all other methods. Of the range of machine learning methods we compare, we find that neural networks consistently perform the best across text sources, horizons, and target variables.

Several other papers have explored the link between text and economic activity (Gentzkow, Kelly and Taddy, 2017), for instance in the case of firms' annual reports and their returns (Jegadeesh and Wu, 2013; Loughran and McDonald, 2011, 2013), and between newspaper text and levels of uncertainty (Alexopoulos and Cohen, 2015; Baker, Bloom and Davis, 2016). We find that newspaper text contains stronger signals of economic sentiment than economic uncertainty, although text-based measures of uncertainty have received far greater attention to date. A particularly relevant example of uncertainty

and text, due to its use of supervised machine learning, is that of Manela and Moreira (2017), who retrospectively forecast the VIX based on font-page articles in *The Wall Street Journal*. There is evidence that text is more strongly linked to financial market activity during periods of stress. Nyman et al. (2018) show that text-based measures of excitement rose substantially before the financial crisis and note that they may be an important warning sign of impending financial system distress. The uncertainty indicator of Manela and Moreira (2017), which is a news implied volatility, peaks in financial crises and rises just before transitions into economic disasters. Garcia (2013) shows that news-derived sentiment can affect asset prices, and that the effect is particularly strong during recessions. We find that text adds most value to macroeconomic forecasting during periods of stress.

The closest paper to ours is Shapiro, Sudhof and Wilson (2018), which looks at the ability of a number of dictionary (or lexicon) based sentiment analysis methods to predict the same 5 classifications (very negative to very positive) as human subjects on 800 newspaper articles. Although they acknowledge that machine learning may have advantages, their training set is small and therefore, as they note, not amenable to a supervised machine learning approaches. They also construct a sentiment index and show that, like those extracted from consumer confidence surveys, positive news-based sentiment shocks are associated with increases in consumption, output, and interest rates. Here, we assess the performance of a broader range of methods (dictionary, Boolean, algorithms from the computer science literature, and supervised machine learning) covering both sentiment and uncertainty. Our sample size is large enough for machine learning to be effective and so we are able to use text to forecast the (continuous) economic variables that policymakers are most concerned with.

Papers that have used text for forecasting include Antweiler and Frank (2004); Tetlock (2007) in the context of financial markets and firms. For forecasting macroeconomic activity, Thorsrud (2018) and Larsen and Thorsrud (2019) employ Latent Dirichlet Allocation (LDA) to create a one-off set of machine-generated topics that are then used to choose representative articles for sentiment analysis (via the Harvard IV-4 Psychological Dictionary). 80 different sentiment-by-topic normalised time series are produced and used to create nowcasts that are broadly competitive with those based on expert judgement or a model combination framework. Similarly, Ardia, Bluteau and Boudt (2019) use a combination of expert opinion and tags from a data provider to create a one-off set of topics that filter a corpus of articles to those relevant to a forecast target (classified by topic). A range of dictionary (or lexicon) based methods are then used to obtain sentiment scores for each topic-labelled text. The different sentiment indicators are fed into a penalised linear regression (elastic net) to form a single,

combined forecast that beats models not augmented with text at horizons greater than 9 months ahead.

While the results of nowcasting and forecasting the real economy with topic models are very promising, they have a number of inherent limitations. As Thorsrud (2018) points out, recursive updating of the topic model is computationally infeasible and introduces an identification problem: even in dynamic topic models (Blei and Lafferty, 2006), the same topics cannot be guaranteed to appear, or to be linked, when the model is re-estimated. The results of topic models are also sensitive to the choice of the number of topics (see Turrell et al. (2018) for a discussion). We restrict the methods that we compare or introduce to those that can be reasonably re-estimated during an out-of-sample forecasting exercise as this is closest to how the methods would be used in policy-making institutions.

We make several key contributions relative to the existing literature. What policymakers and practitioners need to know is how to get the best out of text for economic forecasting and we attempt to answer this in the context of newspapers. Our forecast environment is what is used in practice in policy-making institutions: a rolling window re-estimation with direct h-step ahead out-of-sample forecasts.

We compare many existing methods from the literature, including popular Boolean and dictionary-based methods, but also introduce our own. We also go beyond these methods and use supervised[1] machine learning methods that, combined with text feature engineering, completely dominate the other approaches in terms of forecast performance. This approach is also much more likely to be transferable to other applications. Combining text analysis and forecasting presents a number of subtle pitfalls related to information leakage, and we comment on how to avoid these. It is worth emphasising that our proposed approach of retaining a large number of terms from text, turning each into a time series and then applying sophisticated nonlinear machine learning methods, appropriate for high dimensional datasets, to produce predictions is, to the best of our knowledge, novel in the text analysis literature and since it also produces the best results, is by far our major contribution.

While it has been established that text can provide useful forecasts in financial markets, there has been less focus on the real economy, to the news sources of consumers as opposed to investors, or to what is of most use in policy. Relatedly, we show that the improvement to forecasts with text-based measures is also true for the real economy and is a general property of text applied to economic forecasting rather than deriving specifically from looking at either sentiment or uncertainty.

The rest of the paper is organised as follows: we describe our newspaper text data in §2; we discuss

[1]While topic modelling is a type of unsupervised machine learning and looks for patterns within inputs, supervised machine learning is more analogous to regression: it looks for a specific pattern between inputs and outputs.

the different methods to turn text into time series in §3, beginning with our discussion of the pitfalls of using text data in real-time is in §3.1 and then defining the simpler algorithm-based text metrics in §3.2 and the more sophisticated machine learning based measures in §3.3. In §4 we look at whether the simpler algorithm-based text metrics can function as indicators by comparing them to a suite of existing indicators used by policymakers. §6 uses the algorithm-based text metrics in forecast exercises, while §7 looks at the forecast performance of the machine learning based approach forecasting. A discussion of the results may be found in §8.

2 Data

Our data are from Dow Jones Factiva. Newspaper articles are retrieved through an application programming interface (API) which filters for the subjects Commodity/Financial Market News, Corporate/Industrial News, and Economic News. The allowed article types included editorials and commentaries/opinions. We discarded any articles that were updates of previous articles on the basis that the most salient information, if newsworthy, would have been in the first release. We also discarded any articles that had exactly the same text content as another article, keeping only the first occurrence of such articles, and removed any remaining duplicates through string matching. Such articles are not uncommon in this corpus (Eckley, 2015). Descriptive summary statistics of the newspapers are shown in Table 1, which shows the number of unique articles after de-duplication. The circulations shown in the table are for June 2018 (Newsworks, 2018).

	Circulation/10^3	Unique articles	% of total	⟨articles/month⟩	First article	Last article
The Guardian	138	266,757	54.5	764	1990-01-06	2019-01-23
The Daily Mirror	563	129,210	26.4	450	1995-03-01	2019-01-23
The Daily Mail	1,265	93,281	19.1	267	1990-01-11	2019-01-23
Total	1,966	489,248	100.0	1,482	-	-

Table 1: Descriptive statistics of articles from selected UK newspapers. Source: Dow Jones Factiva.

3 Turning text into time series

For the methods that we use, the text of each newspaper article must be cleaned before being transformed into numbers. We remove punctuation and digits, enforce lower case, and remove a large number of stopwords.[2] We use two approaches to turn cleaned text into quantitative time series that are then used as inputs into forecasts: algorithm-based text metrics and term frequency vectors. Throughout,

[2]Words that are not by themselves informative, typically conjunctions such as 'and'.

we refer to terms rather than words, as these are more flexible. A term could be composed of one word, two words[3], e.g. 'bank run', or the stem of a word, e.g. 'econom' for 'economics' and 'economy'.[4]

We only include those methods from the existing literature that can be computed in real time, including being re-estimated at every time step. Where necessary, we have also modified existing methods to not include information about the future, a phenomenon known as information leakage or (in this context) as look-ahead bias. We now turn to a more detailed discussion of how these can arise with text analysis, and how they may be avoided.

3.1 Avoiding information leakage with text-based time series

The simplest example of information leakage with time series is when a continuous time series variable is normalised, i.e. $x_t \longrightarrow \frac{x_t - \mu_x}{\sigma_x}$ where the mean, μ_x, and standard deviation, σ_x, take $\{x_t\}_{t=0}^{t=T}$ as their domain. In real time, at time t, information on times $> t$ is not available and so the transform should be time-dependent, i.e.

$$x_t \longrightarrow \frac{x_t - \mu\left(\{x_{t'}\}_0^t\right)}{\sigma\left(\{x_{t'}\}_0^t\right)}$$

This example may be trivial, but there are subtle ways for information leakage to occur with text.

The most common occurs during pre-processing of text. It is usually undesirable to track every single possible term or combination of terms in a corpus. Usually, a decision is made to omit certain words from analysis. In a static analysis of text it is typical - for practical reasons - for words that occur very frequently or very infrequently in the corpus to be omitted, usually by specifying both a minimum and maximum threshold frequency. This would omit frequent but uninformative words such as 'the' as well as words that are so rare as to be statistically irrelevant. Only words that have middling frequencies are included in the analysis. But threshold frequencies assume and require knowledge of all words in the corpus, which isn't possible in real-time. Terms that suddenly appear at one point in time can be correlated with macroeconomic developments but they may only be tracked because they began to appear at a certain point in time. A good example would be the term 'sub-prime' that might pass a whole-corpus threshold filter but would be far less likely to pass the same filter applied only to text from before 2007. Tracking such a word might appear to produce very strong results that would not have been possible in real time. To avoid this, we explain in §3.3 how we do not use the corpus of newspaper text itself to determine which terms to track from the newspaper text.

Although we exclude topic models for other reasons, they can also be susceptible to this problem

[3]This is known as a 2-gram, and a phrase of length N as an N-gram.
[4]More details of text cleaning may be found in Appendix A.

– the entire corpus being used to train a topic model that is subsequently used to make out-of-sample predictions on a future that the topic model has extracted information from.

Dictionary, or lexicon, and Boolean methods are all also susceptible to the benefits of hindsight. As an example, a Boolean method that tracks the terms 'dodd-frank' and 'bank stress test' would be particularly suited to picking up macroeconomic events that co-occur with the appearance of those terms but this would not have been known in real time.

Just as typical time series can undergo global transforms that should account for time-dependent means and standard deviations, so too can text based transformations. The most common, and most pertinent to our work, is the term frequency – inverse document frequency transform that we define in §3.2. Used naively, this sees terms from the whole corpus as part of the inverse frequency weighting. We explain in the subsequent sections how we modify this commonly used transformation in order to remove this channel for information leakage.

Finally, when we use machine learning we train (do in-sample estimation of a model) on the same text source as that which will be used in (out-of-sample) testing. This ensures that the text based features used by the machine learning models have as similar distributions of features in test and train sets as possible.

3.2 Algorithm-based text metrics

Algorithm-based text metrics are the product of pre-defined rules, or algorithms, that turn text into numbers. They are by far the most commonly used method to extract information from text. The simplest example that we use in this paper is the number of times a specific term appears in each article divided by the number of words in the article. The numerical scores for a particular month are found from the mean of the scores of the articles that were published in that month.

The set of algorithms we use to create text metrics is summarised in Table 2. They fall into three broad categories (see Appendix B for formal definitions of each). Dictionary methods typically associate specific terms with specific scores (positive or negative for sentiment) and count the net score per article. A variation of this is a measure that just counts a single term and weights it by article length, as with the single term counts of "uncertain" and "econom". We also use a more sophisticated weighting, the term frequency – inverse document frequency (tf-idf). This seeks to control for the frequency of the term in each article ($\text{tf}(a)_w$), the number of articles per day (N_t), and the number of articles in which the term appears per day ($n_t < N_t$). It has commonly been used with an inverse document frequency that applies across the whole corpus, we use N_t and n_t to avoid information leakage. It uses a log

transform, partly mindful of the power law for the frequency of different terms in the English language (Zipf, 1950):

$$\text{tf-idf}(a)_t = \frac{\ln\left(1 + \text{tf}(a)_w\right)}{\ln\left(1 + N_t/n_t\right)}$$

Positive and negative dictionary	**Boolean**	**Computer science-based**
Financial stability (Correa et al., 2017)	Economic Uncertainty (Alexopoulos, Cohen et al., 2009)	VADER sentiment (Gilbert, 2014)
Finance oriented (Loughran and McDonald, 2013)	Monetary policy uncertainty (Husted, Rogers and Sun, 2017)	'Opinion' sentiment (Hu et al., 2017; Hu and Liu, 2004)
Afinn sentiment (Nielsen, 2011)	Economic Policy Uncertainty (Baker, Bloom and Davis, 2016)	Punctuation economy (this paper)
Harvard IV (used in Tetlock (2007))		
Anxiety-excitement (Nyman et al., 2018)		
Single word counts of "uncertain" and "econom"		
tf-idf applied to "uncertain" and "econom"		

Table 2: The three broad categories of algorithm-based text metrics used.

Boolean methods provide a count of articles only if the terms in an article satisfy some logical condition, for instance that three distinct and pre-defined terms all appear in the same article. In the most simple case, this just counts any article that contains a specific term. The most notable examples of Boolean methods are the Economic Uncertainty index of Alexopoulos, Cohen et al. (2009) and the similar UK version of the Economic Policy Uncertainty (EPU) index of Baker, Bloom and Davis (2016). However, note that while we apply the text analysis methodology of the UK EPU index, Baker, Bloom and Davis (2016)'s paper uses *The Times* and *The Financial Times*, different publications to ours, and they include all articles, not just those about economic developments.

The third type of metric used in this paper draws on the computer science literature. Two of the metrics that we implement are from previous works; the VADER metric (Gilbert, 2014) is rule-based and designed for sentiment as expressed on social media while the opinion sentiment metric (Hu et al., 2017; Hu and Liu, 2004) combines machine learning and product reviews to develop a dictionary-based method. To these we added a new metric that tries to measure the sentiment within individual sentence fragments if and only if those sentences mention a particular term, in this case the term 'econom'. Before texts are split into fragments, coreference resolution picks up any indirect references to the term in question. Once sentence fragments are isolated, the sentiment is computed with a combination of other dictionary methods. More details of all of the computer science based methods may be found in Appendix B.1.4.

In Table 3, we show the scores produced by some of the algorithms for example articles. There are examples from each of the three types of metric shown in Table 2. The first piece of text is taken

from a Bank of England *Inflation Report* and, according to the metrics, is positive in sentiment.[5] The second is fictional and designed to encapsulate high uncertainty and negative sentiment. Note that only the second text entry triggers the Boolean Alexopoulos metric, because that text contains the word 'uncertainty' and 'economy'. The third and fourth text examples are very similar, but with 'bad' replaced by its antonym, 'good', and, in consequence, almost reversed sentiment scores.

Text	TFIDF economy	Vader	Counts economy	Alexopoulos	Stability
Global GDP growth picked up during 2016 and has been strong over the past year (Section 1.1). Weighted by countries' shares of UK exports, global growth is estimated to have remained at 0.8% in 2017 Q4. That pace of growth is expected to persist in the near term, above expectations in November. Survey indicators of output (Chart 1.1) and new orders remain robust, particularly in the euro area and United States. Measures of business and consumer confidence are also healthy...	-0.00	0.97	0	0	0.03
The economy has struggled and is in a bad state with disappointing performance, unhappy consumers, low confidence with high uncertainty. Policy faces a number of risks which could transmit to the real economy, and pundits are increasingly concerned about a crash.	-0.15	-0.93	-2	1	-0.11
The current direction of policy is very bad.	-0.00	-0.54	0	0	-0.25
The current direction of policy is very good.	-0.00	0.44	0	0	0.25

Table 3: Selected algorithm-based text metrics applied to example text. In the interests of space, the first text example is truncated. Note that we have given pre-factors of -1 or 1 to some metrics so that positive sentiment has a positive score and heightened uncertainty has a positive score. Negative signs before zero indicate that the scores were more than -0.01 but less than zero.

3.3 Machine learning methods

Here we describe our alternative method of employing text for economic forecasting. The method does not create time series that function as indicators, unlike the algorithm-based text metrics, but is well-suited to forecasting with text. There are two motivating principles to the alternative approach to forecasting with text metrics that we take here: i) we wish to extract as much of the rich information available in the text as possible and ii) we want to allow a model to decide which terms to put weight on in real-time, rather than fixing this ahead of time. These two principles are carried out in two steps, feature engineering, and supervised machine learning.

The former step creates a large set of features (in the language of machine learning) or regressors (in the language of econometrics) to use as the inputs to a machine learning algorithm that can operate with a greater number of features than observations. This large feature space allows for a broader set

[5] In the interest of space only part of the text is shown in the table.

of the information in the text to be captured. The feature engineering that we choose represents each article as a term frequency vector. Term frequency vectors extend the idea of counting terms to a large number of terms. Here we use 9660 terms, with up to 3-grams. These terms are taken from the union of words from several sentiment dictionaries combined with common terms taken from a dictionary of economic terms – see Appendix B for more details. The dictionaries that we use are drawn from other studies and are independent of our corpus. Because of this, some of the terms never appear in our corpus. To use term frequency vectors as inputs into forecasts, each article is represented as a vector (one dimension for each term) of counts of terms that occur within it. Each vector may be denoted

$$\overrightarrow{\mathrm{tf}(a)} = (\mathrm{tf}(a)_{w_1}, \mathrm{tf}(a)_{w_2}, \cdots)$$

The term frequency vector for a month is the mean of the vectors of the articles published in that month.

In the second step, we use supervised machine learning models to automatically decide which of this large set of terms (or combinations of terms) to put weight on by using the term frequency vectors as features (regressors). In the case of forecasts with dictionary-based text metrics, a pre-determined set of weights are effectively applied to terms to create a net score, and then, when used in forecasts, a regression model decides what overall weight to put on that aggregate net score. In this case, the weights on individual terms are set directly by the supervised machine learning model, a more flexible solution. In general, this is likely to produce better predictions than specifying some of the weights in advance.

The machine learning stage uses the term frequency vectors to predict a target variable y at time $t + h$:

$$\hat{y}_{t+h} = f_{\mathrm{ML}}\left(\ldots, \overrightarrow{\mathrm{tf}_t}\right)$$

where f_{ML} is the function obtained through machine learning. We use a number of machine learning algorithms detailed in Appendix H; lasso regression, ridge regression, elastic net regression, support vector machine regression (SVM), random forests, and artificial neural networks (NN).

4 Algorithm-based text metrics as proxies

We now turn to our first set of results and ask whether algorithm-based text metrics can be used[6] as plausible forward looking indicators and, if so, which are the most effective? We separate analysis into those text metrics that proxy either sentiment or uncertainty.

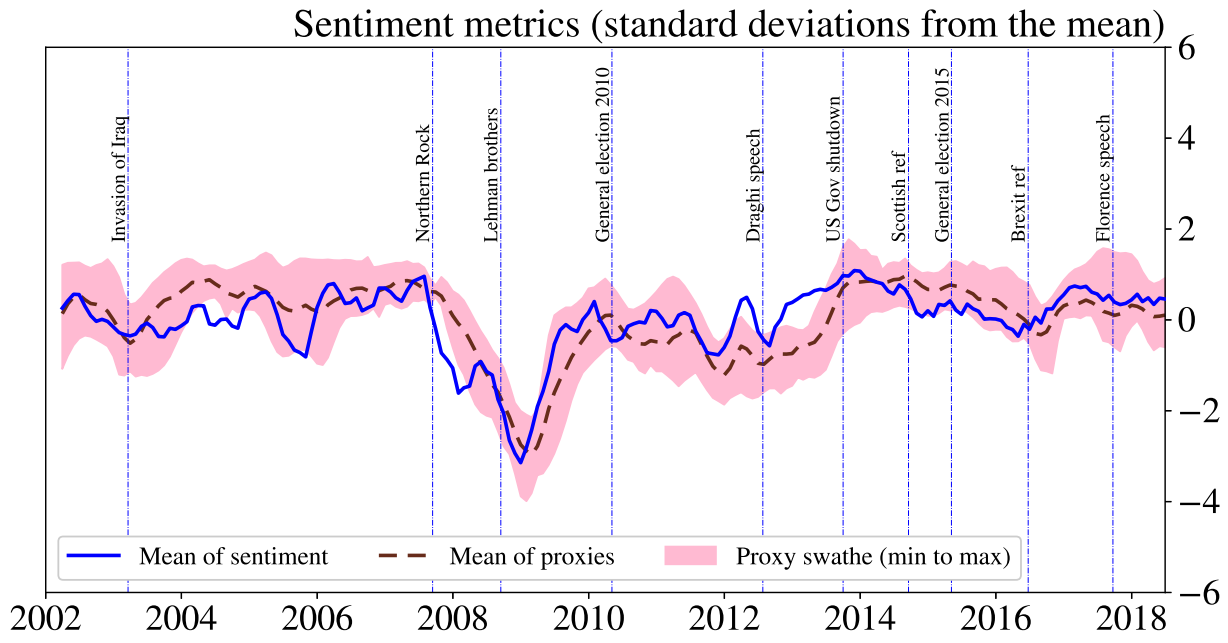

Figure 1: Three month rolling mean of the macroeconomic sentiment text metrics created from the text of *The Daily Mail* (solid line) plotted against the three month rolling mean of the proxies for macroeconomic sentiment (broken line) and a swathe defined by the maximum and minimum values across proxies at each point in time.

We compare the series to existing time series that proxy for sentiment and uncertainty. The proxies are chosen as being representative of the indicators policymakers might currently use to understand each of these. Full descriptions of the proxies may be found in Table D.3 of Appendix D. To give an overall view of the effectiveness of newspaper text based indicators as proxies, we plot the average of all text metrics over time against a swathe from the existing numerical proxies from Table D.3. All text series are aggregated to monthly frequency using a 3 month rolling mean. In the interests of space, we show only two example swathe plots, for *The Daily Mail* for sentiment and for *The Guardian* for uncertainty. These are shown because they have features of interest. In each plot, Fig. 1 for sentiment and Fig. 2 for uncertainty, the solid line shows the mean of all of either the algorithm-based sentiment

[6]We run an augmented Dickey-Fuller test for stationarity in Appendix B.2

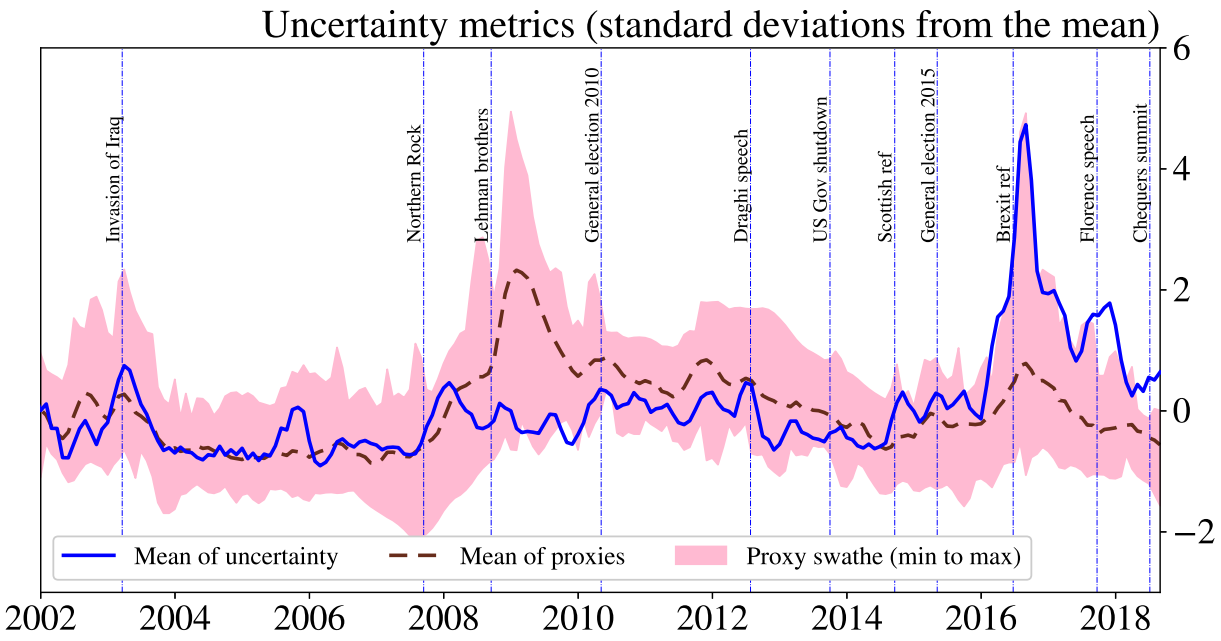

Figure 2: Three month rolling mean of the macroeconomic uncertainty text metrics created from the text of *The Guardian* (solid line) plotted against the three month rolling mean of the proxies for macroeconomic uncertainty (broken line) and a swathe defined by the maximum and minimum values across proxies at each point in time. The very large increase in uncertainty during 2016 precedes the UK's referendum on whether to leave the European Union.

or uncertainty text metrics.

Fig. 1 shows that this broad measure of sentiment taken by averaging the text metrics has a striking qualitative correlation to the swathe of proxies. Of particular note is the sharp deterioration of sentiment that slightly leads, and then tracks, the financial crisis. The leading nature of the text-based sentiment proxy is seen during the recovery too. There are periods when the sentiment indicator diverges from the mean of other indicators substantially.

Fig. 2, showing uncertainty, reflects uncertainty proxies less well, especially during the financial crisis, but still exhibits a surprisingly strong relationship between text and proxies. Overall, the uncertainty measures based on text put more weight on events that are UK-specific. For example, they respond more strongly to the invasion of Iraq, Northern Rock, then British Prime Minister Theresa May's Florence Speech (in which the UK's Brexit policy positions of the time were set out), and public votes within the UK. The lack of a strong increase in uncertainty during the crisis is consistent with other text based measures of uncertainty, and with the non-Euro area uncertainty index of Mumtaz and Musso (2018). The sensitivity to political events causes a disconnect from other measures of un-

certainty at the end of the series. In contrast, the uncertainty time series based on *The Daily Mail* (not shown) remains much more in step with other proxies for uncertainty. It should be noted that different newspapers have different editorial stances on Britain's exit from the EU.

We also look at the correlation of each text metric in turn against proxies 3 months ahead. We use the mean of each text metric across the three newspapers. These heatmaps may be seen in Figures 3 and 4, representing sentiment and uncertainty respectively. For both heatmaps, the structure of the correlations persist 6 to 9 months ahead, becoming slightly weaker as the horizon increases. The correlations for sentiment are substantially stronger than for uncertainty, and the sentiment measures are best correlated to macroeconomic sentiment proxies. A similar result is shown by Kozeniauskas, Orlik and Veldkamp (2018) who document the weak correlations across a wide range of uncertainty proxies used in the literature. Note that the investment grade corporate bond spread could be considered to contain signals of both uncertainty and sentiment, and so we include it in both heatmaps. But the sign of the correlation should be negative for (positive) sentiment. The EPU UK index shares the same method as the Baker-Bloom-Davis text metric but the former is constructed from a different set of newspapers.[7]

For macroeconomic sentiment, the correlations between the text metrics and the business confidence measure of the OECD are highest, and the most highly correlated text metrics are Stability and TFIDF (term frequency - inverse document frequency) economy. For uncertainty, the measure using the method from Alexopoulos, Cohen et al. (2009) and the similar measure from Baker, Bloom and Davis (2016) are also highly correlated. The measure of Husted, Rogers and Sun (2017) measures monetary policy uncertainty specifically and this is likely behind its lower levels of correlation with the more general uncertainty metrics. This suggests that counts of the word uncertainty are providing most of the power of the indicator. In general, the correlation between the text metrics and the proxies is appreciable and of the expected sign, but there are also a number of weak correlations.

We also ask whether our sentiment or uncertainty text metrics Granger cause any of their relevant proxies and vice versa. The results are in Tables D.4 and D.5 of Appendix D.1. Table D.4 shows that the uncertainty metrics tend to Granger cause the EPU UK index and the UK version of the macroeconomic uncertainty indicator of Jurado, Ludvigson and Ng (2015) from Redl (2018), but not the equivalent financial uncertainty indicator. The simplest uncertainty metric, counting the stem

[7]While Baker, Bloom and Davis (2016)'s paper uses all articles from *The Times* and *The Financial Times*, different publications to ours, the time series available on their website that we use here is based on "about 650 U.K. newspapers, ranging from large national papers like the Guardian to small local newspapers across the United Kingdom."

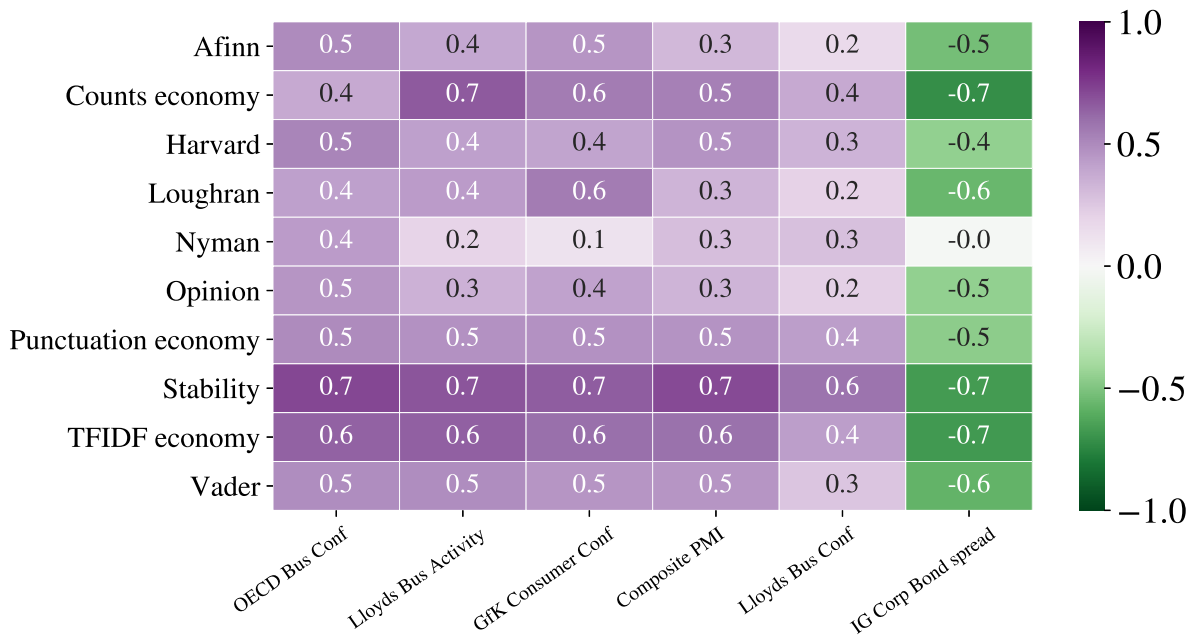

Figure 3: Heatmap of correlations between text metrics, averaged over newspapers, and proxies for macroeconomic sentiment at a three month horizon. Full definitions of the proxies may be found in Table D.3 of Appendix D.

of the word uncertainty, also Granger causes the investment grade corporate bond spread. For the sentiment metrics, the Stability metric, counts of the stem of the word economy, and TFIDF economy all Granger cause a large number of proxies. The Stability metric is the strongest performer overall as it Granger causes a large number of proxies at the 1% significance level, is strongly stationary, and has the highest average correlations with proxies. This is unexpected as the Stability metric is a dictionary designed for a financial stability context, specifically the *Financial Stability Reports* of many countries' central banks. It is not designed for use with the text of newspapers aimed at the general public. Yet, many of its words could plausibly be used to describe the economy in newspapers, for instance 'rebounding', 'sluggishness', and 'over-heated'. In general, the performance of macroeconomic or financial uncertainty is weaker, and much more mixed across the analysis.

In Appendix E we run (in-sample) regressions of GDP on both types of indicator (sentiment and uncertainty) and find that all but two text metrics are highly significant (1%) as confounders and most substantially increase the goodness-of-fit beyond a baseline AR(1) regression with no text included (and fixed effects for newspapers). However, the F-statistic of many of the models with text is lower than the model without except in the case of TFIDF economy and Stability.

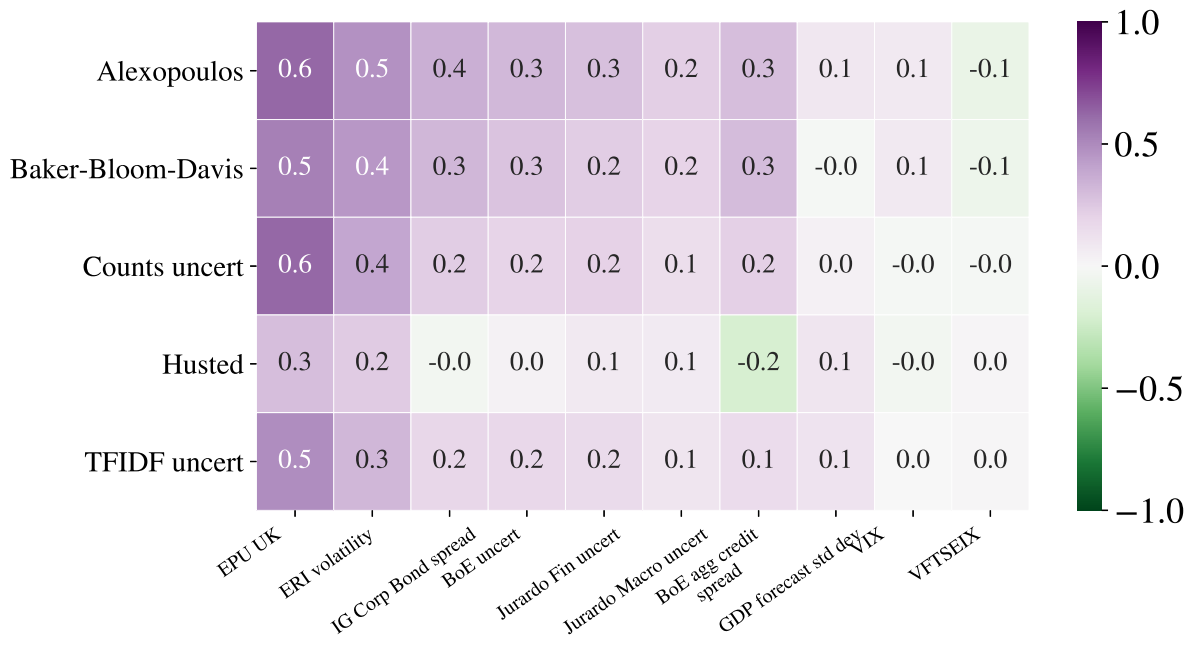

Figure 4: Heatmap of correlations between text metrics, averaged over newspapers, and proxies for financial sentiment at a three month horizon. Full definitions of the proxies may be found in Table D.3 of Appendix D.

Taken collectively, this section shows that text metrics can and do capture, in a forward-looking and timely way, some of the same information as the proxies that policymakers typically look at. Text metrics for macroeconomic sentiment show the strongest relationship with existing proxies and this is likely to be due to the nature of the news sources. Across these tests, the text metrics that perform consistently well are TFIDF economy and Stability for sentiment, with a more mixed picture for uncertainty.

One risk with these conclusions is that because the Stability metric is designed to capture financial stress it does well because it either tracks terms that its creators could only have known about with hindsight or because a substantial amount of our variation occurs over the financial crisis, for which this metric is naturally suited. The former risk is low as the dictionary behind the Stability metric contains no proper nouns and almost all of its words are general, e.g. 'sluggish', with only a small number of specialist financial words, e.g. 'write-downs', and no words that would specifically and solely tie it to the Great Financial Crisis. The latter risk is more material but, as we will see in the next section, our preferred way of obtaining information from text will not rely on any pre-constructed text metric.

5 Forecasting exercises

Forecast exercises involve fitting of a model over a limited training period followed by out-of-sample predictions of target variables at given horizons. Models are re-estimated at every step in time according to a 36 month rolling window. If any transforms of the features are carried out, for instance normalisation, they are only performed with data from the past or present of the model. A detailed description of the training and testing exercise we run may be found in Appendix F.

Our forecast exercise seeks to answer whether a model with text included outperforms a very similar model with text not included. To reflect the timeliness of text, each forecast is done as if a policymaker at time t has information from text at time t, given by a scalar or vector indexed by time x_t, but information on the target variable y from time $t-1$, y_{t-1}, and before only. The policymaker wishes to forecast what y will be h steps in the future, y_{t+h}. This time t scenario of having potentially stale information on y but having access to newspaper text is of great relevance to policy where many official series only appear with a lag. The baseline model without text that we use for comparison will be either an AR(1) or a factor model regardless of the target of the forecasting exercise. While AR(1) models are simple there is overwhelming evidence that on average across series and time periods they are tough to beat (Carriero, Galvão and Kapetanios, 2018).

Our targets are monthly GDP, the unemployment rate, business investment (quarterly), household consumption (quarterly), consumer price inflation (CPI), the index of production (IOP), the index of services (IOS), the financial stress index of Chatterjee et al. (2017), and the IMF financial conditions index for the UK. We use a rolling window of 36 months for model estimation and horizons of $h = 3, 6, 9$. Most forecasts are performed at monthly frequency and we up-sample quarterly variables using interpolation through time from in-sample data points only. In the charts in this section, we plot error bars as the standard deviation of the forecast performance across both horizons and the different newspapers. Better forecast performance across horizons and newspapers is more indicative that the forecast gains are generally realisable.

We now turn to the two types of regressor (and specification) that we use in forecasts: algorithm-based text metrics and term frequency vectors.

6 Forecasting with algorithm-based text metrics

We evaluate the forecasting power of each text metric in turn using the model

$$y_{t+h} = \alpha + \beta \cdot y_{t-1} + \eta \cdot x_t + \epsilon_t$$

and we compare the performance of this model to the same one without the term in x_t (i.e. we force $\eta \equiv 0$). Figure 5 shows forecast RMSEs (root mean squared errors) relative to our AR(1) baseline by metric and target variable. While all target variables show an improvement relative to an AR(1) for some text metric, the top row shows three for which many text metrics provide performance improvements. Given the performance on GDP is good, it is not surprising that performance on the components of GDP is good too – especially the Index of Services (IOS) which accounts for a large fraction of GDP. More generally, real economy variables excluding CPI are improved by the addition of text while indicators of financial conditions are little improved. Further statistics for the AR(1) benchmark may be found in Appendix G.1.

We now test the addition of text to a model that includes additional macroeconomic information. We find that the added value of text degrades significantly when the benchmark is changed to a richer factor model that has highly statistically significant confounders. We utilise the macroeconomic factors derived from a dataset comprising 33 series covering real output, international trade, the labour market, inflation, house prices, retail sales, capacity utilisation, and business and household expectations (Redl, 2017). The factors are denoted by F. As before, the text model also includes a single algorithm-based text metric and an autoregressive term. The model is given by

$$y_{t+h} = \alpha + \beta \cdot y_{t-1} + \sum_j \gamma_j \cdot F_{jt} + \eta \cdot x_t + \epsilon_t$$

where x is the text metric. The benchmark against which this is compared is the same model above but without the term in x_t and we use $J = 2$ factors.

Figure 6 shows the forecast performance of the text and factors model. Across all targets, the results are weaker than in the case of just using an AR(1) as a benchmark. Most notably, very few target-metric pairs offer forecast improvements across all horizons and newspapers. Some of the text metrics that perform well against the AR(1) benchmark retain their position in the rankings, such as the Stability metric, while others, such as TFIDF Economy, rank far worse. In general, the simple and

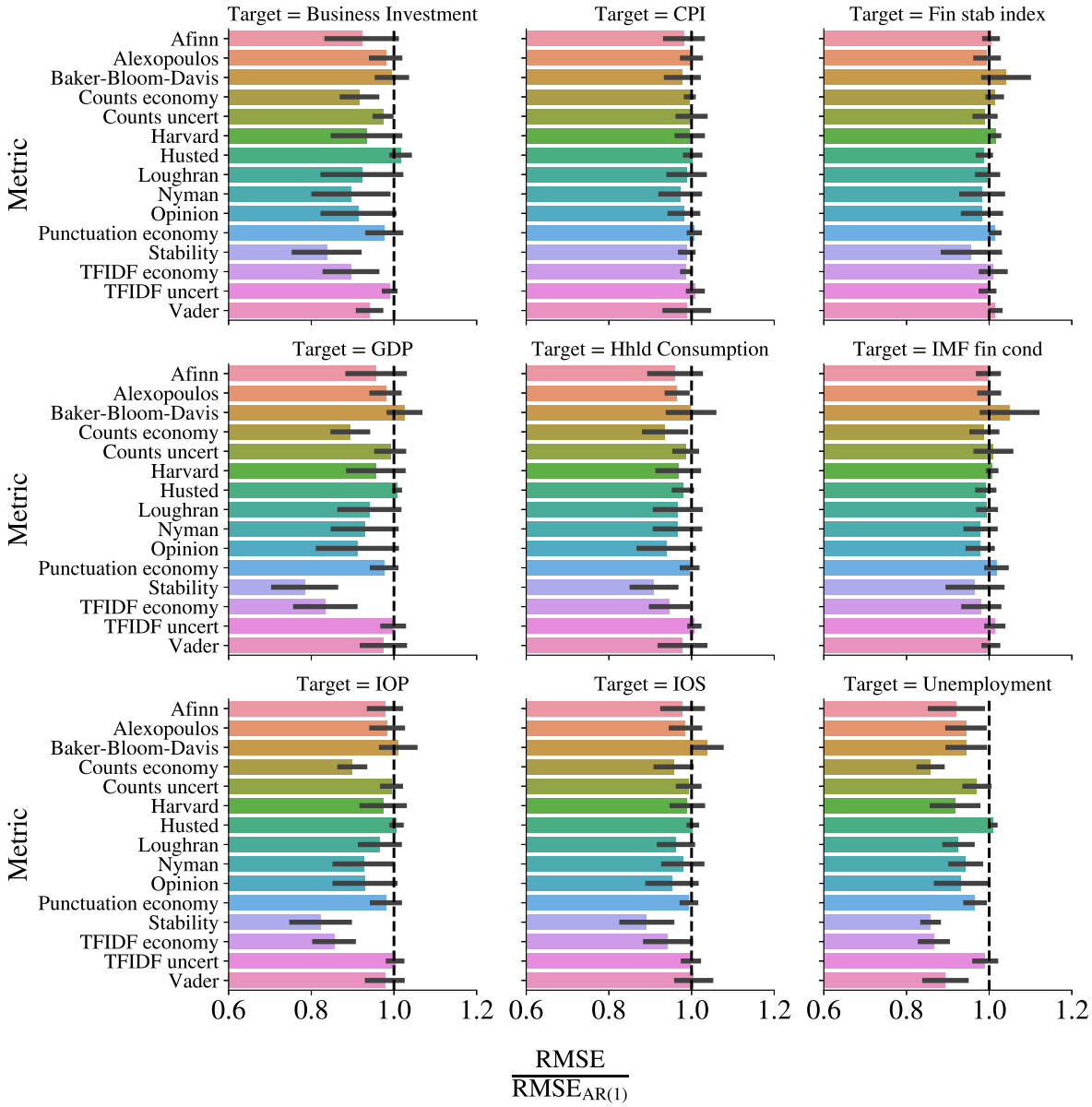

Figure 5: Results from the forecast exercise using algorithm-based text metrics. The plot shows RMSEs (x-axis) of a forecasting model with text in versus a benchmark AR(1) forecast without text. Facets are different target variables, the y-axis shows different algorithmic text metrics. Bars to the left of the dashed line indicate an improvement in forecast performance conducted with the given text metric. The confidence intervals are standard deviations over both the newspapers and the different horizons (3, 6 and 9 months ahead).

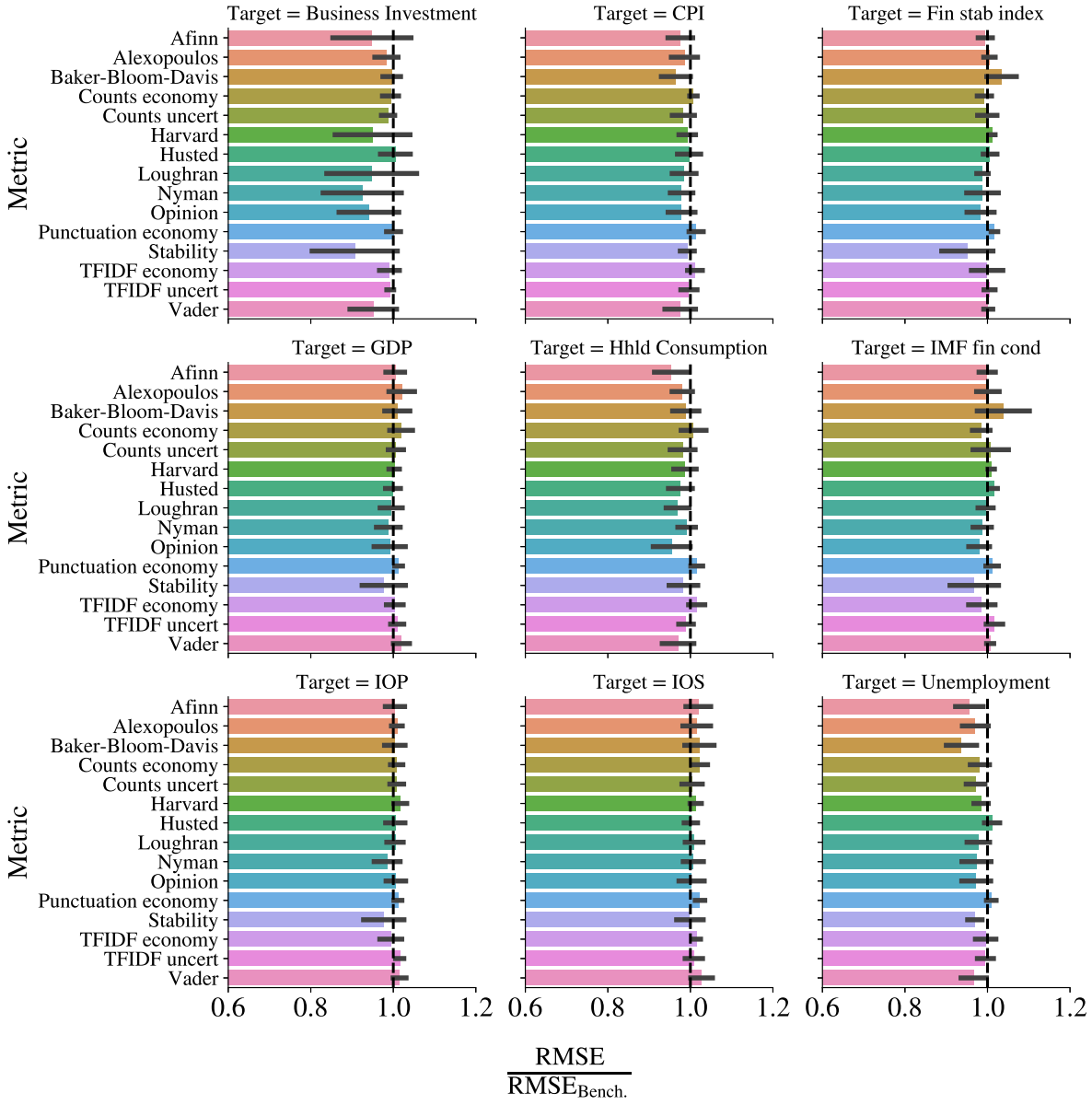

Figure 6: RMSEs relative to a benchmark AR(1) with factors by text metric and target variable. Bars to the left of the dashed line indicate an improvement in forecast performance conducted with the given text metric. The confidence intervals are standard deviations over both the newspapers and the different horizons (3, 6 and 9 months ahead).

transformed counts of single terms do not seem to add as much information and this is not unsurprising given that factor models are suited to capturing general macroeconomic trends that are likely to also be reflected in the number of times the economy is mentioned in newspapers. Those metrics associated with financial markets and finance (Stability, Loughran, Nyman) seem to perform relatively better with this benchmark, perhaps reflecting that our factors are based on time series that mostly capture information on the real economy. Further statistics for the factor model benchmark may be found in Appendix G.2.

7 Forecasting with text and machine learning

Here we use term frequency vectors and supervised machine learning, with its ability to handle a large feature space, to make forecasts. The models we employ from the machine learning literature are the least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996), ridge regression (Hoerl and Kennard, 1970), support vector regression (svm) (Chang, 2011; Drucker et al., 1997), elastic net (Zou and Hastie, 2005), artificial neural networks (Rumelhart, Hinton and Williams, 1985), and random forests (Breiman, 2001). The exact specification of each is defined in Appendix H.

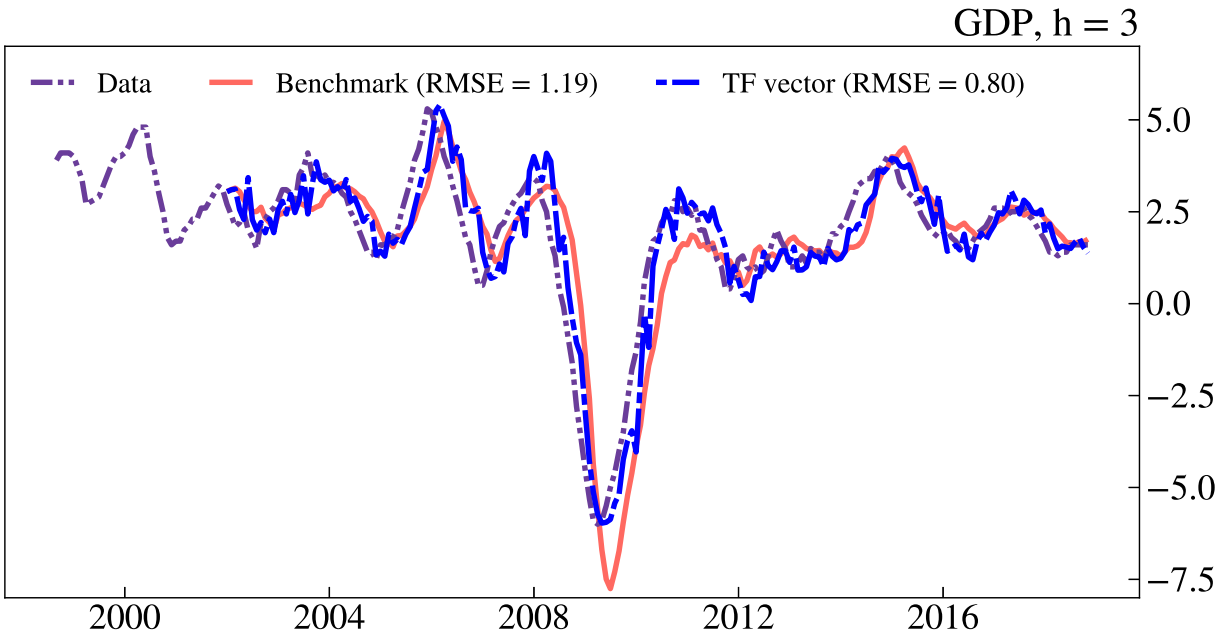

Figure 7: Forecasts for GDP three months ahead using OLS with a single lag (benchmark model, solid line) versus an artificial neural network that uses term frequency vectors from newspaper text in addition to a single lag of GDP (TF vector, dash line). The data are also shown (dot-dash line). Both models are estimated using a rolling window and newspaper text is taken from *The Daily Mail*.

Let X represent the full $N \times T$ matrix of features and $\vec{x}_t$ the same set of features at time t. We evaluate the forecasting power of each supervised machine learning model with text in turn using

$$\hat{y}_{t+h} = f\left(y_{t-1}, \vec{x}_t\right)$$

and we compare the performance of this model to the equivalent OLS model without the term in x_t, i.e. $y_{t+h} = \alpha + \beta \cdot y_{t-1} + \epsilon_t$ that we refer to as the OLS-AR(1) model. It is natural to ask why not compare against the same machine learning model without the text. We do make this comparison in Appendix I.2, but it is not an entirely fair one for several reasons. Firstly, the machine learning algorithms are most suited to a high number of features and our experience of using them with a small number of likely informative regressors is that they may do no better than OLS and so would not provide as difficult a benchmark to beat. Secondly, machine learning models have hyperparameters that can be tuned to produce better forecasts and we wish to avoid any ambiguity about hyperparameter tuning (for instance if the superior performance were down to hyperparameters that worked better with text than without it). Additionally, OLS is almost exclusively used in practice. The results in Appendix I.2 and I.3 show that the performance of the machine learning models with text is even stronger relative to a machine learning model without text benchmark but we choose to use OLS as our main benchmark model. For computational reasons, we do not perform hyperparameter tuning via in-sample cross-validation but instead opt for fixed hyperparameters as described in Appendix H.

An example forecast that uses machine learning, an artificial neural network, and term frequency vectors versus an OLS-AR(1) benchmark may be seen in Figure 7 for monthly GDP. In both cases, a single lag of GDP is included as a feature. There is a visually noticeable improvement in the goodness of fit and a substantially lower root mean squared error (RMSE). Note also that the machine learning model is also much quicker to respond to turning points.

In Figure 8 we show the forecast performance relative to the OLS-AR(1) benchmark for a range of machine learning models. Shown error bars are standard deviations over horizons of three to nine months ahead and the different newspapers. In contrast to forecasts with the algorithmic text metrics (see Figure 5), there are performance improvements relative to the benchmark for every target variable. The magnitude of these improvements is far larger too – few of the text metrics reached an improvement of 20 percentage points on any target variable while a substantial number of the machine learning forecasts have improvements of 30 percentage points or more. Two models perform consistently well:

neural network and ridge regression. The performance of Lasso being very similar with and without text suggests that it is not putting any weight on the term frequency vector, and the elastic net seems to perform similarly for similar reasons. Dense models perform best. This is in line with the findings of Giannone, Lenza and Primiceri (2017) who find that models that put some weight on a broad set of macro variables do best at forecasting macro variables. Appendix I.1 presents further statistics on this model.

We now look at a more stringent test of text, as we did in the case of the algorithmic text metrics in §6. We examine whether text still adds value when the baseline model includes two macroeconomic factors derived from a dataset comprising 33 series covering real output, international trade, the labour market, inflation, house prices, retail sales, capacity utilisation, and business and household expectations (Redl, 2017). The results of this are shown in Figure 9.

Unlike for the simpler text metrics, we find that the added value of text and machine learning versus an OLS-AR(1)-factor model is qualitatively similar to text and machine learning versus the OLS-AR(1) alone, suggesting that the combination of a rich set of features and sophisticated modelling can extract information from text that goes beyond what is available in a wide range of macroeconomic time series. As with the simpler machine learning specification, forecasts for every target variable can be improved and it is the support vector machine, neural network, and ridge regression that offer the best performance. Comprehensively and consistently offering forecast improvements versus a rich factor model suggests that this combination of a large number term frequency features and machine learning models is one of the best ways to get economic insight out of text.

One potential concern is that in running forecasts with so many methods, targets, horizons, and newspapers, our results may show forecast improvements that are statistical flukes. The error bars imply that this is not the case. To demonstrate this point more formally, we run a Diebold-Mariano test (Diebold and Mariano, 1995), with a small sample adjustment from Harvey, Leybourne and Newbold (1997), to check whether our results are statistically distinguishable from forecasts with the factor benchmark model in Table 4. Note that this test is still applicable to nested models in our case because we use rolling window estimation (Giacomini and White, 2006).

We show only those forecasts for which at least one target-model combination per newspaper had a smaller RMSE than the benchmark model. We find statistically significant results across newspapers at all horizons (the table shows $h = 9$). While not all combinations of target, model, and newspaper individually obtain significant results, the three models that do most well consistently are the neural

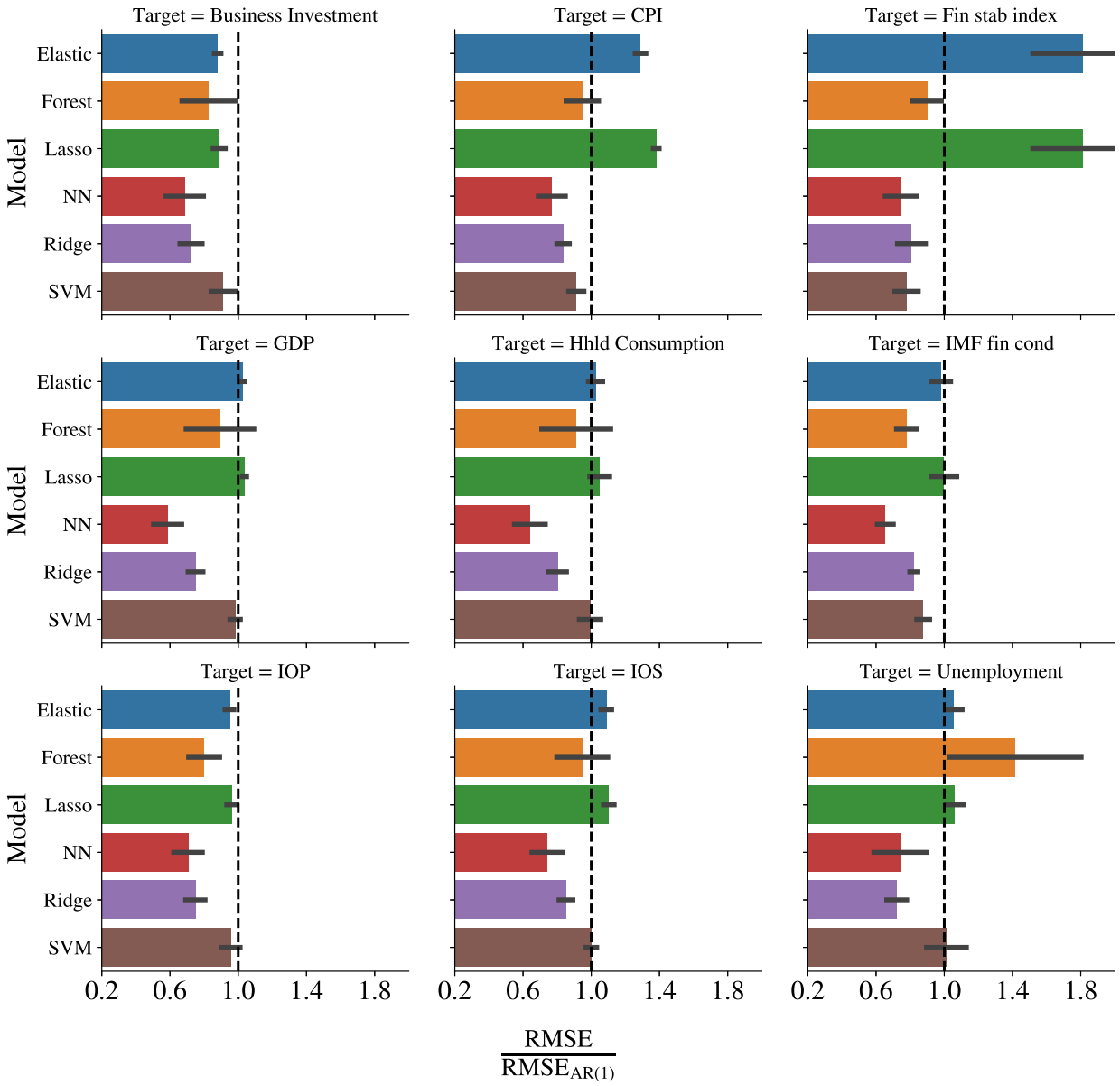

Figure 8: The relative improvement in root mean square error of a machine learning model that uses text and an AR(1) term versus OLS with the AR(1) term only. The facets are different target variables. Bars to the left of the dashed line indicate an improvement in forecast performance conducted with the given text metric. The confidence intervals are standard deviations over both the newspapers and the different horizons (3, 6 and 9 months ahead).

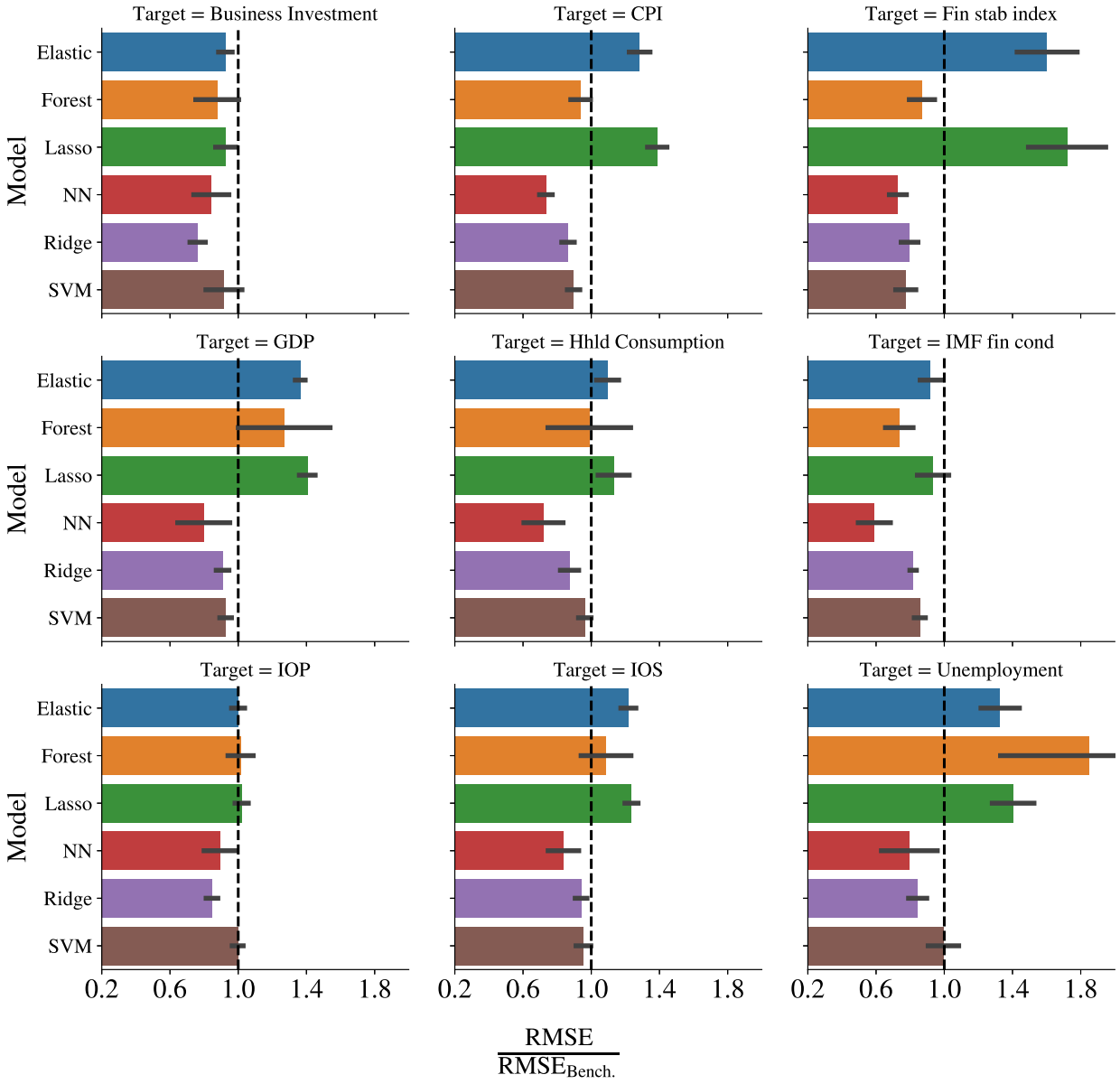

Figure 9: The relative improvement in root mean square error of a machine learning model that uses text, an AR(1) term, and factors versus OLS with the AR(1) and factors but no text. The facets are different target variables. Bars to the left of the dashed line indicate an improvement in forecast performance conducted with the given text metric. The confidence intervals are standard deviations over both the newspapers and the different horizons (3, 6 and 9 months ahead).

network, ridge regression, and support vector regression, reflecting what is visually represented in Figure 9. In Appendix I.3 we show that these results hold, and even improve, versus a machine learning benchmark with the same features.

Paper	Model	Target Horizon	Business Investment	CPI	Fin stab index	GDP	Hhld Consumption	IMF fin cond	IOP	IOS	Unemployment
The Daily Mail	Forest	9	-2.24**				-2.13**				
	NN	3		-2.07**			-2.06**	-1.83*			
		6		-2.13**	-1.68*	-2.25**	-2.07**				-1.91*
		9	-2.23**	-1.99**		-2.19**	-2.28**			-2.22**	-2.09**
	Ridge	3	-2.52**				-2.32**	-1.84*	-2.02**		
		6	-2.22**	-2.02**	-1.79*		-2.14**				
		9	-2.84***				-2.56**			-1.76*	
	SVM	3	-1.96*		-1.80*		-1.89*	-2.15**			
		6			-1.90*						
		9		-2.19**							-2.01**
The Daily Mirror	Forest	6			-1.98**						
		9					-2.10**				
	NN	3		-1.96*				-1.66*			
		6		-1.78*	-1.70*	-2.58**	-2.05**			-2.09**	-2.18**
		9	-1.82*	-2.05**		-1.92*	-1.74*				-2.03**
	Ridge	3	-1.84*					-1.76*	-1.74*		
		6	-3.38***				-1.85*				
		9	-2.33**				-2.11**				
	SVM	3						-1.69*			
		9		-1.74*							
The Guardian	Elastic	3	-1.86*								
	Forest	9					-2.09**				
	Lasso	3	-1.81*								
	NN	3		-2.18**				-1.65*			
		6		-2.29**	-1.79*		-1.99**				-1.70*
		9		-2.15**			-2.19**				-2.65***
	Ridge	3	-2.15**	-1.72*				-1.90*			-2.21**
		6	-1.76*	-1.91*	-1.77*		-2.16**				
		9	-1.67*	-1.87*			-2.73***				-1.68*
	SVM	3	-2.38**					-1.81*			
		9		-2.70***							-2.43**

Table 4: Results from a Diebold-Mariano test on the factor model using machine learning. Statistically significant differences in RMSE are shown. *, **, *** denote rejection of the null, of no difference in RMSE relative to an AR(1) and factors, at the 10%, 5%, 1% levels respectively. In the interests of space, only those targets for which at least one model-newspaper pair had a p-value of less than 10% are included.

7.1 When does text improve forecast performance?

Having shown that various different methods allow text to contribute to forecasts, we ask when does text count most for forecasts? We look at the breakdown of differences in squared error between OLS with only an AR(1) term and the most effective machine learning models with text and an AR(1) are shown in Figure 10 and denoted by $\varepsilon^2_{\text{Bench.}} - \varepsilon^2_{\text{Text}}$. When the lines are above zero, the model with text is performing better than the model without. This shows that most of the improvement in performance comes from the crisis period, which coincides when the costs of mistakes in forecasts are the highest. The same pattern is seen with the best performing text metrics (Appendix J).

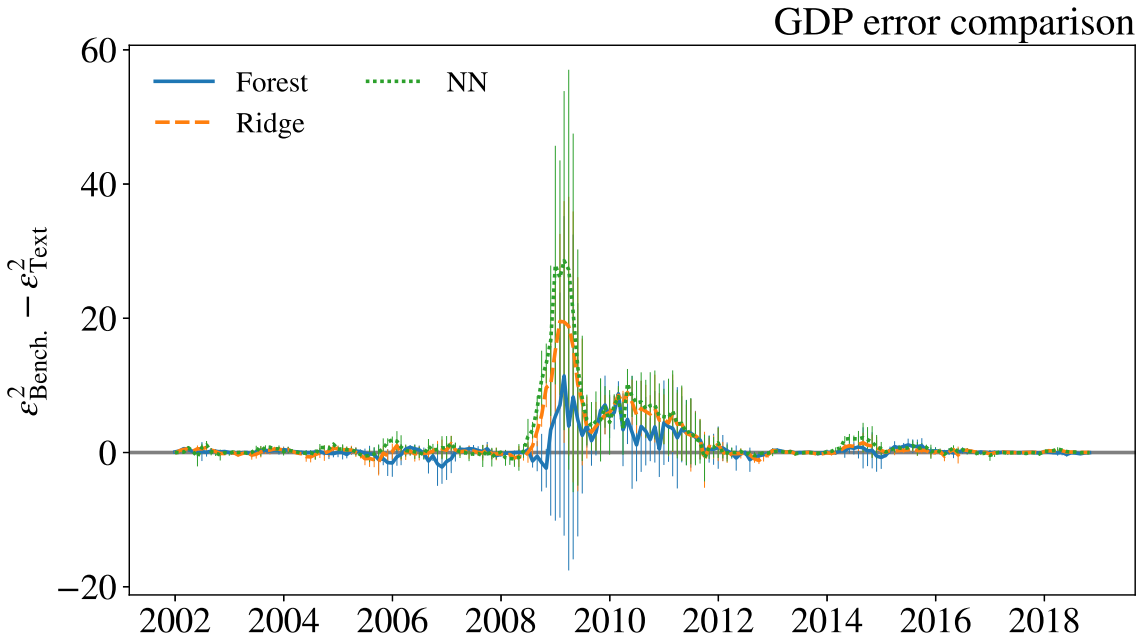

Figure 10: Mean squared error differences between a benchmark model and text over the time-dependent union of h-month ahead out-of-sample forecasts, with horizon $h = 3, 6, 9$. The target variable is monthly GDP. The benchmark is an OLS-AR(1) model. The plotted error bars are standard deviations over the different horizons and newspapers. A solid line above zero means that the model with text produces smaller errors than the benchmark model. Two different text metrics are shown. The majority of the forecast gains are during the crisis.

8 Discussion, summary, and conclusion

We set out to discover whether newspaper text could provide information about future economic activity that is relevant to policymakers and, if it can, what methods make text count the most, toward that end. Our results show that, across a range of methods, text can indeed provide forward-looking information about economic activity, and that this is robust both to horizons from 3 to 9 months and across three UK newspapers that span socio-economic groups and the political spectrum.

Text based indices of both sentiment and uncertainty are able to capture, in a forward looking way, some of the information that policymakers might usually get from other proxies for both. Although much attention has previously focused on the extraction of information about uncertainty from text we find that the signals of sentiment are both much better correlated with proxies for sentiment (than text based measures of uncertainty are with proxies for uncertainty), and seem to carry more information that is useful for predicting macroeconomic and financial target variables.

As well as being useful as proxies for sentiment and uncertainty, we show how text can add value

to economic forecasts in a statistically significant way. Comparing the pre-defined algorithms that turn text into time series and the feature engineering plus machine learning approach that we introduce in this paper we can make recommendations as to what methods one should use generally to get the most out of text. In increasing order of both complexity and performance, we recommend log transformed counts of words that are widely used and have an unambiguous meaning, the financial stability dictionary of Correa et al. (2017), and, finally, our approach of creating a very large set of features from terms in text and supplying that to a non-linear machine learning model such as a neural network.

In the case we explore here, we showed how even a log transform of the counts of the term economy ('econom' to be precise) was able to add a lot of value to a simple forecasting model and correlated well with other proxies for economic activity. This should not be surprise – when people talk about the economy a lot, it is likely to be because it is in trouble. The log transform provides extra stationarity and stability, and using a word that is not rare and does not have an unambiguous meaning protects from the use of the word in irrelevant contexts. A simple count of the word uncertainty did almost as well to the more complex Boolean methods for uncertainty suggested by Alexopoulos, Cohen et al. (2009) and Baker, Bloom and Davis (2016).

Only slightly more difficult to compute than a simple word count, the dictionary method of Correa et al. (2017) was originally designed to be an index for financial stability but we find that it performs the best of all other algorithmic methods as a proxy for sentiment and as an input into forecasts. As the newspapers used in our analysis are not geared towards specialists in financial markets or toward regulators, but toward the general public, this is a good indication of its general power to capture economic sentiment. Note that this method, like the previous one, collapses the information in each article down to a single number.

Finally, to get the most out of text, we recommend the approach that retains thousands of terms from text and turns each into a time series that can be used with a machine learning model. This is a departure from simpler models that collapse article text into a single number. While this approach is not appropriate for the construction of an indicator, because the machine learning model learns as it goes, it is by far and away the approach that produces the best improvements in forecasts with texts. The reasons for this are likely that more of the text gets into the model, the model decides which terms to put weight on, and the model itself is very powerful at prediction. We choose term frequencies to construct a time series per term appearing in the newspaper text, and we find that the best machine

learning model is also the most complex – the neural network. This method also has the advantage of being transferable to the prediction of whatever continuous variable using whatever text the researcher is interested in: that is, it is transferable to many other domains and applications away from the macroeconomic examples we present here. It is important to emphasise that the approach of retaining a large number of terms from text, turning each into a time series and then applying sophisticated nonlinear machine learning methods to produce predictions is, to the best of our knowledge, novel in the text analysis literature and since it also produces the best results, is by far our major contribution.

Regardless of whether using pre-defined text indices or our machine learning approach we find that newspaper text adds the most to forecasts during stressed times, perhaps reflecting the mantra that "if it bleeds, it leads". This is also when economic forecasting matters most, particularly for policymaking where mistakes made during stressed times are more costly. These findings echo those of Garcia (2013) for stocks and may also suggest that newspaper articles are where fast moving developments in the economy appear first, or just that the feedback loops between newspaper reports and real economic activity become more important during stressed times. Indeed there is evidence that periods of stress correspond to times of greater sensitivity to news (Akerlof and Shiller, 2010) and that newspapers can significantly influence their readers' views (Kennedy and Prat, 2017). Shiller (Shiller, 2017) has suggested that narratives may spread like viruses and play a causal role in economic activity, and newspapers could act as spreaders of such epidemiological narratives.

Although we have shown the power of combining a large set of text features and machine learning, feature engineering is a whole area of enquiry in itself and although our method is easily transferable to analysis in other contexts it is quite likely that we have not found the optimal process for creating features from text. The disadvantages of term frequencies include that they preserve neither the order nor the context of the words used, unlike transfer learning models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018). Future work could usefully explore the most effective features to use. Here, we focused on predicting the first moment of our target variables rather than anything to do with how uncertain they are over time. Another limitation is that here we used timely text to look at time scales that are policy relevant – the near future. But what is happening now is also relevant for policymakers and the evidence that we present suggests text would add a lot of value to as-if real-time nowcasts too.[8]

Our results have an immediate application in policy situations where decisions must be made on the basis of the near future but there is no official data, or even survey data, available on current

[8]There is informal evidence that it does – see https://bankunderground.co.uk/2019/02/28/whats-in-the-news-text-based-confidence-indices-and-growth-forecasts/

conditions and text can provide a more timely read on economic activity. For the three time series that the Bank of England publish forecasts of in the *Inflation Report* – GDP, unemployment, and CPI – we show that our approach gives forecast improvements versus a factor model benchmark that are as large as 30 percentage points of the benchmark RMSE and are also statistically significant. Furthermore, that text adds the most value during stressed times is of great significance to both fiscal and monetary policymakers.

References

Akerlof, George A, and Robert J Shiller. 2010. Animal spirits: How human psychology drives the economy, and why it matters for global capitalism. Princeton University Press. 27

Alexopoulos, Michelle, and Jon Cohen. 2015. "The power of print: Uncertainty shocks, markets, and the economy." International Review of Economics & Finance, 40: 8–28. 1

Alexopoulos, Michelle, Jon Cohen, et al. 2009. "Uncertain times, uncertain measures." University of Toronto Department of Economics Working Paper, 352. 7, 12, 26, 33

Antweiler, W., and M. Z. Frank. 2004. "Is all that talk just noise? The information content of internet stock message boards." Journal of Finance, 59(3): 1259–1294. 2

Ardia, David, Keven Bluteau, and Kris Boudt. 2019. "Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values." International Journal of Forecasting. 2

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring economic policy uncertainty." The Quarterly Journal of Economics, 131(4): 1593–1636. 1, 7, 12, 26, 33, 36

Bird, Steven, and Edward Loper. 2004. "NLTK: the natural language toolkit." 31, Association for Computational Linguistics. 32

Blei, David M, and John D Lafferty. 2006. "Dynamic topic models." 113–120, ACM. 3

Breiman, Leo. 2001. "Random forests." Machine learning, 45(1): 5–32. 19

Carriero, Andrea, Ana Galvão, and George Kapetanios. 2018. "A comprehensive evaluation of macroeconomic forecasting methods." International Journal of Forecasting (Forthcoming). 15, 40

Chang, Chih-Chung. 2011. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology, 2:3(27). 19

Chatterjee, Somnath, Ching-Wai (Jeremy) Chiu, Sinem Hacioglu-Hoke, and Thibaut Duprey. 2017. "A financial stress index for the United Kingdom." Bank of England Staff Working Paper 697. 15

Chauvet, Marcelle, and Simon Potter. 2013. "Forecasting output." In Handbook of Economic Forecasting. Vol. 2, 141–194. Elsevier. 40

Clark, Kevin, and Christopher D Manning. 2016. "Deep reinforcement learning for mention-ranking coreference models." arXiv preprint arXiv:1609.08667. 34

Correa, Ricardo, Keshav Garud, Juan M Londono, Nathan Mislang, et al. 2017. "Constructing a Dictionary for Financial Stability." Board of Governors of the Federal Reserve System (US). 7, 25, 26, 33, 35

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805. 27

Diebold, Francis X, and Robert S Mariano. 1995. "Comparing predictive accuracy." Journal of Business & economic statistics, 20(1): 134–144. 22

Drucker, Harris, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. 1997. "Support vector regression machines." 155–161. 19

Eckley, Peter. 2015. "Measuring economic uncertainty using news-media textual data." 4

Elango, Pradheep. 2005. "Coreference resolution: A survey." University of Wisconsin, Madison, WI. 34

Faust, Jon, and Jonathan H Wright. 2013. "Forecasting inflation." In Handbook of economic forecasting. Vol. 2, 2–56. Elsevier. 40

Garcia, Diego. 2013. "Sentiment during recessions." The Journal of Finance, 68(3): 1267–1300. 2, 26

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2017. "Text as data." National Bureau of Economic Research. 1

Giacomini, Raffaella, and Halbert White. 2006. "Tests of conditional predictive ability." Econometrica, 74(6): 1545–1578. 22

Giannone, Domenico, Michele Lenza, and Giorgio E Primiceri. 2017. "Economic Predictions with Big Data: The Illusion Of Sparsity." C.E.P.R. Discussion Papers CEPR Discussion Papers 12256. 20

Gilbert, CJ Hutto Eric. 2014. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." 7, 34

Harvey, David, Stephen Leybourne, and Paul Newbold. 1997. "Testing the equality of prediction mean squared errors." International Journal of forecasting, 13(2): 281–291. 22

Hoerl, Arthur E, and Robert W Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." Technometrics, 12(1): 55–67. 19

Hu, Guoning, Preeti Bhargava, Saul Fuhrmann, Sarah Ellinger, and Nemanja Spasojevic. 2017. "Analyzing Users' Sentiment Towards Popular Consumer Industries and Brands on Twitter." 381–388, IEEE. 7, 33, 34, 35

Hu, Minqing, and Bing Liu. 2004. "Mining and summarizing customer reviews." 168–177, ACM. 7, 33, 34, 35

Husted, Lucas F., John Rogers, and Bo Sun. 2017. "Monetary Policy Uncertainty." Board of Governors of the Federal Reserve System (U.S.) International Finance Discussion Papers 1215. 7, 12, 33

Jegadeesh, Narasimhan, and Di Wu. 2013. "Word power: A new approach for content analysis." Journal of Financial Economics, 110(3): 712–729. 1

Jurado, Kyle, Sydney C Ludvigson, and Serena Ng. 2015. "Measuring uncertainty." The American Economic Review, 105(3): 1177–1216. 13, 36

Kennedy, Patrick, and Andrea Prat. 2017. "Where Do People Get Their News?" Columbia Business School Research Papers 17-65. 27

Kozeniauskas, Nicholas, Anna Orlik, and Laura Veldkamp. 2018. "What are uncertainty shocks?" Journal of Monetary Economics, 100: 1 – 15. 12

Larsen, Vegard H, and Leif A Thorsrud. 2019. "The value of news for economic developments." Journal of Econometrics, 210(1): 203–218. 2

Loughran, Tim, and Bill McDonald. 2011. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." The Journal of Finance, 66(1): 35–65. 1

Loughran, Tim, and Bill McDonald. 2013. "IPO first-day returns, offer price revisions, volatility, and form S-1 language." Journal of Financial Economics, 109(2): 307–326. 1, 7, 33, 35

Manela, Asaf, and Alan Moreira. 2017. "News implied volatility and disaster concerns." Journal of Financial Economics, 123(1): 137–162. 2

Mumtaz, Haroon, and Alberto Musso. 2018. "The evolving impact of global, region-specific and country-specific uncertainty." European Central Bank Working Paper Series 2147. 11

Newsworks. 2018. "Circulation of newspapers in the United Kingdom (UK) as of June 2018 (in 1,000 copies)." Statista, Retrieved August 30, 2018. https://www.statista.com/statistics/529060/uk-newspaper-market-by-circulation/. 4

Nielsen, Finn Årup. 2011. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." arXiv preprint arXiv:1103.2903. 7, 33, 35

Nothman, Joel, Hanmin Qin, and Roman Yurchak. 2018. "Stop Word Lists in Free Open-source Software Packages." 7–12. 32

Nyman, Rickard, Sujit Kapadia, David Tuckett, David Gregory, Paul Ormerod, and Robert Smith. 2018. "News and narratives in financial systems: exploiting big data for systemic risk assessment." Bank of England Staff Working Papers, 704. 2, 7, 33, 35

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365. 27

Puurula, Antti. 2013. "Cumulative progress in language models for information retrieval." 96–100. 32

Redl, Chris. 2017. "The impact of uncertainty shocks in the United Kingdom." Bank of England Bank of England Staff Working Papers. 16, 22

Redl, Chris. 2018. "Uncertainty matters: evidence from close elections." Bank of England Bank of England Staff Working Papers. 13, 36

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams. 1985. "Learning internal representations by error propagation." California Univ San Diego La Jolla Inst for Cognitive Science. 19

Shapiro, Adam Hale, Moritz Sudhof, and Daniel Wilson. 2018. "Measuring news sentiment." Federal Reserve Bank of San Francisco. 2

Shiller, Robert J. 2017. "Narrative economics." The American Economic Review, 107(4): 967–1004. 27

Tetlock, Paul C. 2007. "Giving content to investor sentiment: The role of media in the stock market." The Journal of finance, 62(3): 1139–1168. 2, 7, 33, 36

Thorsrud, Leif Anders. 2018. "Words are the new numbers: A newsy coincident index of the business cycle." Journal of Business & Economic Statistics, 1–17. 2, 3

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological), 267–288. 19

Turrell, Arthur, James Thurgood, Jyldyz Djumalieva, David Copple, Bradley Speigner, et al. 2018. "Using online job vacancies to understand the UK labour market from the bottom-up." Bank of England Staff Working Paper 742. 3

Zipf, George K. 1950. "Human behavior and the principle of least effort." 7

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology), 67(2): 301–320. 19

Making text count: economic forecasting using newspaper text

Appendix

Eleni Kalamara Arthur Turrell Chris Redl George Kapetanios Sujit Kapadia

A Text cleaning

Text must be processed in order for it to be used in any quantitative application. Except where stated otherwise, for the algorithm-based text metrics, we use the following methods to pre-process newspaper text.

1. remove punctuation, hyperlinks, hyper text markup language (HTML) tags, special characters, leading or trailing white space characters, and digits;
2. set all characters in lower case; and
3. drop words which are in our list of stop words.

Note that we do not use stemming or lemmatisation. It is common practice to drop a large number of words from a corpus before turning it into a quantitative measure over text. One of the reasons is that a large number of words in any text corpus is uninformative, either because it occurs very rarely or very frequently. Words in the latter category are often known as 'stop words' and include 'and', 'is', 'in', and so on (see Nothman, Qin and Yurchak (2018) for a discussion). As noted in §3, one of the common approaches to excluding words is to use threshold frequencies (both high and low) applied to the entire corpus. However, this requires knowledge of the entire corpus ahead of time and is not suitable for real time forecasting (see §3.1 for a discussion of this). Instead, it is necessary to define ahead of time a set of words that will not be retained. We drop words from the union of two popular lists of stop words: the NLTK word list (Bird and Loper, 2004) and the list proposed by Puurula (2013).

B Turning text into time series

B.1 Algorithm based text metrics

B.1.1 Dictionary methods

Dictionary methods measure sentiment using a pre-defined list of words associated with scores. The scores are usually positive and negative scores with values of +1 and -1, respectively, in the simplest case. These scores are counted for each article. The net score, weighted by the number of words with scores, is the sentiment score for each article. For each news source, let articles – which consist of a group of (possibly repeated) terms – be denoted a. Each dictionary D is split into positive, D^+, and negative, D^- parts and defines a mapping $D : W \to C$ such that $w \in W$ has an associated score $c \in C$. Not all terms in every article are in the domain of D. The sentiment score for an article a with terms

Table B.1: Lists of words in the UK BBD metric.

E, Economics words	economic, economy
U, Uncertainty words	uncertainty, uncertain
P, Policy words	spending, policy, deficit, budget, tax, regulation, bank of england

w is given by

$$S = \frac{1}{|w|}\left(\sum_{w} D^{+}(w) - \sum_{w} D^{-}(w)\right)$$

The purely dictionary based text metrics with positive and negative words which we use are from Nyman et al. (2018), Loughran and McDonald (2013), Nielsen (2011), Hu and Liu (2004) and Hu et al. (2017), and Correa et al. (2017) in addition to the Harvard IV psychological dictionary used by Tetlock (2007).

B.1.2 Boolean methods

These metrics are typically counts of articles which satisfy a logical condition (within a given time period). They may also be normalised by the total number of articles within a time period. As the most simple example of this, we count the number of occurrences of "uncertain" and "econom" aggregated over the relevant time scale.

We also use more elaborate Boolean metrics. For instance, ones in which, given two sets of words, E and U, and w a term in article a, article a is counted if and only if

$$(w \in E) \wedge (w' \in U) \quad \forall \quad w, w' \in a$$

A daily measure is created from the ratio of the number of counts each day to the number of articles satisfying the condition each day.

The uncertainty measure of Alexopoulos, Cohen et al. (2009) falls into this category, with $U = \{\text{uncert}, \text{uncertainty}\}$ and $E = \{\text{econom}, \text{economy}\}$.

Baker, Bloom and Davis (2016) describe 'Economic Policy Uncertainty'. The UK measure uses counts of the logical combination of three lists. We use a very similar measure to theirs, denoted as 'baker_bloom_davis'. If terms from all three of the lists shown in Table B.1 appear in an article, a count is recorded.

We also use the Boolean logic monetary policy uncertainty measure of Husted, Rogers and Sun (2017) with a slight modification for real time forecasting. Their measure counts the number of articles containing the triple of (i) "uncertainty" or "uncertain," and (ii) "monetary policy(ies)" or "interest rate(s)" or "Bank rate" and (iii) " Bank of England" or "BoE". This is normalised by the total number of articles mentioning category (iii) words for a given newspaper-period. The index is then rescaled to have a standard deviation of unity across the entire sample. For our purposes, the latter step is not appropriate as it introduces information leakage. Instead, where we normalise, we only use data up to and including that point, or the in-sample in a forecast test environment. We divide by the number of articles mentioning category (iii) words within each day.

B.1.3 Word counts

We include in our text metrics some simple counts of the number of words, and also transforms of those simple counts. We use two metrics that are transforms of counts: TFIDF economy and TFIDF uncert which, as part of their construction, look for the strings 'econom' and 'uncertain' respectively. The details of the tfidf transforms are in §3.2.

B.1.4 Methods from computer science

We use the Valence Aware Dictionary for sEntiment Reasoning (VADER) metric of Gilbert (2014). This is a rule based metric that embodies grammatical and syntactical conventions that humans use when expressing or emphasising sentiment intensity. It is oriented to small snippets of text, such as tweets, and produces a magnitude of sentiment in addition to a sign. The (unnormalised) sentiment intensity is on a scale from -4 to +4. For example, the word "okay" has a positive score of 0.9, "good" is 1.9, and "great" is 3.1, whereas "horrible" is -2.5, and the frowning emoticon ":(" is -2.2. The sentiment scores are calculated on a sentence level and we create per article sentiment by averaging the scores and dividing by the total number of sentences in each article[9].

We also adopt a metric based on the opinion mining literature (Hu et al., 2017; Hu and Liu, 2004). Although strictly speaking a dictionary method, the words have not been selected *a priori* by a researcher. Instead, the 'opinion sentiment' dictionary is constructed from words which have strong positive or negative connotations as discovered by text summarisation techniques applied to web reviews of products. As such, the dictionary reflects consumer preferences. The series are constructed by subtracting the positive and negative counts of words and normalising by the total number of words in each article.

Finally, we develop a metric based on measuring sentiment within individual sentences, discarding the information contained in the remainder of the article. Given a specific term, the metric returns the sentiment of the words of the surrounding sentence fragment. In our results, we use the term 'econom', the root of the word 'economy', as the search term and call this metric 'punctuation economy'. This approach retains and processes snippets of text up to the closest punctuation characters if they contain the term(s) of interest.

We do not take notice of punctuation associated with titles, such as 'Mr.', 'Mrs.', 'Dr.', 'etc.', etc., as being the end of a sentence or segment of text. Exclamation marks and periods not associated with titles or well-used abbreviations are counted as ending sentences, while punctuation such as ',', '?', ';' is counted as ending or beginning a snippet of text.

For this metric, we *first* search the raw text for the term(s) of interest, keeping the punctuation and stop words. This allows us to perform coreference resolution (Clark and Manning, 2016; Elango, 2005). Coreference resolution allows for any linguistic expressions that refer to the same real-world entity indirectly to be replaced by explicit references to that real world entry[10]. This ensures that we can find all fragments of sentences which refer to a particular term before then analysing them for sentiment.

[9]The model is available as a part of NLTK sentiment analysis Python package.

[10]An example would be "The cat is on the mat. It looks hungry.", which would be converted to "The cat is on the mat. The cat looks hungry."

For the sentiment analysis, this metric uses the words from the union of the dictionaries of Nyman et al. (2018), Nielsen (2011), Correa et al. (2017), and the Harvard IV psychological dictionary, keeping separate the positive and negative version of each and scoring their words as $+1$ and -1 respectively.

B.2 Stationarity of text metrics

We determine whether our algorithm-based series are stationary. An augmented Dickey-Fuller test was run, using the Akaike information criterion to choose the number of lags, to test the null hypothesis of a unit root against the alternative hypothesis of stationarity. At a 1% significance level, we can reject the null hypothesis for all metrics for at least one of the three newspapers. The null cannot be rejected at the 10% significance level for a small number of newspaper-text metric pairs, mostly those based on raw counts of occurrences. *The Guardian* had the fewest significant results. The null can be rejected most strongly for dictionary and computer-science methods, followed by Boolean methods, followed by word counts.

	The Daily Mirror	No. obs.	The Daily Mail	No. obs.	The Guardian	No. obs.
TFIDF uncert	-02.29	253	-07.74***	272	-05.00***	321
Counts uncert	-04.72***	255	-04.31***	268	-01.98	308
Alexopoulos	-02.01	241	-04.33***	267	-01.90	309
Baker-Bloom-Davis	-05.97***	255	-05.44***	271	-01.55	312
Husted	-08.60***	256	-08.96***	272	-04.27***	319
Opinion	-04.09***	254	-04.51***	269	-02.93**	320
Harvard	-07.12***	255	-04.11***	269	-03.82***	319
Loughran	-04.39***	255	-04.51***	268	-02.40	320
Vader	-03.81***	253	-04.29***	271	-03.18**	320
Afinn	-03.75***	254	-02.57*	264	-02.68*	320
Counts economy	-01.92	249	-03.52***	270	-03.48***	320
Stability	-04.35***	255	-04.65***	272	-03.17**	315
TFIDF economy	-03.23**	256	-04.33***	272	-03.50***	311
Nyman	-05.42***	255	-04.61***	271	-03.30**	312
Punctuation economy	-06.21***	254	-06.26***	269	-08.28***	321

Table B.2: Results of an Augmented Dickey-Fuller test on all text metrics. The number of observations differ as the number of lags to include is chosen using the AIC information criterion. Asterisks denote p-values; 1%: ***, 5%: **, 10%: *.

C Term frequency vectors

The term frequency for a term w in an article a is denoted $\text{tf}(a)_w$ and is simply the counts of term w in that article. Term frequency vectors are the vector representation of all (tracked) terms in an article or across articles in a given time period t. For example, for articles, the term frequencies define a vector space: $V : a \to \mathbb{R}^N$ with N the dimension of the vector space and, equivalently, the number of tracked terms. A complete matrix, tf, may also be defined in which each column is a term from the pre-defined set of all terms, and each row is an observation (an article or collection of articles within a time period).

The pre-defined list of terms used to construct the term frequency matrix uses the union of several dictionaries. These are those dictionaries found in Nyman et al. (2018), Loughran and McDonald (2013), Nielsen (2011), Hu and Liu (2004) and Hu et al. (2017), and Correa et al. (2017). We add

to this a collection of words related to economics and finance[11] and the Harvard IV psychological dictionary used by Tetlock (2007). We use n-grams up to trigrams if they already exist individually in these dictionaries. This gives 9660 unique terms of which 8030 appear in our corpus.

D Proxies

We use a number of proxies from private and public data providers. These are series often used as indicators by policymakers. To these we also add recently developed series focusing on uncertainty from the academic literature. The complete list of proxies appears in Table D.3.

Name	Description	Type
Lloyds Bus Conf	Lloyds Business Barometer – confidence	Sentiment
Lloyds Bus Activity	Lloyds Business Barometer – activity over next 12 months	Sentiment
OECD Bus Conf	OECD UK business confidence	Sentiment
Composite PMI	Composite measure of PMI	Sentiment
GfK Consumer Conf	GfK Consumer Confidence	Sentiment
IG Corp Bond spread	Investment Grade Corporate Bond spread	Uncertainty, sentiment
Jurardo Fin Uncert	UK version of Jurado, Ludvigson and Ng (2015) from Redl (2018); financial uncertainty, $h = 3$	Uncertainty
Jurardo Macro Uncert	UK version of Jurado, Ludvigson and Ng (2015) from Redl (2018); macroeconomic uncertainty, $h = 3$	Uncertainty
BoE agg credit spread	Bank of England measure of aggregate credit spread	Uncertainty
VIX	CBOE volatility index	Uncertainty
VFTSEIX	FTSE volatility	Uncertainty
EPUUK	Baker, Bloom and Davis (2016) economic policy uncertainty index for UK	Uncertainty
GDP forecast std dev	UK Treasury collected standard deviation of professional forecasts of GDP, 3 months ahead	Uncertainty
BoE Uncert	Bank of England uncertainty index	Uncertainty
ERI volatility	GBP Exchange Rate Index volatility	Uncertainty

Table D.3: Descriptions of the proxy time series and what they are used for.

D.1 Granger Causality Tests

Tables D.4 and D.5 show results of Granger causation tests with text metrics and proxies for both sentiment and uncertainty.

	Husted	Stability	Counts economy	TFIDF economy	Counts uncert	Alexopoulos	Punctuation economy	Baker-Bloom-Davis	TFIDF uncert	Harvard	Vader	Nyman	Afinn	Loughran	Opinion
BoE agg credit spread	48.29***	4.69***	2.46*	3.17**	2.77**	1.75	4.63***	3.28**	1.43	0.13	2.10*	1.37	2.27*	1.07	0.80
Lloyds Bus Activity	0.79	7.07***	12.34***	12.09***	1.56	2.34*	6.28***	2.85**	1.12	1.73	1.05	1.71	0.40	0.90	0.73
OECD Bus Conf	32.24***	2.11*	1.65	2.06	0.29	1.72	1.52	0.79	0.20	0.65	1.06	1.14	1.20	2.27*	0.52
EPU UK	1.89	1.81	4.15***	1.84	12.52***	6.86***	0.69	3.08**	6.36***	2.21*	0.75	0.63	0.08	0.06	0.50
VFTSEIX	1.44	4.74***	1.12	0.92	2.76**	3.78**	1.30	0.97	1.49	3.99***	4.39***	2.47*	4.15***	2.17*	2.93**
Lloyds Bus Conf	0.88	3.65**	5.69***	7.17***	0.59	0.78	4.35***	1.01	0.61	1.23	1.21	1.80	1.28	1.30	0.88
Jurardo Macro uncert	5.44***	3.13**	1.18	0.68	3.62**	4.25***	0.84	1.04	3.77**	1.00	1.08	0.91	0.74	1.22	1.05
IG Corp Bond spread	0.64	3.95***	3.78**	4.39***	2.77**	2.21*	1.52	1.85	2.01	1.27	0.34	0.57	0.85	1.16	1.80
Composite PMI	0.85	5.85***	3.07**	2.77**	1.32	1.88	1.52	1.52	1.91	3.06**	1.22	1.89	0.63	0.10	0.36
Jurardo Fin uncert	3.43**	1.72	2.07	1.06	0.93	1.31	1.74	3.84**	1.66	0.98	0.72	0.04	0.28	0.89	0.07
GfK Consumer Conf	2.77**	2.08	1.34	0.77	2.44*	2.27*	3.00**	1.03	1.04	0.39	0.33	0.41	0.46	0.22	0.47
GDP forecast std dev	2.14*	2.20*	0.44	0.74	1.12	2.40*	1.25	1.74	0.57	1.01	0.45	1.40	0.43	0.65	0.55
ERI volatility	0.49	1.48	1.60	1.04	0.22	1.11	1.39	2.82**	0.19	0.92	2.07	0.23	0.97	0.97	0.37
BoE uncert	2.03	3.38**	1.33	1.11	0.91	1.03	0.61	0.94	0.86	0.95	0.11	0.38	0.22	0.53	0.27
VIX	0.62	0.78	0.34	0.16	0.77	0.78	0.27	1.55	0.14	0.30	0.08	0.26	0.17	0.22	0.26

Table D.4: Test of whether text metrics Granger cause proxies, at a three month horizon. The text metrics are averaged across the three newspapers. Asterisks denote p-values; 1%: ***, 5%: **, 10%: *.

E Regressions of GDP on text metrics

Here we look at the relationship between the different text-based measures and GDP, as a variable of primary interest, to try and ascertain whether there is any information about its future path (9 months

[11] Most of these come from https://home.ubalt.edu/ntsbarsh/stat-data/KeywordsPhra.htm and http://home.ubalt.edu/ntsbarsh/Business-stat/stat-data/KeysPhrasFinance.htm.

	Husted	Stability	Counts economy	TFIDF economy	Counts uncert	Alexopoulos	Punctuation economy	Baker-Bloom-Davis	TFIDF uncert	Harvard	Vader	Nyman	Afinn	Loughran	Opinion
BoE agg credit spread	48.29***	4.69***	2.46*	3.17**	2.77**	1.75	4.63***	3.28**	1.43	0.13	2.10*	1.37	2.27*	1.07	0.80
Lloyds Bus Activity	0.79	7.07***	12.34***	12.09***	1.56	2.34*	6.28***	2.85**	1.12	1.73	1.05	1.71	0.40	0.90	0.73
OECD Bus Conf	32.24***	2.11*	1.65	2.06	0.29	1.72	1.52	0.79	0.20	0.65	1.06	1.14	1.20	2.27*	0.52
EPU UK	1.89	1.81	4.15***	1.84	12.52***	6.86***	0.69	3.08**	6.36***	2.21*	0.75	0.63	0.08	0.06	0.50
VFTSEIX	1.44	4.74***	1.12	0.92	2.76**	3.78**	1.30	0.97	1.49	3.99***	4.39***	2.47*	4.15***	2.17*	2.93**
Lloyds Bus Conf	0.88	3.65**	5.69***	7.17***	0.59	0.78	4.35***	1.01	0.61	1.23	1.21	1.80	1.28	1.30	0.88
Jurardo Macro uncert	5.44***	3.13**	1.18	0.68	3.62**	4.25***	0.84	1.04	3.77**	1.00	1.08	0.91	0.74	1.22	1.05
IG Corp Bond spread	0.64	3.95***	3.78**	4.39***	2.77**	2.21*	1.52	1.85	2.01	1.27	0.34	0.57	0.85	1.16	1.80
Composite PMI	0.85	5.85***	3.07**	2.77**	1.32	1.88	1.52	1.52	1.91	3.06**	1.22	1.89	0.63	0.10	0.36
Jurardo Fin uncert	3.43**	1.72	2.07	1.06	0.93	1.31	1.74	3.84**	1.66	0.98	0.72	0.04	0.28	0.89	0.07
GfK Consumer Conf	2.77**	2.08	1.34	0.77	2.44*	2.27*	3.00**	1.03	1.04	0.39	0.33	0.41	0.46	0.22	0.47
GDP forecast std dev	2.14*	2.20*	0.44	0.74	1.12	2.40*	1.25	1.74	0.57	1.01	0.45	1.40	0.43	0.65	0.55
ERI volatility	0.49	1.48	1.60	1.04	0.22	1.11	1.39	2.82**	0.19	0.92	2.07	0.23	0.97	0.97	0.37
BoE uncert	2.03	3.38**	1.33	1.11	0.91	1.03	0.61	0.94	0.86	0.95	0.11	0.38	0.22	0.53	0.27
VIX	0.62	0.78	0.34	0.16	0.77	0.78	0.27	1.55	0.14	0.30	0.08	0.26	0.17	0.22	0.26

Table D.5: Test of whether proxies Granger cause text, at a three month horizon. The text metrics are averaged across the three newspapers. Asterisks denote p-values; 1%: ***, 5%: **, 10%: *.

ahead) within the news. We assume the same model as is used in the baseline case in §6, that is:

$$y_{t+h} = \alpha + \beta \cdot y_{t-1} + \eta \cdot x_t + \epsilon_t$$

where x_t is a text metric using time t text. Tables E.6 and E.7 look at this for sentiment and uncertainty respectively. All measures we look at are significant even when controlling for a lag of GDP. However, this does not translate into better R^2 values for uncertainty and sentiment indicators. For uncertainty, only the Alexopoulos and Baker-Bloom-Davis metrics provide a very marginal improvement in R^2 but have smaller F-statistics. The performance of text is better for sentiment metrics: all provide a boost in R^2, particularly TFIDF economy and Stability. These also have higher F-statistics than the baseline case with no text, while the other sentiment metrics do not. The important takeaway from this exercise is that there is a demonstrable link between text and real economic activity across the board.

F Forecast environment

Features are indexed by $k = 0, \ldots, K$, time by $t = 0, \ldots, T$, the window step size by $s \geq 1$, the initial training period length as $\alpha + s$, and train (and associated test) periods by $\mu = 1, \ldots, \frac{T-s-\alpha}{s}$. $\alpha = 0$ implies that the initial training period is of length s. Define $\{y_t\}_{t=0}^{t=T}$ as the target variable shifted h steps ahead, for h the desired horizon of the forecast. It is denoted $\vec{y}$ for short. Let $\{x_{tk}\}_{t=0}^{t=T}$ represent feature k, also denoted $\vec{x}_k$. The entire set of features of all time form a matrix X. Though we use rolling window estimation for all results presented we define below the cuts of the data for both expanding and rolling window estimation. Also, in both cases, the test set is composed of data points that have never been used for estimation and lie in the future (in time) of the training set, i.e. for a rolling window from $t = 20$ to $t = 25$ the test set would run from $t = 26$ to $t = T$.

Our in-sample and out-of-sample results as presented are created from the union of the last in-sample prediction of each estimation window and the first out-of-sample prediction of the same estimation window, respectively. These are defined formally below.

F.1 Expanding window

Define

$$I_\mu^e(\vec{z}) = \left\{ z_t \right\}_{t=0}^{t=\mu \cdot s + \alpha - 1}$$

Dep. variable: GDP	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Afinn	16.74***										
	(5.41)										
Counts economy										2.79***	
										(0.29)	
Harvard		86.06***									
		(14.87)									
Loughran			67.93***								
			(13.52)								
Nyman				186.38***							
				(47.27)							
Opinion					51.14***						
					(10.68)						
Punctuation economy						68.73***					
						(10.14)					
Stability							381.32***				
							(25.13)				
TFIDF economy								1038.10***			
								(74.37)			
Vader									2.76***		
									(0.70)		
Lagged GDP	0.24***	0.22***	0.22***	0.27***	0.23***	0.22***	0.09***	0.06*	0.23***	0.17***	0.28***
	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Intercept											1.37***
											(0.10)
The Daily Mail	1.58***	0.38*	3.53***	1.95***	2.41***	1.79***	5.72***	4.03***	0.89***	3.26***	
	(0.16)	(0.21)	(0.45)	(0.21)	(0.26)	(0.15)	(0.31)	(0.23)	(0.18)	(0.24)	
The Daily Mirror	1.76***	0.73***	3.43***	1.53***	2.48***	1.63***	4.95***	2.80***	1.32***	2.11***	
	(0.19)	(0.17)	(0.43)	(0.14)	(0.27)	(0.14)	(0.27)	(0.16)	(0.14)	(0.15)	
The Guardian	1.55***	-0.18	3.49***	1.81***	2.25***	1.63***	4.80***	3.90***	0.88***	3.84***	
	(0.15)	(0.30)	(0.44)	(0.17)	(0.23)	(0.14)	(0.25)	(0.22)	(0.19)	(0.29)	
R-squared	0.09	0.12	0.11	0.10	0.11	0.13	0.30	0.27	0.10	0.18	0.08
Adj. R-squared	0.09	0.11	0.10	0.09	0.10	0.13	0.30	0.27	0.09	0.18	0.08
No. observations	724	724	724	724	724	724	724	724	724	724	724
F-statistic	17.88	24.38	22.14	19.51	21.51	27.76	77.75	68.14	19.47	39.79	61.37

Table E.6: Regression of GDP growth, at nine months ahead, on text-based sentiment measures. Asterisks denote p-values; 1%: ***, 5%: **, 10%: *.

as the in-sample expanding window slice μ for an arbitrary time vector $\vec{z}$. Similarly, define the associated out of sample slice μ as:

$$O^e_\mu(\vec{z}) = \left\{ z_t \right\}_{t=\mu\cdot s+\alpha}^{t=T}$$

Transformations T are labelled by whether they are expanding (e) or rolling (r), and for the feature, k, they are based on. For instance, a normalisation transformation is given by

$$T^e_{\mu k}(\vec{z}) = T^e_{\mu k}(\vec{z}; I^e_\mu(\vec{x}_k)) = T^e_{\mu k}\left(\vec{z}; \{x_{kt}\}_{t=0}^{t=\mu\cdot s+\alpha-1}\right) = \frac{\vec{z} - \langle I^e_\mu(\vec{x}_k)\rangle}{\sigma_{I^e_\mu(\vec{x}_k)}}$$

Transformations are indexed by μ to avoid information leakage (aka look-ahead bias). In general, the feature index on T will be implicit.

Define f_μ as the model which results from trying to fit $T^e_\mu(I^e_\mu(X))$ to $\vec{y}$. In-sample tests are based on $f_\mu\left(T^e_\mu(I^e_\mu(X))\right)$, while out-of-sample tests are performed on

$$f_\mu\left(T^e_\mu(O^e_\mu(X))\right)$$

To create a unified in-sample set from the end of each in-sample estimation window (recall that

Dep. variable: GDP	I	II	III	IV	V	VI
Alexopoulos	-21.60***					
	(5.19)					
Baker-Bloom-Davis		-43.34***				
		(11.31)				
Counts uncert					-6.38**	
					(2.64)	
Husted			-70.75**			
			(29.73)			
TFIDF uncert				-3229.52***		
				(1219.57)		
Lagged GDP	0.26***	0.26***	0.28***	0.27***	0.27***	0.28***
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Intercept						1.37***
						(0.10)
The Daily Mail	1.80***	1.84***	1.56***	1.71***	1.65***	
	(0.17)	(0.19)	(0.16)	(0.19)	(0.18)	
The Daily Mirror	1.48***	1.45***	1.38***	1.51***	1.45***	
	(0.14)	(0.14)	(0.14)	(0.15)	(0.14)	
The Guardian	1.87***	1.82***	1.51***	1.65***	1.74***	
	(0.18)	(0.18)	(0.15)	(0.17)	(0.20)	
R-squared	0.10	0.10	0.09	0.09	0.09	0.08
Adj. R-squared	0.10	0.09	0.08	0.08	0.08	0.08
No. observations	724	724	724	724	724	724
F-statistic	19.99	19.27	16.83	17.19	16.88	61.37

Table E.7: Regression of GDP growth, at nine months ahead, on text-based uncertainty measures. Asterisks denote p-values; 1%: ***, 5%: **, 10%: *.

each these is indexed by μ), take

$$\mathcal{I}^e = \bigcup_{\mu} \left\{ f_\mu \left(T^e_\mu (I^e_\mu (X)) \right) \right\}_{t=(\mu-1)s+\alpha}^{t=\mu s-1+\alpha}$$

This takes, for each possible value of t, the model prediction with the index label that has the highest possible value of μ. The final test, or out-of-sample, set that we use is constructed similarly: for each possible value of t, it is the model prediction with the lowest possible value of μ:

$$\mathcal{O}^e = \bigcup_{\mu} \left\{ f_\mu \left(T^e_\mu (O^e_\mu (X)) \right) \right\}_{t=\mu s+\alpha}^{t=(\mu+1)s-1+\alpha}$$

Equivalently, the in-sample and out-of-sample sets are composed of the last step of each training window indexed by μ, and the first step of each test set indexed by μ.

F.2 Rolling window

A window of size $\alpha + s$ is used to estimate the model.

$$I^r_\mu(\vec{z}) = \left\{ z_t \right\}_{t=(\mu-1)\cdot s}^{t=\mu\cdot s+\alpha-1}$$

$$O^r_\mu(\vec{z}) = \left\{ z_t \right\}_{t=\mu\cdot s+\alpha}^{t=T}$$

$$T^r_{\mu k}(\vec{z}) = T^r_{\mu k}(\vec{z}; I^r_\mu(\vec{x}_k)) = \frac{\vec{z} - \langle I^r_\mu(\vec{x}_k)\rangle}{\sigma_{I^r_\mu(\vec{x}_k)}}$$

The unified, one-step ahead dataset is created from

$$\mathcal{I}^r = \bigcup_\mu \left\{ f_\mu \left(T^r_\mu(I^r_\mu(X)) \right) \right\}_{t=(\mu-1)s+\alpha}^{t=\mu s-1+\alpha}$$

and

$$\mathcal{O}^r = \bigcup_\mu \left\{ f_\mu \left(T^r_\mu(O^r_\mu(X)) \right) \right\}_{t=\mu s+\alpha}^{t=(\mu+1)s-1+\alpha}$$

Note that, because the global transformations depend on the training data, $\mathcal{I}^r \neq \mathcal{I}^e$ and $\mathcal{O}^r \neq \mathcal{O}^e$.

G Algorithm-based text metrics – further forecast results

This section present further results related to §6.

G.1 Performance versus an AR(1) model benchmark

For the case in which the benchmark model for the algorithmic text based metrics is an AR(1), we run a Diebold-Mariano test to check whether the results are statistically distinguishable from forecasts with the benchmark model. In the table, we show only those forecasts for which at least one target-metric combination per newspaper had a statistically significantly smaller RMSE than the benchmark model and we look at $h = 9$. We find statistically significant results across newspapers, although *The Guardian* does more poorly than the other two. The most consistent pattern of forecast performance is across targets. Although the gains for CPI look small in Figure 5, Table G.8 makes it clear that they are significant for some combinations of newspaper, metric, and target.

G.2 Performance versus a factor model benchmark

In Table G.9 we present results from a Diebold-Mariano test for a model including a text metric, two factors, and an AR(1) versus the same model without the text metric at $h = 9$. Combinations of targets and metrics that were not statistically significant with the simpler AR(1) model do reach statistical significance in this test, somewhat counter-intuitively. However, this is not an unusual finding. Several studies drawn from the forecasting literature suggest that univariate time series models have better forecasting power than richer models, especially for macroeconomic time series (Chauvet and Potter, 2013; Faust and Wright, 2013). Including more information does not necessarily improve forecasts at a horizon longer than one quarter. In particular, Carriero, Galvão and Kapetanios (2018) show that the choice of the best forecasting model class may vary with the forecast horizon.

Which targets can be forecast better using text are somewhat consistent with the AR(1) model: business investment, CPI, and household consumption feature heavily. In the case of the factor model, we also see significant results for unemployment, and less strong results for GDP. As GDP is a composite measure, and the factors are designed to track many variables that go into its construction, this is unsurprising.

Paper	Metric	Target Horizon	Business Investment	CPI	Fin stab index	GDP	Hhld Consumption	IMF fin cond	IOP	IOS	Unemployment
The Daily Mail	Afinn	6	-1.70*								
		9	-1.72*								
	Alexopoulos	9	-1.71*								
	Counts economy	9					-1.91*				
	Harvard	9	-1.66*								
	Husted	9							-5.89***		
	Loughran	3			-1.86*						
		6				-1.78*					
		9	-1.86*			-1.73*	-1.78*				
	Nyman	3			-2.10**	-1.67*		-1.74*		-1.95*	
		6				-1.66*					
		9				-1.76*					
	Opinion	3			-2.15**						
		6	-1.76*			-1.72*					
		9	-1.69*			-1.67*					
	Stability	3			-2.00**	-1.68*					
		6				-1.69*					
		9	-1.67*			-1.67*					
	TFIDF economy	9				-1.67*	-1.99**				
	Vader	3									-1.72*
The Daily Mirror	Afinn	9		-2.12**							
	Harvard	9		-2.10**							
	Loughran	6		-1.69*							
		9		-1.69*							
	Nyman	3		-2.08**							
		6		-2.32**							
		9		-1.78*							
	Opinion	9		-3.11***							
	Punctuation economy	3									-3.00***
	Stability	9					-1.83*				
	TFIDF economy	3				-1.81*					
	TFIDF uncert	6	-1.98**								
	Vader	6		-1.86*							
		9		-2.19**							
The Guardian	Alexopoulos	3									-1.93*
	Counts economy	3									-1.82*
		6									-1.81*
		9									-1.66*
	Counts uncert	6			-1.65*						
	Loughran	3									-1.95*
	Punctuation economy	3									-12.57***
	Stability	3	-1.65*			-1.72*					-2.11**
		9						-2.10**			
	TFIDF economy	3									-1.97**
		6									-1.94*
		9									-1.70*

Table G.8: Results from a Diebold-Mariano test of an OLS-AR(1) model with text metrics versus an AR(1) model without them (the benchmark). Statistically significant differences in RMSE are shown. *, **, *** denote rejection of the null, of no difference in RMSE relative to the benchmark model, at the 10%, 5%, 1% levels respectively. Only those targets for which at least one metric-newspaper pair had a p-value of less than 10% are included.

H Machine learning models

Here we present the specifications of the machine learning models and their hyperparameters. Throughout, let $\{x_{tk}\}_{t=0}^{t=T}$ represent feature k, also denoted $\vec{x}_k$, and the entire set of features of all time form a matrix X. The time series of the target variable is denoted $\vec{y}$. Define

$$\|\beta\|_p = \left(\sum_{k=1}^{K} |\beta_k|^p \right)^{1/p}$$

as the ℓ^p norm.

Paper	Metric	Target Horizon	Business Investment	CPI	Fin stab index	GDP	Hhld Consumption	IMF fin cond	Unemployment
The Daily Mail	Afinn	6							-1.68*
		9					-2.04**		-1.95*
	Alexopoulos	9					-2.88***		
	Baker-Bloom-Davis	9					-4.02***		-1.90*
	Counts uncert	9		-2.16**					
	Harvard	9	-1.81*						
	Husted	9		-2.15**			-3.14***		
	Loughran	9							-2.12**
	Nyman	3			-1.82*			-1.68*	
		9					-2.41**		
	Opinion	3			-2.09**				
		6			-2.00**				-1.77*
		9					-1.97**		-2.01**
	Stability	3			-1.69*				
		6							-1.88*
		9					-1.95*		-2.18**
The Daily Mirror	Afinn	6							-1.80*
		9		-1.79*					-2.07**
	Baker-Bloom-Davis	9							-1.81*
	Harvard	6					-1.70*		
		9		-1.84*					
	Loughran	9					-1.69*		
	Opinion	9							-1.92*
	Vader	9		-1.76*		-2.12**	-1.87*		-1.95*
The Guardian	Afinn	3	-1.68*						
		9							-1.67*
	Alexopoulos	3							-1.75*
		6							-2.47**
		9		-1.99**					-2.45**
	Baker-Bloom-Davis	6							-2.11**
		9							-2.32**
	Counts uncert	9		-1.81*					-1.71*
	Harvard	3	-1.74*						
	Loughran	3	-1.75*						
		9							-1.67*
	Nyman	3	-1.73*						
	Opinion	3	-1.72*						
	Punctuation economy	9		-1.70*			-1.95*		
	Stability	3	-1.91*						
		9							-1.69*
	TFIDF uncert	9			-2.20**				-1.73*

Table G.9: Results from a Diebold-Mariano test on the factor model with algorithm-based text metrics. Statistically significant differences in RMSE are shown. *, **, *** denote rejection of the null, of no difference in RMSE relative to an AR(1) and factors (the benchmark model), at the 10%, 5%, 1% levels respectively. Only those targets for which at least one metric-newspaper pair had a p-value of less than 10% are included.

H.1 Lasso

The least absolute shrinkage and selection operator solves

$$\min_{\beta} \left\{ \frac{1}{T} \|y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq \kappa$$

with $\kappa = 1$.

H.2 Ridge

Ridge regression solves

$$\min_{\beta} \left\{ \|y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_2^2 \leq \kappa$$

with $\kappa = 1$.

H.3 Elastic net

Elastic net regression solves

$$\min_{\beta} \left\{ \|y - X\beta\|_2^2 \right\} \text{ subject to } \alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \leq \kappa$$

with $\alpha = 0.5$ and $\kappa = 1$.

H.4 Support vector regression

Support vector machine regression solves

$$\min_{w,b,\vec{\xi},\vec{\xi^*}} \frac{1}{2}\|w\|^2 + C\sum_t \xi_t + C\sum_t \xi_t^*$$

subject to

$$\begin{aligned} y_t - \vec{w}^\intercal \phi(\vec{x}_t) - b &\leq \epsilon + \xi_t^*, \\ \vec{w}^\intercal \phi(\vec{x}_t) + b - y_t &\leq \epsilon + \xi_t, \\ \xi_t^*, \xi_t &\geq 0 \quad \forall \quad t \end{aligned}$$

where $K(\vec{x}_t, \vec{x}_{t'}) = \phi(\vec{x}_t)^\intercal \phi(\vec{x}_{t'})$ is a kernel function. We use $\epsilon = 0$, $C = 800$, and choose the radial basis function as our kernel.

H.5 Artificial Neural Network

We use a multilayer perceptron that minimises the squared error loss. We use two hidden layers, tanh as our activation function, and an ℓ^2 penalty of 2000. To solve for the weights, we use the lbfgs solver.

H.6 Random forest

We use a bootstrapped random forecast regressor with 200 trees, a max depth of 8, and a minimum sample split of 2.

I Machine learning and text models – further forecast results

This section present further results related to §7.

I.1 OLS AR(1) benchmark

In Table I.10 we present results from a Diebold-Mariano test for a machine learning model including text features and an AR(1) versus AR(1) OLS without the text.

I.2 ML-AR(1) benchmark

In Figure I.1 we present results for a machine learning model including text features and an AR(1) versus the same machine learning model without text. In Table I.11 we present results from a Diebold-Mariano test for the same specification.

I.3 ML-factor model and AR(1) benchmark

In Figure I.2 we present results for a machine learning model including text features, an AR(1) and factors versus the same machine learning model without text. In Table I.12 we present results from a Diebold-Mariano test for the same specification.

Paper	Target Model	Business Investment	CPI	GDP	Hhld Consumption	IMF fin cond	IOP	Unemployment
The Daily Mail	Elastic	-2.25**						
	Forest	-2.48**	-1.69*		-2.03**		-1.82*	
	Lasso	-1.96*						
	NN	-2.44**	-2.54**	-1.72*	-1.83*	-1.75*		-2.02**
	Ridge	-2.55**	-1.88*		-1.81*		-1.70*	-2.07**
	SVM		-2.42**		-1.67*			-1.71*
The Daily Mirror	Elastic	-1.84*						
	Forest	-2.16**	-1.78*		-1.75*		-1.72*	
	Lasso	-1.82*						
	NN	-2.29**	-2.75***	-1.72*				-1.87*
	Ridge	-2.40**			-1.72*			-1.72*
	SVM		-1.99**					
The Guardian	Elastic	-1.69*						
	Forest				-1.93*			
	NN		-2.77***		-1.88*			-2.31**
	Ridge		-2.23**		-1.93*	-1.67*		-1.99**
	SVM		-2.69***					

Table I.10: Results from a Diebold-Mariano test on forecasts using term frequency vectors with an AR(1) versus an AR(1) alone using OLS. Statistically significant differences in RMSE are shown. *, **, *** denote rejection of the null, of no difference in RMSE relative to the OLS AR(1), at the 10%, 5%, 1% levels respectively. Only those targets for which at least one of the machine learning models had a p-value of less than 10% are shown.

J Breakdown of forecast performance through time

The breakdown of differences in squared error between OLS with only an AR(1) term and OLS with text metrics and an AR(1) are shown in Figure J.3 and denoted by $\epsilon^2_{\text{Bench.}} - \epsilon^2_{\text{Text}}$. When the lines are above zero, the model with text is performing better than the model without. This shows that most of the improvement in performance comes from stressed periods.

Paper	Model	Target Horizon	Business Investment	CPI	Fin stab index	GDP	Hhld Consumption	IMF fin cond	IOP	IOS	Unemployment
The Daily Mail	Forest	6	-1.80*							-2.16**	
		9	-2.21**			-1.66*	-2.29**		-1.76*	-1.70*	
	NN	3	-2.52**	-1.87*	-2.24**	-2.11**	-2.22**		-1.80*	-1.95*	
		6		-1.92*	-1.73*	-1.66*	-1.66*				-1.82*
		9		-2.30**	-1.74*	-1.79*	-2.20**		-1.77*	-1.78*	
	Ridge	3		-1.94*	-1.93*	-2.18**	-2.42**	-1.75*	-2.12**	-3.37***	-1.90*
		6	-1.98**	-2.22**	-1.78*	-2.47**	-2.01**	-2.11**	-1.73*	-3.61***	-2.30**
		9	-2.20**	-2.31**		-3.14***	-2.09**	-1.71*		-4.23***	-2.52**
	SVM	3	-2.01**			-2.21**	-2.46**	-1.72*	-3.54***	-2.86***	
		6	-3.79***	-1.86*					-3.20***	-3.46***	
		9	-3.96***	-1.67*			-1.69*	-2.27**	-2.30**	-2.85***	
The Daily Mirror	Elastic	9									-1.68*
	Forest	6								-2.07**	
		9	-1.99**				-2.31**		-1.78*		
	NN	3	-2.36**		-1.97*	-1.72*	-1.88*		-2.05**	-2.33**	
		6	-2.40**	-2.14**	-1.98**		-2.12**			-2.04**	
		9	-1.79*	-2.24**	-1.76*	-1.75*	-2.03**		-1.66*		
	Ridge	3		-1.71*	-1.75*	-1.69*			-1.78*	-2.24**	-2.42**
		6	-1.79*	-2.34**	-1.72*	-2.44**	-2.03**			-3.09***	-2.54**
		9	-2.23**	-2.20**		-4.33***	-2.12**	-1.67*		-3.26***	-2.59**
	SVM	3	-5.68***			-2.19**	-2.23**		-3.64***	-2.87***	
		6	-3.78***	-1.90*					-2.80***	-3.47***	
		9	-4.07***					-2.34**	-2.37**	-2.51**	
The Guardian	Elastic	9		-1.72*			-1.91*				
	Forest	6					-1.70*			-2.28**	
		9				-1.76*	-2.34**		-1.85*	-1.83*	
	Lasso	6					-1.67*				
		9					-1.69*				
	NN	3	-1.85*		-1.92*	-2.06**	-2.00**		-1.71*	-1.92*	-3.09***
		6		-1.71*	-1.94*	-1.67*	-1.98**			-1.78*	-1.94*
		9		-2.35**		-1.82*	-2.20**		-1.69*	-1.79*	-1.92*
	Ridge	3		-2.47**		-2.06**	-2.01**	-1.84*	-1.87*	-2.95***	-3.11***
		6		-2.39**		-2.58**	-1.94*	-1.78*	-1.72*	-4.12***	-2.76***
		9		-2.24**		-3.16***	-2.11**			-4.29***	-2.69***
	SVM	3				-2.24**	-2.77***		-3.55***	-2.62***	-1.91*
		6	-2.12**	-1.80*			-1.86*		-3.26***	-3.73***	-1.79*
		9		-1.97*				-2.90***	-2.38**	-3.08***	-1.96*

Table I.11: Results from a Diebold-Mariano test on forecasts using term frequency vectors with an AR(1) versus an AR(1) alone with the same machine learning model. Statistically significant differences in RMSE are shown. *, **, *** denote rejection of the null, of no difference in RMSE relative to the benchmark model, at the 10%, 5%, 1% levels respectively. Only those targets for which at least one of the machine learning models had a p-value of less than 10% are shown.

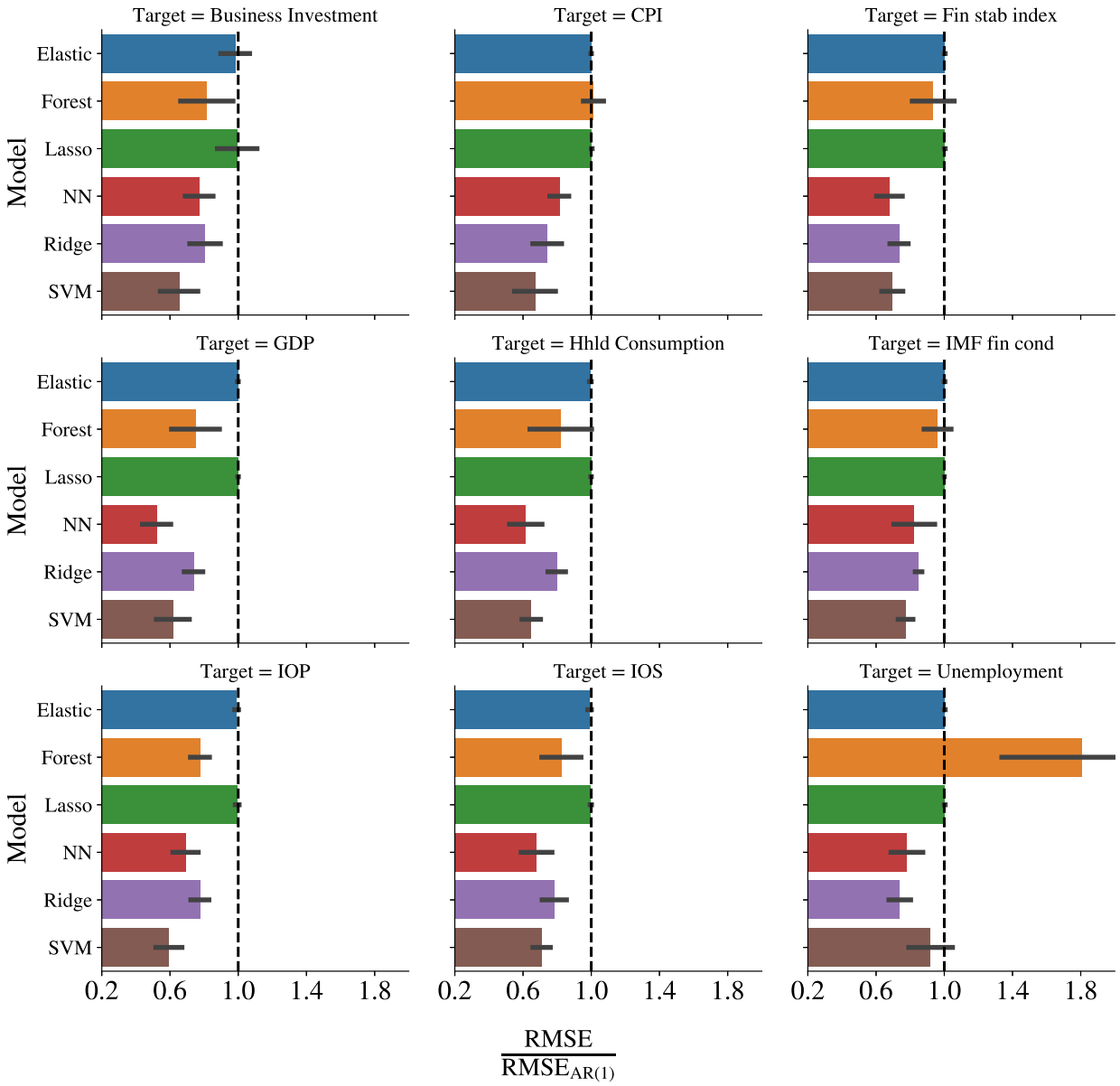

Figure I.1: RMSEs relative to a benchmark AR(1) by machine learning model and target variable. The same machine learning model (with the same hyperparameter settings) is used with text and without. Bars to the left of the dashed line indicate an improvement in forecast performance conducted with the given text metric. The confidence intervals are standard deviations over both the newspapers and the different horizons (3, 6 and 9 months ahead).

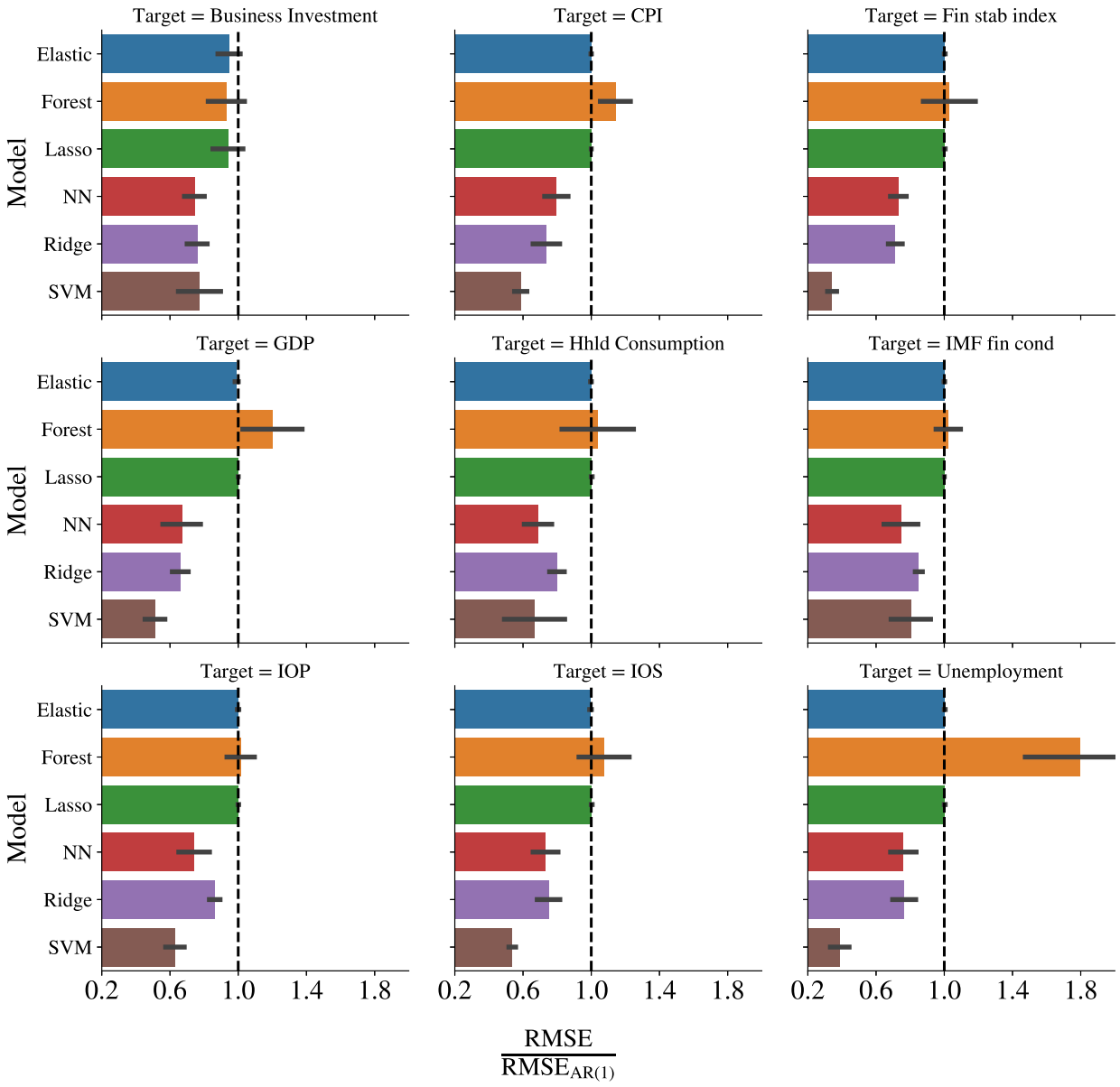

Figure I.2: The relative improvement in root mean square error of a machine learning model that uses text, an AR(1) term, and factors versus the same machine learning model with the AR(1) and factors but no text. The facets are different target variables. Bars to the left of the dashed line indicate an improvement in forecast performance conducted with the given text metric. The confidence intervals are standard deviations over both the newspapers and the different horizons (3, 6 and 9 months ahead).

Paper	Model	Target Horizon	Business Investment	CPI	Fin stab index	GDP	Hhld Consumption	IMF fin cond	IOP	IOS	Unemployment
The Daily Mail	Elastic	3				-1.81*					
	Lasso	3								-2.72***	
	NN	3	-2.53**	-2.74***	-2.48**	-1.86*	-2.92***	-1.85*	-2.15**	-3.18***	-3.21***
		6	-2.01**	-2.67***		-2.39**	-2.73***				-2.74***
		9		-2.06**		-2.23**	-2.67***	-1.73*	-1.81*	-2.40**	-2.32**
	Ridge	3	-2.42**	-2.02**	-1.88*	-3.20***	-2.42**	-1.93*	-2.16**	-3.34***	-2.25**
		6	-2.22**	-2.22**		-3.12***	-2.65***	-1.85*		-2.99***	-3.72***
		9	-2.34**	-2.26**		-3.21***	-2.81***	-1.73*		-3.49***	-3.82***
	SVM	3	-3.52***	-2.95***	-3.95***	-3.14***	-4.99***	-1.74*	-3.80***	-3.42***	-5.01***
		6	-2.13**	-2.28**	-2.28**	-3.54***	-3.85***		-4.08***	-3.65***	-2.84***
		9	-1.68*	-3.52***	-2.26**	-3.66***			-2.58**	-2.51**	-2.22**
The Daily Mirror	Elastic	3	-1.94*					-1.86*			
	Forest	9					-2.04**				
	NN	3	-2.25**	-2.17**	-2.59**	-2.48**	-2.27**			-2.50**	-3.61***
		6	-2.00**		-1.83*	-2.20**	-2.77***		-2.20**	-3.09***	-2.32**
		9	-2.14**	-2.25**		-1.73*	-2.75***	-1.79*		-2.16**	-2.31**
	Ridge	3	-2.79***	-1.79*		-2.96***	-1.72*	-1.92*	-1.71*	-3.18***	-2.53**
		6	-2.19**	-2.44**		-2.86***	-2.20**			-3.34***	-3.45***
		9	-1.89*	-2.16**		-2.94***	-2.46**			-3.49***	-2.86***
	SVM	3	-2.56**	-2.92***	-3.90***	-3.11***	-5.55***		-3.62***	-3.07***	-4.89***
		6	-1.81*	-1.81*	-2.40**	-3.60***	-3.74***		-4.02***	-3.59***	-2.68***
		9		-3.47***	-2.31**	-4.17***			-2.32**	-2.23**	-2.09**
The Guardian	Elastic	3				-2.01**					
		6								-1.72*	
		9				-1.83*				-1.79*	
	NN	3		-1.77*	-2.17**	-2.12**	-2.70***				-2.82***
		6		-2.63***		-3.01***	-2.41**			-2.28**	-1.70*
		9		-2.34**		-2.10**	-3.53***		-1.83*	-2.67***	-2.63***
	Ridge	3		-1.93*		-3.51***	-2.09**	-2.02**	-1.87*	-3.70***	-2.64***
		6		-2.33**		-3.25***	-2.68***			-3.26***	-3.44***
		9		-2.18**		-3.12***	-2.78***			-3.24***	-3.68***
	SVM	3		-3.86***	-3.98***	-3.18***	-3.72***	-1.84*	-3.74***	-3.75***	-6.12***
		6		-2.55**	-2.26**	-3.74***	-4.05***		-4.23***	-4.13***	-3.22***
		9		-3.89***	-2.62***	-4.12***			-2.54**	-2.59**	-2.46**

Table I.12: Results from a Diebold-Mariano test on forecasts using term frequency vectors with an AR(1) and factors versus an AR(1) and factors without text using the same machine learning model. Statistically significant differences in RMSE are shown. *, **, *** denote rejection of the null, of no difference in RMSE relative to the benchmark model, at the 10%, 5%, 1% levels respectively. Only those targets for which at least one of the machine learning models had a p-value of less than 10% are shown.

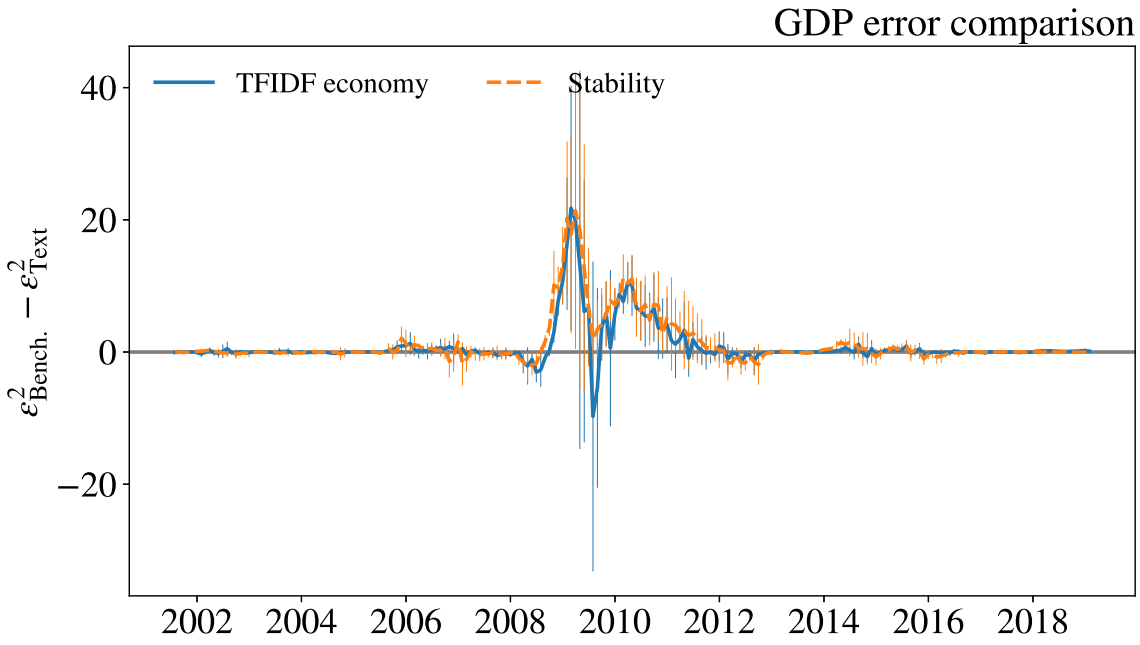

Figure J.3: Mean squared error differences between a benchmark model and text over the time-dependent union of h-month ahead out-of-sample forecasts, with horizon $h = 3, 6, 9$. The target variable is monthly GDP. The benchmark is an OLS AR(1) model. The plotted error bars are standard deviations over the different horizons and newspapers. A solid line above zero means that the model with text produces smaller errors than the benchmark model. Two of the best all round performing text metrics are shown. The majority of the forecast gains are during the crisis.