

# Forecasting the IBEX-35 stock index using deep learning and news emotions\*

Sergio Consoli<sup>1</sup>, Matteo Negri<sup>2</sup>, Amirhossein Tebbifakhr<sup>2</sup>, Elisa Tosetti<sup>3</sup>, and Marco Turchi<sup>2</sup>

<sup>1</sup> European Commission, Joint Research Centre (JRC), Ispra, Italy.  
[name.surname]@ec.europa.eu

<sup>2</sup> Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy.  
{atebbifakhr,negri,turchi}@fbk.eu

<sup>3</sup> Università Ca' Foscari Venezia, Department of Management, Cannaregio 873, Fondamenta San Giobbe, 30121 Venezia, Italy. elisa.tosetti@unive.it

\*Authors listed in alphabetic order.

**Abstract.** Measuring the informational content of text in economic and financial news is useful for market participants to adjust their perception and expectations on the dynamics of financial markets. In this work, we adopt a neural machine translation and deep learning approach to extract the emotional content of economic and financial news from Spanish journals. To this end, we exploit a dataset of over 14 million articles published in Spanish newspapers over the period from 1st of July 1996 until 31st of December 2019. We then examine the role of these news-based emotions indicators in forecasting the Spanish IBEX-35 stock market index by using DeepAR, an advanced neural forecasting method based on auto-regressive Recurrent Neural Networks operating in a probabilistic setting. The aim is to evaluate if the combination of a richer information set including the emotional content of economic and financial news with state-of-the-art machine learning can help in such a challenging prediction task. The DeepAR model is trained by adopting a rolling-window approach and employed to produce point and density forecasts. Results look promising, showing an improvement in the IBEX-35 index fitting when the emotional variables are included in the model.

**Keywords:** IBEX-35 stock index; Deep Learning; Neural forecasting; DeepAR; Machine translation; Emotion classification; News analysis.

## 1 Introduction

Forecasting economic and financial variables is a challenging task for several reasons including, among others, the effect of volatility, regime changes, and low signal-to-noise ratio [15]. In addition, modern economies are subject to numerous shocks that make such prediction task extremely hard, particularly during times of economic turmoil like the ones we are currently experiencing with the Covid-19 pandemic [50]. In this context, the incorporation in forecasting models of economic and financial information coming from news media, like in particular emotions and sentiment, has already demonstrated great potentials [1, 23, 8, 33, 4, 42, 49].

Our endeavour is to explore the predictive power of news emotions for forecasting economic and financial time series by leveraging on the recent advances in

the literature on deep learning [31, 9, 39, 38]. We believe that news are a promising forecasting tool since they describe recent economic events and trends, represent the updated expectations of market participants about the future, thus significantly influencing investors' perception and decisions.

In this work, we first extract sentences referring to specific economic and financial aspects from Spanish news, over a period of time ranging from July 1996 to December 2019<sup>4</sup>. We then use a neural machine translation (NMT) [43] approach based on deep learning for classifying sentences into groups, according to the Ekman's six basic emotions [19]: *fear*, *anger*, *joy*, *sadness*, *disgust*, *surprise*. The method leverages on the recent NMT approach by Tebbifakhr et al. [44] used for sentiment classification, readapting it to the massive annotation task of Spanish news with emotion labels. This approach originally adapts a NMT system, which is trained for general translation purposes, to produce translation that is easier to process by a downstream classifier in the target language. In our case, we use this methodology to adapt a Spanish-to-English NMT system targeting an emotion classifier in English, fine-tuned on the economic and financial domain. This approach allows us first to translate the Spanish news into English and then to annotate them using the English classifier. The advantage of translating before annotating is that the English classifier can be trained on a large amount of available annotated data. Although in this work we focus on the IBEX-35 index and Spanish news [14], the methodology is easily generalizable to other languages, and portable to other domains and evaluation scenarios.

We adopt this approach to extract a set of emotion indicators from Spanish economic and financial news and explore whether they provide useful predictive signals to improve the accuracy of existing forecasting models for financial variables [20]. Specifically, we focus on the Spanish IBEX-35 stock index, aiming at predicting its daily fluctuations by using as inputs the past time series values along with the daily emotion indicators.

The forecasting methodology employed for this task is *DeepAR*, a recent neural forecasting methodology proposed by Salinas et al. [39] that builds upon previous work on deep learning for time series data [31, 9, 30]. This powerful approach produces accurate probabilistic forecasts, based on training an auto-regressive Recurrent Neural Network (RNN) model with feedback connections on a given number of related time series, which in our case are the emotion signals. The aim is to disentangle the improvement in the forecasting power due to the inclusion of the emotions indicators extracted from news within the DeepAR approach. Since our forecasting method calculates the probability attached to each forecast, the output can help investors in their decision making according to their individual risk tolerance.

Our results show that the emotion indicators extracted from news, used in combination with DeepAR, improve the performance of forecasting for the analysis of the

---

<sup>4</sup> We focus on Spain since it gives, from our commercial provided, the largest coverage of news relative to other EU countries. The largest covered period is indeed the used one, that is from July 1996 to December 2019.

IBEX-35 stock index. The method is also benchmarked against two traditional approaches (simple moving average and a naïve method for random forecasts). Overall, the obtained results look promising. We believe that the combination of the cutting-edge technologies used in our approach has high potential for the implementation of effective solutions for the prediction of other economic and financial indexes.

## 2 Background

Recent literature has pointed at the important role of financial investor’s sentiment and emotions in anticipating interest rates dynamics [45, 34]. News articles, in particular, represent a relevant data source to model economic and financial variables, and several studies have already exploited this additional source of information. A seminal work that has used a sentiment variable calculated on news articles from the Wall Street Journal is that by Tetlock [45]. The author shows that high levels of pessimism are a relevant predictor of convergence of the stock prices towards their fundamental values. Several other papers have tried to understand the role that news play in predicting, for instance, company news announcements, stock returns and volatility. Agrawal et al. [2] and Dridi et al. [17] have recently used emotions extracted from social media, financial microblogs, and news to improve predictions of the stock market. Hansen and McMahon [26] has looked at the emotional content of the Federal Reserve statements and the effects of these statements on the future evolution of monetary policy. Other papers ([46], [47] and [41] among others) have classified articles in topics and extracted emotional signals that showed to have a predictive power for measures of economic activity, such as GDP, unemployment and inflation [25]. These results have shown the high potential of emotions extracted from news on monitoring and improving the forecasts of economic developments [17].

Studies adopting machine learning approaches to forecast financial indexes have also increased exponentially in recent years (see, e.g. [15, 7, 24, 37, 22, 10]). Among these, several studies have successfully adopted various versions of deep learning [31, 9, 39]. For a survey on the use of deep learning methods for time series forecasting, the reader is referred, among others, to [9, 28, 38, 30]. Recently, in [39] the authors proposed DeepAR, an RNN-based forecasting model using Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) cells, the latter being a simplification of LSTMs that does not use a separate memory cell and may result in good performance for certain applications. At each time step, DeepAR takes as input the previous time points and covariates, and estimates the distribution of the values of the next period. This is done via the estimation of the parameters of a pre-selected distribution. Training and prediction follow the general approach for auto-regressive models [39].

A common trait of the other machine learning approaches employed to forecast financial and economic time series is that they are generally constrained to a point forecasting setting [29, 28, 21]. Differently, the approach described in the following section focuses on the full predictive distribution, not just a single best realization, making the analysis more robust and reducing uncertainty in the downstream

decision-making flow. In addition, a large part of existing studies that predict stock market indices has addressed the forecasting problem as a classification task, where the goal is predicting the direction of the stock index rather than its actual variations (see e.g. [51, 18, 40, 12]). Indeed, forecasting the variations of the stock index directly, as we aim in this work, is an extremely challenging task given that the series behaves similarly to a random walk process. All these characteristics make our approach very innovative.

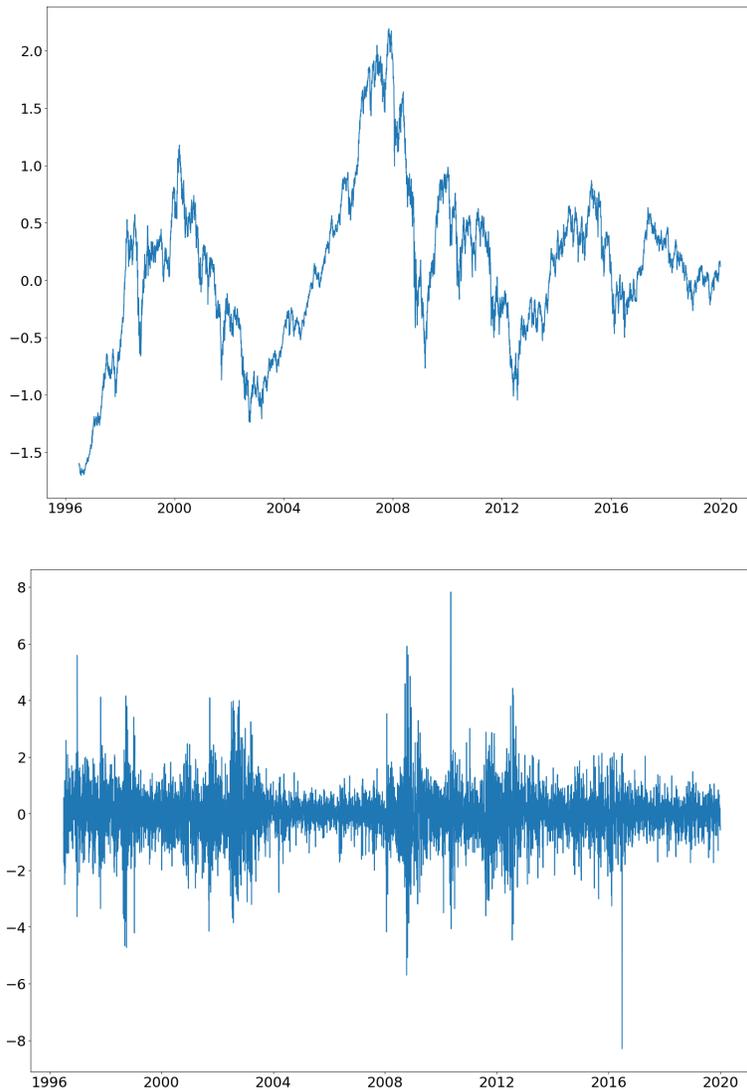


Fig. 1: IBEX-35 stock index (top), and the daily fluctuations of the IBEX-35 index defined as the log-difference between consecutive observations (bottom).

### 3 Data

The source of economic news is obtained from a commercial provider.<sup>5</sup> The dataset consists of around 14 million articles, full-text, from July 1, 1996 until December 31, 2019. Their source is Spanish newspapers, selected so as to achieve a good national as well as regional coverage.<sup>6</sup> We extract sentences referring to specific economic and financial aspects, by using a keyword-based information extraction procedure with search keywords broadly related to the Spanish economy, monetary and fiscal policies.<sup>7</sup> In order to filter out only sentences referring to Spain, we also use a location detection heuristic [6] assigning the location to which a sentence is referring as its most frequent named-entity location detected in the news text, and then selecting only sentences with specific assigned location labels related to Spain.<sup>7</sup> With this procedure we obtain a total of over 4.2 million sentences.

We obtain the IBEX-35 index from Yahoo Finance.<sup>8</sup> This is a free-float, capitalization-weighted stock market index that tracks the performance of the 35 most liquid stocks traded on the continuous market on the Bolsa de Madrid. For our study, we consider the close price of the index, that is the price of the stock at the closing time of the market on a given day. Being the IBEX-35 a highly persistent and non-stationary index (see Figure 1, top), we have applied a log-difference transformation to obtain a stationary series of daily changes representing our prediction target shown (Figure 1, bottom). Missing data related to weekends have been dropped from the target time series, giving a final number of 5,950 data points for the time period of interest.

## 4 Methods

### 4.1 Machine Translation

We classify the 4.2 million according to the Ekman’s basic emotions [19] using a recent NMT approach [44], adapted from a sentiment to an emotion classification setting [1]. The lack of emotion-annotated data in languages different from English requires to first translate the Spanish sentences into English using NMT, and then to annotate them with an English emotion classifier.

NMT systems [43] are based on neural technology that requires large quantities of training data, from which they learn how to translate text in the source into text in the target language. The training data consists of “parallel” sentence pairs, where the source and the target sentences are one the translation of the other.

<sup>5</sup> Dow Jones Data, News and Analytics Platform: <https://www.dowjones.com/dna/>

<sup>6</sup> *El Mundo, ABC, Expansión, La Vanguardia, Cinco Días, El País, Actualidad Económica, Agencia EFE - Servicio Económico, Alimarket, Aseguranza, El Comercio, Córdoba, El Correo, El Diario de León, Diario Montañés, El Diario Vasco, Europa Press, Hoy, Ideal, El Periódico, Las Provincias, La Rioja, Sur, La Verdad, La Voz de Galicia.*

<sup>7</sup> The complete list includes around 300 terms, such as *inflation, consumer prices, bankruptcy, economic volatility, housing market, competitiveness, debt, employment, bubble bust, bond market*, etc, and can be obtained upon request from the authors.

<sup>8</sup> <https://finance.yahoo.com/quote/%5EIBEX>

During training, the NMT system iterates over the data for a number of epochs until the performance measured in terms of training loss (e.g. cross-entropy) reaches a plateau. The “MT for machines” paradigm proposed in [44] represents a variant of this general approach, which is oriented to generating automatic translations that, instead of targeting human readers, should be easy to process by NLP tools. The underlying idea is to orient NMT adaptation towards “machine-oriented” criteria (i.e. maximising the performance of a downstream NLP component fed with the NMT output) rather than the standard “human-oriented” quality criteria of output fluency and semantic adequacy. This is done by applying Reinforcement Learning techniques [15, 44], in which NMT model optimization is driven by rewards collected in terms of final performance in the downstream task of interest (in our case, emotion classification).

It is known that the larger the training set, the better the capability of the model to translate and generalize to unseen inputs [43]. For our work, we first train a generic Spanish to English NMT system, built using the freely available parallel data (around 84M parallel sentences) present in the OPUS website,<sup>9</sup> tokenized and encoded using 32K byte-pair codes. To develop the NMT system, we use the Transformer architecture by Vaswani et al. [48] with the original parameter settings. We perform the training step until convergence, that is until we reach a plateau in the computed loss. For training the NMT system, we use the OpenNMT-tf (v2.0.1) toolkit.<sup>10</sup> We also embed a pre-trained Spanish BERT model<sup>11</sup> inside the NMT system [52, 13] to give a better representation of the source Spanish text during translation; this experimental choice has shown to be advantageous for our task.

We then adapt the generic Spanish-English NMT model to the economic domain by fine-tuning it on parallel sentences derived by the news, using two approaches:

1. *Back-translating the English in-domain data.* We automatically translate the English economic sentences back into Spanish. Doing so, we generate translation pairs in which the Spanish source side is the output of the NMT system and the English target side contain human-quality sentences. The Spanish-English sentence pairs are then used to adapt the NMT system.

2. *Selecting in-domain translation pairs from the training data.* We train a language model using the English economic sentences and use it to rank the English side of the 84M parallel sentences from OPUS. A higher rank means, for a given sentence, a higher similarity to the economic sentences. To have a fair comparison, from the top of this parallel ranked list, we select an amount of data equal to the original economic sentences in terms of number of tokens on the English side. The Spanish-English selected sentence pairs are then used to adapt the NMT system.

The fine-tuning [31, 15] is performed continuing the training of the generic model using the economic parallel data created with the two methods, considering them both together and in isolation.

---

<sup>9</sup> OPUS, the open parallel corpus. Available at: <https://opus.nlpl.eu/>

<sup>10</sup> <https://github.com/OpenNMT/OpenNMT-tf>

<sup>11</sup> BETO - available at: <https://github.com/dccuchile/beto>

## 4.2 Emotion classification

The English classifier is based on a BERT language model [16] adapted to our classification problem. Recent works on sentence classification have shown that neural language models can be used to transform a textual sentence into a special token vector (CLS) that represents the input sentence. This vector can be used for any sentence-level classification task, including emotion detection. To build the classifier, the language model needs to be trained on a large unsupervised corpus, then embedded in a neural network that transforms the token vector in one of the desired categories. We use in particular one fully-connected neural layer, which maps the CLS token of the English BERT into the emotion classes. We train the resulting architecture until convergence on a freely available database with English sentences annotated with emotion labels.<sup>12</sup>

This general model is then further fine-tuned for our emotion detection task by re-training it using a set of 5,100 Spanish economic sentences annotated with the Ekman’s six basic emotions [19] by 8 different annotators<sup>13</sup> and translated into English using the European Commission *eTranslation* service.<sup>14</sup> For the annotation task we use the commercial Amazon AWS SageMaker service.<sup>15</sup> In order to realize a unique label for each sample, the following steps are followed: (i) For each sample, we detect the label with the highest vote. In case of equal votes between classes, the priority is given to the least represented class in the dataset; (ii) We select a positive threshold equal to 3 and assign the corresponding label to all the samples with the identified vote higher than this threshold; (iii) The remaining samples are discarded. Given that the number of “Disgust” samples obtained with this procedure resulted to be very low (only 26 samples), this class has been removed from the dataset. We then split the labelled data into training, development and test sets with the following proportions: 0.6, 0.1, and 0.3. Overall, the classifier is able to reach an average F1 score on test of nearly 70%, which represents quite an acceptable classification performance for our final task.

For the final massive emotion annotation of the 4.2M sentences selected from the Spanish news, we then use the developed NMT system for translating them into English, and then the English classifier to annotate the translated sentences with the considered emotions. The final distribution of emotion labels in the annotated Spanish sentences is: *surprise* = 25%, *joy* = 19%, *sadness* = 13%, *anger* = 10%,

---

<sup>12</sup> The annotated English corpus is taken from the *unified emotion datasets* by Bostan and Klinger [11], available at: <https://github.com/sarnthil/unify-emotion-datasets>. It is a large English collection of different emotion datasets [11], mapped to a unified set of emotion classes from which we select the those considered in our study. Note that, after some pre-processing, each considered emotion category contains at least 2K samples.

<sup>13</sup> Annotators have been also allowed to assign a “no emotion” class in the case they believe the sentence is not providing any emotion connotation.

<sup>14</sup> More details about the *eTranslation* service are available at: [https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-ettranslation\\_en](https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-ettranslation_en).

<sup>15</sup> Amazon AWS SageMaker service, available at: <https://aws.amazon.com/sagemaker/>.

and  $f_{\text{ear}} = 9\%$ , overall resulting to be quite balanced between positive and negative classes.<sup>16</sup>

### 4.3 Forecasting method

Classic techniques in economy and finance do not scale well when the dimension of the data is big, noisy, and highly volatile [15]. In these cases, we need “good”, “acceptable” answers even if input data are extremely complex, working out of the box to recognize patterns among data and give improved quality predictions. Following this direction, Salinas et al. [39] proposed DeepAR, a forecasting method based on auto-regressive RNNs and leveraging on previous work on deep learning to time series data [31, 9, 28]. The approach is data-driven, that is, DeepAR learns a global forecasting model from historical data of all time series under consideration in the dataset. The model tailors an RNN architecture into a probabilistic forecasting setting, in which predictions are not restricted to point forecasts only, but density forecasts are also produced accordingly to a user-defined distribution (in our case a student  $t$ -distribution is experimentally selected). The outcome is more robust with respect to point forecasts alone, and uncertainty in the downstream decision making flow is reduced by minimizing expectations of the loss function (negative log-likelihood) under the forecasting distribution.

Similarly to classic RNNs, DeepAR is able to produce a mapping from input to output considering the time dimension. This mapping, however, is no longer fixed [27]. In addition to providing more accurate forecasts, DeepAR has also other advantages compared to classical approaches and other global methods [39]: (i) As the model learns seasonal behavior and dependencies on given covariates across time series, manual feature engineering is drastically minimized; (ii) DeepAR makes probabilistic forecasts in the form of Monte Carlo samples that can be used to compute consistent quantile estimates for all sub-ranges in the prediction horizon; (iii) By learning from similar items, DeepAR is able to provide forecasts for items with little history, a case where traditional single-point forecasting methods fail; (iv) DeepAR does not assume Gaussian noise, but can incorporate a wide range of likelihood functions, allowing the user to choose one that is appropriate for the statistical properties of the data.

DeepAR supports both LSTM and GRU cells. In our case we use an LSTM architecture to ensure overcoming vanishing gradients issues, typical of RNNs [31]. All the network weights and bias coefficients are estimated as usual during the training phase by back-propagating and minimizing the negative log-likelihood loss function. The DeepAR model that we use in our study has been deployed by using the implementation available in Gluon Time Series (GluonTS) [3], an open-source library for deep learning-based time series approaches<sup>17</sup> interfacing Apache MXNet.<sup>18</sup> We adopt a rolling window estimation technique for training and validation, with a window

<sup>16</sup> The remaining 24% of samples are classified as “no emotion” and removed from the dataset.

<sup>17</sup> <https://gluon-ts.mxnet.io/#gluonts-probabilistic-time-series-modeling>

<sup>18</sup> <https://mxnet.apache.org/>

length equal to half of the full sample, that is 2,975 data points. For each window, we calculate one step-ahead forecasts. We also set a re-training step for the model equal to 7 days, meaning that every 7 consecutive data points the DeepAR model is completely retrained. Hyperparameter tuning for the model has been performed through Bayesian hyperparameter optimization using the Ax Platform [32, 5] on the first estimation sample, providing the following best configuration: 1 RNN layer having 20 LSTM cells, 500 training epochs, and a learning rate equal to 0.001.

## 5 Results

In this section, we show our early empirical findings on the application of DeepAR to the prediction of the IBEX-35 daily changes, augmented with the emotions expressed within the Spanish economic news [14]. Note that forecasting the log-differences of the IBEX-35 index is an extremely challenging task, as the series behaves similarly to a random walk process. The goal is to assess whether news emotions contain some predictive power and might help in this difficult job.

We use standardization on training data only, a common requirement in the estimation of machine learning models. Typically, this is done by removing the mean and scaling to unit variance. However, outliers can often influence the sample mean/variance negatively, as it applies to our target variable. In such cases, the median and the inter-quartile range provide better results and, accordingly, we center and scale each feature independently with respect to these robust statistics. Median and inter-quartile range have been computed on training only, and then stored to be used to transform later validation data.

In the experiments, we consider an autoregressive-only DeepAR, that is without any covariates included in the model, referred to as *DeepAR-NoCov*. We then consider a DeepAR model with the news emotions used as covariates, referred to as *DeepAR-Emotions*. The experiments have been computed on an Intel(R) Xeon(R) E7 64-bit server having 40 cores at 2.10 GHz and overall 1TB of shared RAM. The DeepAR model training requires to run for few computational hours in parallel on the available CPU cores.

To benchmark the forecasting power of DeepAR we can compare its performance against those of other classic models, like for example a simple moving average approach (*Seasonal-MA*) and a naïve method (*Naïve*). With the moving average method, the forecasts of all future values are equal to the average (or “mean”) of the historical data. If we let the historical data be denoted by  $y_1, \dots, y_T$ , then we can write the forecasts as  $\hat{y}_{T+h|T} = \bar{y} = (y_1 + \dots + y_T)/T$ . The notation  $\hat{y}_{T+h|T}$  is a short-hand for the estimate of  $y_{T+h}$  based on the data  $y_1, \dots, y_T$ . In our case, we choose  $T = 7$ , that is we do a one-week moving average on the daily changes of the IBEX-35 index. For the naïve forecasts, instead, we simply set all forecasts to be the value of the last observation for our target (i.e. the IBEX-35 log-differences). That is,  $\hat{y}_{T+h|T} = \hat{y}_T$ . Since naïve forecasts are optimal when data observations follow a random walk, these are also called random walk forecasts. The naïve method works

Table 1: Out-of-sample forecasting metrics of the different methods, along with loss function values obtained at 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles.

<i>metrics</i>	RMSE	MAPE	mQL	QL <sub>0.1</sub>	QL <sub>0.3</sub>	QL <sub>0.5</sub>	QL <sub>0.7</sub>	QL <sub>0.9</sub>
<i>Seasonal-MA</i>	1.206	1.525	1.084	1.069	1.221	1.231	1.204	1.049
<i>Naïve</i>	1.287	1.471	1.179	1.046	1.313	1.424	1.209	0.949
<i>DeepAR-NoCov</i>	1.045	1.450	0.982	0.975	1.159	1.185	1.108	0.868
<i>DeepAR-Emotions</i>	<b>0.956</b>	<b>1.415</b>	<b>0.652</b>	<b>0.597</b>	<b>0.987</b>	<b>1.072</b>	<b>0.930</b>	<b>0.516</b>

well for many economic and financial time series, as is the case with our IBEX-35 fluctuations data.

For the evaluation, we use common time series prediction metrics, namely: root mean square error (RMSE), symmetric mean absolute percentage error (MAPE), and mean (weighted) quantile loss (mQL), that is the average quantile negative log-likelihood loss weighted with the density. Our results are shown in Table 1, along with the loss function values obtained by the methods at different quantiles. As the table shows, there is a clear superiority of the DeepAR algorithm with respect to the two other approaches. These improvements get consistently higher when the DeepAR model is combined with the news emotions. This suggests that the emotional content extracted from Spanish news contain a predictive power for our target forecasting. When these features are added to *DeepAR-NoCov*, producing the *DeepAR-Emotions* model, the results clearly improve in terms of all the metrics. These results look also consistent when we evaluate the models on the different quantiles. The best performance is obtained again by *DeepAR-Emotions*, followed by the *DeepAR-NoCov* model. We also note that the models show higher performance at high (0.9) and low (0.1) quantiles, with higher weighted quantile losses in general performing better than the lower quantiles.

Let’s now focus more in detail on the results obtained by *DeepAR-Emotions*, that from our computational experience appears to be the best performing model for the prediction task. Figure 2 shows the observations for the target time series (blue line) in the first 50 days of the testing period, together with the median forecast (dark green line) for the *DeepAR-Emotions* model and the confidence interval (lighter green). Forecasting an interval rather than a point is an important feature of the process since it provides an estimate of the uncertainty involved in the forecast, which allows downstream decisions accounting for such uncertainty. Figure 3 illustrates the median absolute forecast error for the *DeepAR-Emotions* model for the entire forecasting period. We can qualitatively see that the model does a reasonable job at capturing the variability and volatility of the time series for such a challenging scenario, overall achieving an acceptable deviation. The performance of the model drops for periods of particular crisis, which appears to be hardly predictable even if considering the emotional content of news. For example, this happens in

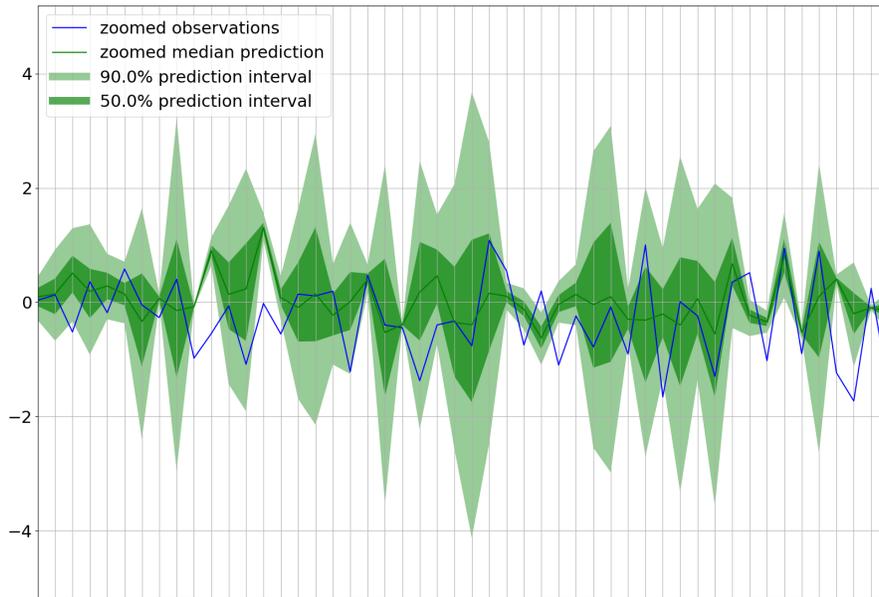


Fig. 2: Probabilistic forecasts (green) for the *DeepAR-Emotions* model and observations for the target series (blue) for the first 50 days in the forecasting period. The green continuous line shows the median of the probabilistic predictions, while the lighter green areas represents an higher confidence interval.

June 2016: the days after the Brexit vote. On that occasion, the IBEX-35 index drops by 12.4%, the biggest one-day drop for the benchmark index in its history, as the political and economic uncertainty unleashed by the UK referendum results hammering the shares of Spanish companies. Our model also fails at capturing this pronounced decline. The model also worsens in performance at the end of 2008 as

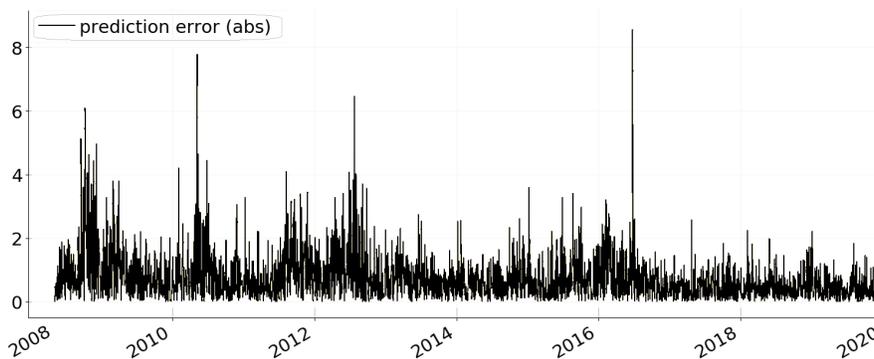


Fig. 3: Mean absolute forecast error for the *DeepAR-Emotions* model.

a result of the bankruptcy of Lehman Brothers that confirms the start of a global financial crisis. This actually corresponds to the other largest falls in the history of the IBEX-35, where the index falls by more than 9% in October 19, 2008, the day there is a crash in almost all world stock exchanges. After that black Friday, the rebound occurs but does not last: a few days later, the stock market crash is world-wide. As any prediction model would have failed on these unpredictable events, also our *DeepAR-Emotions* algorithm underperforms.

To improve interpretability and impact of the considered news emotions with respect to the DeepAR model, we also perform the computation of the Shapley values [35, 36] of the model using the SHAP library available for Python.<sup>19</sup> SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model [35]. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions [36]. In our case, in particular, we use the model agnostic KernelExplainer method of SHAP, which is used to explain any function by using a specially-weighted local linear regression to estimate SHAP values for the considered model.

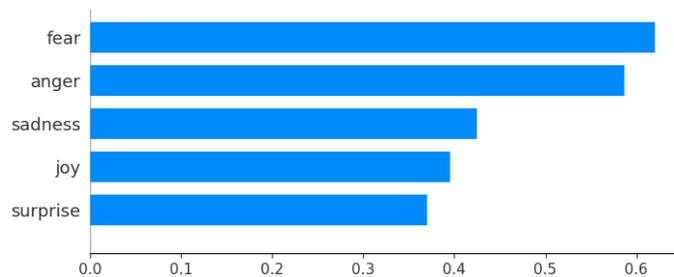


Fig. 4: SHAP bar plot: average impact of news emotions on model output magnitude.

To get an overview of which features result to be the most important for the DeepAR model, in Figure 4 we illustrate a standard bar plot with the mean absolute SHAP values of the top news emotions over the data samples, sorting by importance from the most impactful emotions with respect to the model output to the worst ones. As we can see, negative emotions, i.e., respectively, *fear*, *anger*, and *sadness*, result to be the features with highest impact, followed by the positive emotions of *joy* and *surprise*. This confirms the findings in previous literature that have shown media pessimism, in general, to be a relevant predictor for movements in spreads [4], bond markets interest rates [8, 33], and stock prices [45].

<sup>19</sup> <https://github.com/slundberg/shap>

## 6 Conclusions

In this paper we present an approach aimed at exploring the predictive power of news for economic and financial time series forecasting. In particular, we focus on the IBEX-35, a benchmark stock market index for Spain, and consider the log-differences of its daily close price. We then compute the emotional content of economic and financial news from Spanish outlets using neural machine translation, and use them as covariates of DeepAR, a neural forecasting method operating into a probabilistic setting, opportunely trained and validated with a rolling window approach. After providing an overview of our methodology, we report our results on this use-case application, showing satisfactory performance of the devised approach for such a challenging task. Emotions extracted from news look relevant for the forecasting exercise of the IBEX-35 fluctuations. Overall, DeepAR manages to achieve good trading results, producing better results when the news-based emotions variables are included into the model. The method is also benchmarked against two classic forecasting approaches, confirming its superiority. Results look encouraging, showing an overall validity of the employed methodology. News-based emotions appear to be good proxies for market investor’s expectations and behaviour. Their combination with state-of-the-art machine learning shows to have good potential for the forecasting of other economic and financial time series.

## References

- [1] L. Ackert, B. K. Church, and R. Deaves. Emotion and financial markets. *Economic Review*, 88:33–41, 2003.
- [2] S. Agrawal, P. Azar, A. W. Lo, and T. Singh. Momentum, mean-reversion and social media: Evidence from StockTwits and Twitter. *Journal of Portfolio Management*, 44:85–95, 2018.
- [3] A. Alexandrov, K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. S. Rangapuram, D. Salinas, J. Schulz, L. Stella, A. Caner Türkmen, and Y. Wang. GluonTS: Probabilistic Time Series Models in Python. *Journal of Machine Learning Research*, 21(1):1–6, 2020.
- [4] N. Apergis, M. C. K. Lau, and L. Yarovaya. Media sentiment and CDS spread spillovers: Evidence from the GIIPS countries. *International Review of Financial Analysis*, 47(C):50–59, 2016.
- [5] E. Bakshy, L. Dworkin, B. Karrer, K. Kashin, B. Letham, A. Murthy, and S. Singh. AE: A domain-agnostic platform for adaptive experimentation. In *Proc. Neural Information Processing Systems (NIPS)*, pages 1–8, 2018.
- [6] L. Barbaglia, S. Consoli, and S. Manzan. Monitoring the business cycle with fine-grained, aspect-based sentiment extraction from news. In *Lecture Notes in Computer Science*, volume 11985 LNAI, pages 101–106, 2020.
- [7] A. Beber, M. W. Brandt, and K. A. Kavaiecz. Flight-to-quality or flight-to-liquidity? Evidence from the Euro-area bond market. *Review of Financial Studies*, 22(3):925–957, 2009.
- [8] R. Beetsma, M. Giuliadori, F. de Jong, and D. Widiyanto. Spread the news: The impact of news on the European sovereign bond markets during the crisis. *Journal of International Money and Finance*, 34:83–101, 2013.
- [9] K. Benidis, S. S. Rangapuram, V. Flunkert, B. Wang, D. C. Maddix, A. Caner Türkmen, J. Gasthaus, M. Bohlke-Schneider, D. Salinas, L. Stella, L. Callot, and T. Januschowski. Neural forecasting: Introduction and literature overview. *CoRR*, abs/2004.10240, 2020. URL <https://arxiv.org/abs/2004.10240>.

- [10] O. Bernal, J.-Y. Gnabo, and G. Guilmin. Economic policy uncertainty and risk spillover in the Eurozone. *Journal of International Money and Finance*, 65(C):24–45, 2016.
- [11] L.-A.-M. Bostan and R. Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, 2018.
- [12] S. M. Carta, S. Consoli, L. Piras, A. S. Podda, and D. R. Recupero. Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting. *IEEE Access*, 9:30193–30205, 2021.
- [13] S. Clinchant, K. W. Jung, and V. Nikoulina. On the use of BERT for Neural Machine Translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117. Association for Computational Linguistics, 2019.
- [14] S. Consoli, M. Negri, A. Tebbifakhr, E. Tosetti, and M. Turchi. On neural forecasting and news emotions: the case of the Spanish stock market. In *MIDAS 2021: 6th Workshop on Mining Data for financial applications*, 2021. (Extended abstract).
- [15] S. Consoli, D. R. Recupero, and M. Saisana. *Data Science for Economics and Finance - Methodologies and Applications*. Springer Nature, 2021. doi: 10.1007/978-3-030-66891-4.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.
- [17] A. Dridi, M. Atzeni, and D. R. Recupero. FineNews: Fine-grained semantic sentiment analysis on financial microblogs and news. *International Journal of Machine Learning and Cybernetics*, pages 1–9, 2018.
- [18] C. Dunis, R. Rosillo, D. de la Fuente, and R. Pino. Forecasting IBEX-35 moves using support vector machines. *Neural Computing and Applications*, 23(1):229–236, 2013.
- [19] P. Ekman and D. Cordaro. What is meant by calling emotions basic. *Emotion Review*, 3(4): 364–370, 2011.
- [20] C. Fabbi, A. Righi, P. Testa, L. Valentino, and D. Zardetto. Social mood on economy index. In *XIII Conferenza Nazionale di Statistica*, 2018.
- [21] C. Faloutsos, T. Januschowski, J. Gasthaus, and Y. Wang. Classical and contemporary approaches to big time series forecasting. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 2042–2047, 2019.
- [22] C. Favero, M. Pagano, and E.-L. von Thadden. How does liquidity affect government bond yields? *Journal of Financial and Quantitative Analysis*, 45(1):107–134, 2010.
- [23] M. Fenton-O’Creevy, E. Soane, N. Nicholson, and P. Willman. Thinking, feeling and deciding: The influence of emotions on the decision making and performance of traders. *Journal of Organizational Behavior*, 32(8):1044–1061, 2011.
- [24] A. J. Garcia and R. Gimeno. Flight-to-liquidity flows in the Euro area sovereign debt crisis. Technical report, Banco de Espana Working Papers, 2014.
- [25] M. Gentzkow, B. Kelly, and M. Taddy. Text as data. *Journal of Economic Literature*, 57: 535–574, 2019.
- [26] S. Hansen and M. McMahon. Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99:S114–S133, 2016.
- [27] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–1780, 1997.
- [28] T. Januschowski, J. Gasthaus, Y. Wang, D. Salinas, V. Flunkert, M. Bohlke-Schneider, and L. Callot. Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36(1):167–177, 2020.
- [29] I. Kaastra and M. Boyd. Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, 10(3):215–236, 1996.
- [30] G. Lai, W.-C. Chang, Y. Yang, and H. Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018*, pages 95–104, 2018.
- [31] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

- [32] B. Letham and E. Bakshy. Bayesian optimization for policy search via online-offline experimentation. *Journal of Machine Learning Research*, 20, 2019.
- [33] S. Liu. The impact of textual sentiment on sovereign bond yield spreads: Evidence from the Eurozone crisis. *Multinational Finance Journal*, 18(3/4):215–248, 2014.
- [34] T. Loughran and B. McDonald. When is a liability not a liability? Textual analysis, dictionaries and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.
- [35] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017.
- [36] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.
- [37] S. Manganelli and G. Wolswijk. What drives spreads in the Euro area government bond markets? *Economic Policy*, 24(58):191–240, 2009.
- [38] Y. Qin, D. Song, H. Cheng, W. Cheng, G. Jiang, and G. Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 2627–2633, 2017.
- [39] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.
- [40] D. Senol and M. Ozturan. Stock price direction prediction using artificial neural network approach: The case of turkey. *Journal of Artificial Intelligence Research*, 3:261–268, 2010.
- [41] A. H. Shapiro, M. Sudhof, and D. Wilson. Measuring news sentiment. *Federal Reserve Bank of San Francisco Working Paper*, 2018.
- [42] R. Taffler. Emotional finance: investment and the unconscious. *The European Journal of Finance*, 24(7-8):630–653, 2018.
- [43] G. Tang, M. Müller, A. Rios, and R. Sennrich. Why self-attention? A targeted evaluation of neural machine translation architectures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 4263–4272, 2020.
- [44] A. Tebbifakhr, L. Bentivogli, M. Negri, and M. Turchi. Machine translation for machines: the sentiment classification use case. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1368–1374, Hong Kong, China, 2019. Association for Computational Linguistics.
- [45] P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007.
- [46] L. A. Thorsrud. Nowcasting using news topics. Big Data versus Big Bank. *Norges Bank Working Paper*, 2016.
- [47] L. A. Thorsrud. Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, pages 1–17, 2018.
- [48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-December, pages 5999–6009, 2017.
- [49] H. Yuan, R. Y. K. Lau, M. C. S. Wong, and C. Li. Mining emotions of the public from social media for enhancing corporate credit rating. In *Proceedings - 2018 IEEE 15th International Conference on e-Business Engineering, ICEBE 2018*, pages 25–30, 2018.
- [50] D. Zhang, M. Hu, and Q. Ji. Financial markets under the global pandemic of COVID-19. *Finance Research Letters*, 36:101528, 2020.
- [51] F. Zhou, Q. Zhang, D. Sornette, and L. Jiang. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. *Applied Soft Computing Journal*, 84, 2019.
- [52] J. Zhu, Y. Xia, L. Wu, D. He, T. Qin, W. Zhou, H. Li, and T. Liu. Incorporating BERT into Neural Machine Translation. In *International Conference on Learning Representations (ICLR)*, 2020. URL [https://iclr.cc/virtual\\_2020/poster\\_Hyl7ygStwB.html](https://iclr.cc/virtual_2020/poster_Hyl7ygStwB.html).