A Quantitative Exploration of Natural Language Processing Applications for Electricity Demand Analysis

Yun Bai, Simon Camal, Andrea Michiorri

A Quantitative Exploration of Natural Language Processing Applications for Electricity Demand Analysis

Yun Bai, Simon Camal, Andrea Michiorri

***Abstract*—The relationship between electricity demand and weather has been established for a long time and is one of the cornerstones in load prediction for operation and planning, along with behavioral and social aspects such as calendars or significant events. This paper explores how and why the social information contained in the news can be used better to understand aggregate population behaviour in terms of energy demand. The work is done through experiments analysing the impact of predicting features extracted from national news on day-ahead electric demand prediction. The results are compared to a benchmark model trained exclusively on the calendar and meteorological information. Experimental results showed that the best-performing model reduced the official standard errors around 4%, 11%, and 10% in terms of RMSE, MAE, and SMAPE. The best-performing methods are: word frequency identified COVID-19-related keywords; topic distribution that identified news on the pandemic and internal politics; global word embeddings that identified news about international conflicts. This study brings a new perspective to traditional electricity demand analysis and confirms the feasibility of improving its predictions with unstructured information contained in texts, with potential consequences in sociology and economics.**

***Keywords*—Electricity demand forecasting; Natural language processing; Population behaviour; Social events**

I. Introduction

A. Context

Known dependencies characterise electricity demand to economic activity (e.g., working or non-working days) and weather (e.g., low or high temperatures). The impact of significant social events, such as major sports competitions, has also been identified along with the recent COVID-19 pandemic. This work aims to analyze unstructured information present in news' textual content to find relationships between social events and electricity demand using the techniques of Natural Language Processing (NLP) and numerical predictions.

B. State of the art

The field of text-based forecasting is relatively new and is being explored deeply by researchers, with a visible acceleration after 2010 and a general interest in price predictions. It is possible to identify three influential milestones relevant to this research topic.

Firstly, an earlier expression of the idea could be traced to [1] where the fundamental concepts of text-based forecasting were suggested, and tests were carried out for movie revenues forecasting with the n-grams, part-of-speech n-grams, and the dependency relations from online movie reviews. The study showed an improvement in prediction performance and considered the mechanisms behind the improvement among future developments, aiming to understand why specific words are linked with the predicted variable.

Secondly, it is worth mentioning that the authors of [2] pointed out a series of important conclusions: i) as in forecasting more generally, complex models or methods were not necessarily more successful than simpler ones; ii) fully automatic methods benefitted from replicability, speed, and ease of updating but with the downside of missing more subtle shades of meaning.

Finally, in [3], textual data were used for improving electricity demand prediction using weather reports and tweets, specifically in the context of the sudden demand changes caused by COVID-19 lockdowns in France and Italy. This study revealed that calendar and meteorological information extracted from the text was a beneficial supplement in the absence of these data sources. The paper also showed that the sudden relevance of words related to 'remote working' in the public discourse was strongly correlated to changes in electricity demand.

Independently from these three milestones, the text-based forecasting approach has been tested in several fields such as bankruptcy and fraud [4], stock prices [5], psychological disturbances in college students [6], demand for taxi rides [7], short term apartment rentals [8], street crime [9], COVID-19 evolution [10], tourism demand from online searches [11], and crude oil prices [12].

The works above show that in terms of NLP methods, sentiment analysis and topic modeling are well-established for extracting textual information, and word embedding, also word vectorization, acts as an upstream task for NLP applications such as text classification. The main techniques applied are summarised following.

Sentiment analysis covers different aspects as introduced in [13]. ***Polarity analysis*** portrays the sentiment tendency within a sentence. For supervised methods, sentences are labeled as either negative-positive binary or negative-neutral-positive ternary categories. ***Subjectivity analysis*** quantifies the amount of personal opinion carried in a sentence. The higher proportion of personal opinion in a sentence, the more subjective it is. ***Emotion analysis*** is commonly used in social media analysis, such as Twitter and Weibo, to label sentences with multiple human emotions such as happy, angry, sad, disgusted, and scared [14], [15]. Besides that, researchers also apply advanced neural network models and word embeddings to complete the model classification task [16], [17], [18].

Topic analysis. Topics are abstract expressions of text contents and topic analysis is a technique to discover the hidden semantic structures from the text collections. [19]. Essentially, topic analysis is a dimensionality reduction method. In the

The authors are with the Centre for Processes, Renewable Energies and Energy Systems (PERSEE), MINES Paris - PSL University, Sophia Antipolis, France, e-mail: (yun.bai@minesparis.psl.eu).

NLP field, bag-of-words provides vector representations of text statistically, but it suffers from sparsity. Combined with Principal Component Analysis (PCA), [20] developed Latent Semantic Analysis (LSA), which represents documents as denser vectors and yields principal components, or topics, that express deeper semantic content. Subsequently, [21] built on this foundation and studied the Latent Dirichlet Allocation (LDA), which is currently a popular topic model, and we will describe it in Section II-A3.

Word vectorization is a technique aiming at mapping words to vectors in the same space. Word vectors are closer in distance when they are similar in lexical meaning. Earlier word vectorization methods were frequency-based, such as one-hot encoding, index encoding, and Term Frequency-Inverse Document Frequency (TF-IDF) [22]. Among the recent developments, Global Vectors (GloVe) and Word2Vec models are based on the local terms co-occurrence [23], [24]. Also, other advanced models would pay more attention to the words for prediction, thus resolving sentence ambiguities and inferring word meaning. The attention mechanism spawned the Transformer method, on which pre-trained language models, such as BERT, are based [25]. These word embeddings are prediction-based, obtain a higher-level representation of the text probability distribution, and reduce the dimensionality of word vectors, yet frequency-based methods struggle to do so.

C. Hypotheses and objectives

The field of NLP application for forecasting, particularly electricity demand forecasting, is at its beginning, and this work aims to explore the possibilities and limitations of the approach. In particular, we consider two following hypotheses:

1) Except for weather and economic activity, electricity demand is also influenced by social factors visible in the news.
2) It is possible to quantify textual information and use it in practice via NLP.

The objectives of this paper are summarised as follows:

1) To verify *IF* it is possible to extract valuable information from news to improve electricity demand forecasting.
2) To explore *HOW* to best treat textual information and develop a complete forecasting chain that integrates text and other structured data.
3) To understand *WHY* we have improved performance and uncover the mechanisms of this approach.

D. Structure of the paper

In this paper, after Section I where the problem is introduced with the context, state of the art, and a clarification of the contributions to knowledge, the methodology is presented in Section II with an overview of the NLP and forecasting techniques. Evaluation metrics and model explanations techniques are employed along with the Case study Section II-E. Results are shown in Section III followed by a discussion in Section IV, and conclusions are drawn in Section V.

II. Methodology

An overview of the workflow for this study is described and visually shown in Figure 1. Firstly, electricity load, meteorology and economic activity, and news data are acquired in modules A, B, and C. Then numerical and textual data are pre-processed in modules D and E. In the case of numeric data, time series are cleaned and synchronised for i) the target variable (aggregated electricity demand), ii) ambient temperature, and iii) calendar features (holidays and weekends). To build the benchmark model in module F, we used the features from D, including lags, calendars, and temperatures. The first group of calendar features is represented by the day of the week and day of the year, embedded through their sine and cosine, reflecting the multiple seasonality in electric demand. Additional variables account for weekends and holidays.

Then, textual inputs are pre-processed to create different numerical input features. Subsection II-A shows the detailed textual features. It should be noted that some textual features are redundant for forecasting, and they are filtered out by the Granger test as suggested in [26], which happens in module E. The text-based forecasting model is built in module G with inputs from E and D. Finally, the results of the two models are evaluated in H with an analysis of errors and the explanation from the global, local, and causality aspects.

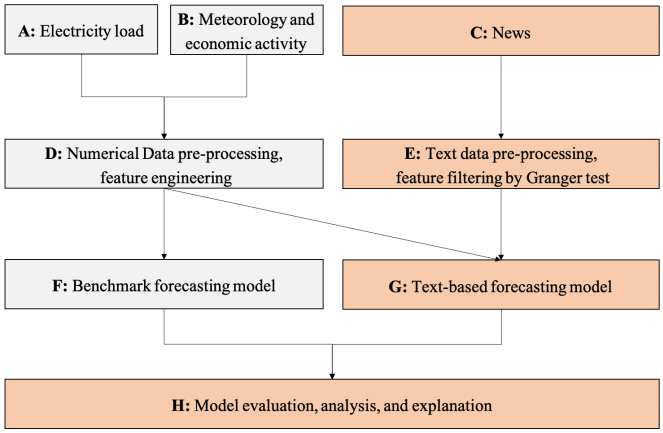

Fig. 1. The forecasting framework of this research.

A. Preprocessing for the text-based model

This section describes the methods used in Module E in 1 to extract numerical features to be fed to the following machine learning prediction model from the raw textual data. For textual data, the pre-processing includes splitting a sentence into words; converting all letters to lowercase; removing stopwords, words with less than three letters, and all numbers. All news is finally transformed into word lists.

This study borrowed three NLP techniques: sentiment and subjectivity analysis, topic modeling, and text vectorisation, to capture sentiment scores, topic distributions, and word embeddings in the news.

1) Simple statistics: **Count features** include 27 features counted daily. For each text, the number of words, sentences,

unique words, non-stopping words, the average number of sentences, and the average number of words for all the sentences each day are calculated. Finally, we made two categorical features for the proportion of news in the 18 sections on the BBC website (e.g., Asia, Business, UK Politics, ...).

Word frequencies analysis consists of the words in text after stop-words and non-words have been removed. To reduce the number of words, only the most relevant has been selected. Considering the different volumes of titles (***T***), descriptions (***D***), and text bodies (***B***), we set separate thresholds $\sigma_T = 200$, $\sigma_D = 400$, $\sigma_B = 5000$, and only words that exceeded this threshold were included. This resulted in the selection of respectively 456, 329, and 550 words for ***T***, ***D***, and ***B***.

2) Sentiment and subjectivity analysis: This analysis is performed with the library TextBlob from Python's Natural Language ToolKit (NLTK) [27], widely used in sentiment analysis and is particularly suitable for corpora without manual labeling [28], [29], [30]. This algorithm calculates a score between [-1, 1] for each word according to its negative or positive meaning. In addition, it calculates a score between [0, 1] according to its subjectivity, considering the modifying effect of adjectives and adverbs. In this study, a distribution of sentiment and subjectivity is calculated for each piece of news. This distribution is then discretised in five quintiles $[0, 0.2)$, $[0.2, 0.4)$, $[0.4, 0.6)$, $[0.6, 0.8)$, $[0.8, 1]$. They also computed the maximum, minimum, average, and standard deviation for 18 sentiment features.

3) Topic distribution: Topic distribution is analysed through LDA model, with the objective to obtain a probabilistic estimation of the belonging of each news article to a specific topic. LDA is a classical unsupervised topic model that mimics human writing by assuming a text-generation process. The LDA model assists in extracting several topics from mass texts and gives the probability distribution of each text under these topics. Based on this, the average topic probability distribution for all daily news is easily obtained. The probability value under each topic reflects how widespread the daily news is on that topic and is a noteworthy feature. According to the topic number selection method in [31], we set the ideal number of topics as κ and obtained $\kappa_T = 87$, $\kappa_D = 100$, $\kappa_B = 69$ for each text type.

4) Text vectorisation: This step is to obtain a vectorial representation of the text according to its content for the other texts in the corpus through the library GloVe described in [23]. Thanks to the pre-trained word embeddings based on a large corpus, the current GloVe not only contains rich global and local semantic information but also facilitates our application on a new corpus without repeating the time-consuming training work.

High-dimensional textual features are extracted and expressed in this work from the textual dataset by transforming words into 100-dimensional vectors with GloVe. A text vector is obtained by averaging all the word vectors in this text. The position of each element in the text vector is an axis in the high-dimensional space. We averaged all the text vectors within a day to get the features.

5) Granger test: The method mentioned above can produce numerous features, totaling 2026 in the configuration. However, it is necessary to prevent some of the features from positively affecting the quality of the forecast. Therefore the Granger test is done preventively before the training of the prediction model in module E.

The Granger test is a measure to test whether a stationary time series *X* contributes to the forecasting of parameter *Y* [32]. It is based on the following autoregressive model:

$$y_t = \theta_0 + \sum_{i=1}^{T} \theta_i y_{t-i} + \sum_{i=1}^{T} \phi_i x_{t-i} + \epsilon_t \qquad E(\epsilon) = 0, \tag{1}$$

where θ_i and ϕ_i are the lag coefficients of *X* and *Y*, and *T* is a chosen lag order, which in this paper is 30. The null hypothesis is

$$H_0 : \phi_1 = \phi_2 = ... = \phi_T = 0, \tag{2}$$

i.e. that the lagged terms of *X* are independent of *Y*. This can be rejected when $p < 0.05$.

As shown in [26], [33], the Granger test assists in the selection of text features that are relevant to the forecasting target. It acts as a dimensionality reduction for the text features. In these cases, the Granger test is usually performed unilaterally, but we do not want *Y* Granger cause *X*. Therefore, the experiments in this paper include a bilateral Granger test, i.e., requiring $p_2 < 0.05$ for the first test and $p_1 \geq 0.05$ for the second.

Note that the Granger test used for initial feature filtering is not an actual causality test, as it involves correlations between lagged and predicted values, and correlations do not necessarily lead to causality. We will discuss the causality of text features and electricity load in more detail in Section IV-C.

B. Forecasting with textual features

Due to the nature of the dataset, the news was available per day, and it was not possible to distinguish an intra-day order. Furthermore, the news may refer to events of the previous days. It was then decided to frame the problem as a forecasting problem for the day ahead. This section describes the forecasting algorithm used in Block F and G. The idea is to verify if textual-based features measured in day *d-1* can provide additional explanation to the behaviour of the demand in day *d+1*, considering a prediction calculated in day *d+1* in a valuable time to take decisions (for example for trading or scheduling).

After an initial comparison with different models, such as Support Vector Regression and Multilayer Perceptron, the ExtraTrees algorithm has been selected because of its higher performance (known in the case of relatively small tabular datasets) and flexibility. ExtraTrees is an ensemble learning method within the decision tree paradigm. Like Random Forest, ExtraTrees creates many decision trees during the training but randomly samples each tree. The features in the trees are also randomly selected by splitting values without using the criterion of optimizing localization, which enables ExtraTrees to achieve faster computational speed and membership diversity [34]. Current applications of ExtraTrees have emerged in

the field of electricity load forecasting, for example, the peak load forecasting [35], day-ahead load demand forecasting case from Spain [36], and medium- and long-term load forecasting [37].

C. Evaluation

The two regression models in blocks F and G are evaluated according to the following criteria. Firstly, deterministic evaluation metrics are calculated to compare the performance of the algorithm and the baseline quantitatively. Then an analysis is carried out to explain the relationship between the features and demand.

Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Symmetric Mean Absolute Percentage Error (SMAPE) are used in this study, and they are calculated as follows:

$$RMSE = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(y_i - \hat{y_i})^2}, \tag{3}$$

$$MAE = \frac{1}{H}\sum_{i=1}^{H}\|y_i - \hat{y_i}\|, \tag{4}$$

$$SMAPE = \frac{100\%}{H}\sum_{i=1}^{H}\frac{\|y_i - \hat{y_i}\|}{\|y_i + \hat{y_i}\|/2}, \tag{5}$$

where H is the forecasting horizons and $H = 48$ for the half-hourly data in our case. y_i and $\hat{y_i}$ are truth and forecased loads at time i.

These metrics are calculated for each time step of the test dataset, but they are then averaged over a whole yearly period and noted as $\overline{rmse}$,$\overline{mae}$, and $\overline{smape}$.

D. Models explaination

Previous research has confirmed that well-selected text features enhance forecasting, yet it is often difficult to explain this improvement deeply. Explainability is necessary to shed light on the behaviour of the trained machine learning models, which otherwise would be completely black boxes. This paper attempts to explore the mechanisms by which text features enhance forecasting in terms of *i)* global, *ii)* local, and *iii)* causality.

Global explainability is analysed through Pearson correlation coefficients. It will not be discussed here due to its popularity. Local explainability is analised through the Local Interpretable Model-agnostic Explanation (LIME), whilst causality is analysed through Double Machine Learning (Double ML). Both methods are described in the following lines.

1) Local explainability: LIME targets a sample of the original data and generates a new, normally distributed local dataset using the current sample [38], [39]. After that, a simple surrogate model, such as linear regression, is used to fit the new dataset, yielding a locally interpretable perspective on the data under perturbation. We can interpret how features in the current local affect the forecasting by viewing the coefficients of the features in the linear model.

2) Causation: Previous analysis, such as the Pearson correlation or the LIME model, verified the correlation between the candidate features and electric demand. Nevertheless, positive or negative correlations between variables may be the result of coincidence, and it is known that correlation does not imply causation [40]. For this reason, a test is proposed to verify the causality between the features and the target. The so-called causality is the effect of a particular feature of interest (Treatment, *T*) on the predictions (Outcomes, *Y*), provided that the rest of the features remain constant (Confounders, *X*). The Double ML method helps find the causality between variables and predictions based on the following partially linear model [41]:

$$Y = f(X) + \rho(X)T + \epsilon \quad E(\epsilon|X,T) = 0, \tag{6}$$

where $f(X)$ a train model and $\rho(X)$ the treatment effect. Although $\rho(X)$ is regarded as a function by some studies [42], we treat it as a constant $\rho(X) = \tau$ for simplicity. The next step is to explore the effect of the text feature *T* on the output *Y* while maintaining the rest of the features *X* constant. In this way, it is possible to observe whether, or to what extent, *T* causes a change in *Y*. To estimate $\rho(X)$, the formula 6 is rewritten in residualised form:

$$Y - \hat{Y} = \tau(T - \hat{T}) + \delta \quad E(\delta|X,T) = 0, \tag{7}$$

in which $\hat{Y} = g(X)$ and $\hat{T} = h(X)$ are forecasts of *Y* and *T*. $g(X)$ and $h(X)$ are nuisance functions that can be replaced by many machine learning methods. In this case, we set $f(X)$, $g(X)$, and $h(X)$ for all ExtraTrees regressors. The treatment effect τ can then be obtained with Ordinary Least Squares (OLS).

E. Datasets

This work uses three datasets covering five years between June 2016 and May 2021. The first four years are used as a training set and the last year as a test set. Aggregated electric demand for the UK is obtained from the ENTSOe transparency platform [43] along with the official day-ahead forecasts. Historical Bank holidays and daily temperatures for the city of London have been taken from commercial websites. According to [44], we used the observed rather than predicted temperatures for the convenience of reproducibility.

Previous studies have mainly used keywords for external texts to filter news related to the forecasting domain [33], [45], [12]. However, this paper proposes to use the entire volume of news from British Broadcasting Corporation (BBC) to explore the impact of broader social events on electricity load forecasting. Over 80,000 news items were collected, thanks to the repository [46], which archives [47].

III. Results

A. Benchmark model

A prediction model based on ExtraTrees Regression (ETR) has been trained on the basic datasets. Grid search and five-fold cross-validation were used to find the optimal parameters and avoid overfitting. The performance of the benchmark

model is summarised in Table I, in which different combinations of features are tested: demand $\mathcal{D}$, calendar features $\mathcal{C}$ and temperature $\mathcal{T}$. These are compared with the official forecasts obtained from the ENTSOe Transparency Platform $\mathcal{ENTSO}$].

TABLE I
BENCHMARK MODEL PERFORMANCE

Features	$\overline{rmse}$(MW)	$\overline{mae}$(MW)	$\overline{smape}$(%)
$\mathcal{ENTSO}$]	2800.50	2544.86	7.65
$\mathcal{D}$	2983.05±5.35	2539.25±5.14	7.75±0.01
$\mathcal{D}+\mathcal{C}$	2896.49±4.81	2468.48±4.22	7.56±0.01
$\mathcal{D}+\mathcal{T}$	2938.40±2.81	2488.93±2.61	7.59±0.01
$\mathcal{D}+\mathcal{C}+\mathcal{T}$	**2800.77±4.84**	**2374.07±4.39**	**7.29±0.01**

As expected, combining these inputs produces the best performance with moderate gains concerning the official forecasts over the three deterministic metrics. However, holidays do not provide improvements. The best improvements range in the region of 200MW in $\overline{mae}$ and 4% in $\overline{smape}$.

B. Textual features enhanced model

1) Impact of textual features: In this subsection, a new model is trained with all the features used in the benchmark model ($\mathcal{D}$, $\mathcal{C}$ and $\mathcal{T}$), and the features extracted from the textual dataset, here divided in terms of Title (**T**), Description (**D**) and Body (**B**). The textual features $\mathcal{F}_t$ explored in the step are: Count Features ($\mathcal{CF}$), Words Frequencies ($\mathcal{WF}$), Sentiment ($\mathcal{SE}$), Topic Distributions ($\mathcal{TD}$), and GloVe Word Embeddings ($\mathcal{GWE}$) from ***T***, ***D***, and ***B***. The experimental results are presented in Table II, along with the original and selected feature numbers after the Granger test in each group.

The results in Table II show that 6 of the 15 sets of experiments outperform $\mathcal{D}+\mathcal{C}+\mathcal{T}$. This reflects that some text features are beneficial for load forecasting. Among these text features, $\mathcal{WF}$ in all three text types reduces forecasting errors, especially in ***T***. In addition, $\mathcal{SE}$, $\mathcal{TD}$ and $\mathcal{GWE}$ from ***B*** also improved $\mathcal{D}+\mathcal{C}+\mathcal{T}$ and are better than those in ***T*** and ***D***. From now on these best-performing features are renamed as follows: $\mathcal{WF}$ from ***T***, ***D***, and ***B*** are renamed as $\mathcal{WF}_T$, $\mathcal{WF}_D$, and $\mathcal{WF}_B$, whilst the other beneficial $\mathcal{F}_t$s are $\mathcal{SE}_B$, $\mathcal{TD}_B$, and $\mathcal{GWE}_B$.

2) Features combination: The performance improvement with different textual features combination is tested. This is done because it is expected that the different textual features have considerable overlap, coming from similar texts, especially if considering the information contained in ***T***, ***D***, and ***B***. The combinations considered are $\mathcal{M}_0 := \mathcal{WF}_T$, $\mathcal{M}_1 := \mathcal{WF}_T + \mathcal{WF}_D + \mathcal{WF}_B$, $\mathcal{M}_2 := \mathcal{WF}_T + \mathcal{SE}_B$, $\mathcal{M}_3 := \mathcal{WF}_T + \mathcal{TD}_B$, $\mathcal{M}_4 := \mathcal{WF}_T + \mathcal{GWE}_B$, $\mathcal{M}_5 := \mathcal{WF}_T + \mathcal{SE}_B + \mathcal{TD}_B$, $\mathcal{M}_6 := \mathcal{WF}_T + \mathcal{SE}_B + \mathcal{GWE}_B$, $\mathcal{M}_7 := \mathcal{WF}_T + \mathcal{TD}_B + \mathcal{GWE}_B$, $\mathcal{M}_8 := \mathcal{WF}_T + \mathcal{SE}_B + \mathcal{TD}_B + \mathcal{GWE}_B$.

For these combinations, results are plotted in Figure 2. Here it is possible to see how by combining $\mathcal{WF}_D$ and $\mathcal{WF}_B$ with $\mathcal{WF}_T$, the forecasting errors increase. This shows that $\mathcal{WF}_T$ is sufficient for forecasting in terms of word frequency.

$\mathcal{M}_2$ shows the addition of $\mathcal{SE}_B$ to reduce the forecasting errors. $\mathcal{GWE}_B$ in $\mathcal{M}_4$ reduce the error spread in the box plot. The combination $\mathcal{M}_6$, obtained from $\mathcal{SE}_B$ and $\mathcal{GWE}_B$, brings together the advantages of both and has the best performance, which is then used for further analysis.

3) Errors Analysis: Errors are analysed according to different hours and day types. This analysis is carried out on the model trained with the $\mathcal{M}_6$ input combination. In Figure 3, it is possible to see the difference in terms of $\overline{rmse}$, $\overline{mae}$ and $\overline{smape}$ between the benchmark model and the advanced model with textual features for different hours of the day. The performance improvement is generally more remarkable in the first and last hours of the day, usually characterised by more giant ramps in demand. In Table III, the same analysis is presented comparing the performance on weekdays and weekends. The error is more significant on weekends, probably because of the lower absolute value of the demand, and the advanced model increases its advantage on weekends.

4) DM-test: The Diebold-Mariano (DM) test was applied to the forecasts across models to evaluate the differences in forecast accuracy statistically. The null hypothesis $\mathcal{H}_0$ is that there is no significant difference between the two models. The alternative hypothesis is that one model is better than another, given the one-sided situation. With a p-value less than 0.05, we can infer a better model. We used four models for the DM-test: the ETR $\mathcal{D}$, benchmark $\mathcal{D}+\mathcal{C}+\mathcal{T}$ in Table I, $\mathcal{M}_0$ and $\mathcal{M}_6$ in Figure 2.

Table IV shows the p-values of DM-test for the four models. The bolded p-values are less than 0.05 where we reject the null hypothesis and take the model in the column better than the one in the row. For example, when comparing the model $\mathcal{D}+\mathcal{C}+\mathcal{T}$ and $\mathcal{M}_0$, the p-value is 0.0404 and less than 0.05. So we found that there is a statistically significant difference between the forecasting accuracy of model $\mathcal{M}_0$ and $\mathcal{D}+\mathcal{C}+\mathcal{T}$, and $\mathcal{M}_0$ is superior to $\mathcal{D}+\mathcal{C}+\mathcal{T}$.

IV. DISCUSSION

In the following subsections IV-A, IV-B, and IV-C, we would explain the relationships between textual features and electricity load, from the global, local, and causality views.

A. Global correlations

Although model $\mathcal{M}_6$ suggested that the combination of $\mathcal{WF}_T$, $\mathcal{SE}_B$, and $\mathcal{GWE}_B$ is the best-performed, we still list all the beneficial textual feature groups: $\mathcal{WF}_T$, $\mathcal{WF}_D$, $\mathcal{WF}_B$, $\mathcal{SE}_B$, $\mathcal{TD}_B$, $\mathcal{GWE}_B$. The detailed descriptions are in Table V.

We then measured the correlation between textual features and loads. Due to a large number of words in $\mathcal{WF}$, we only mention six of the top 3 with the strongest positive and negative correlations overall. Besides that, $\mathcal{SE}_B$, $\mathcal{TD}_B$, and $\mathcal{GWE}_B$ were included. Figure 4 illustrates the correlations scenarios over hours on different seasons, weekdays, and weekends.

Generally, Figure 4 presents more obvious correlations in spring, summer, and weekends. Except for seasons correlations, driver-T shows a positive correlation with load in the dawn and early morning in autumn and winter. The

TABLE II
FORECASTING ERRORS WHEN ADDING $\mathcal{F}_t$ INTO MODEL $\mathcal{D}+\mathcal{C}+\mathcal{T}$

Text Type	$\mathcal{F}_t$	#Original $\mathcal{F}_t$	#Selected $\mathcal{F}_t$	$\overline{rmse}$(MW)	$\overline{mae}$(MW)	$\overline{smape}$(%)
$\mathcal{D}+\mathcal{C}+\mathcal{T}$	—-	0	0	2800.77±4.84	2374.07±4.39	7.29±0.01
T	$\mathcal{CF}$	27	1	2799.65±3.73	2372.63±3.38	7.29±0.01
	$\mathcal{WF}$	456	32	2702.23±4.95**	2283.52±4.24**	6.98±0.01**
	$\mathcal{SE}$	18	2	2803.73±6.15	2376.81±5.82	7.30±0.02
	$\mathcal{TD}$	87	6	2806.27±5.35	2377.14±5.08	7.31±0.01
	$\mathcal{GWE}$	100	5	2795.80±4.34	2367.49±2.88	7.27±0.01
D	$\mathcal{CF}$	27	0	2798.81±4.85	2371.95±4.51	7.29±0.01
	$\mathcal{WF}$	329	25	2760.65±6.85*	2342.98±6.16*	7.14±0.02*
	$\mathcal{SE}$	18	0	2798.81±4.85	2371.95±4.51	7.29±0.01
	$\mathcal{TD}$	100	5	2805.32±5.46	2379.75±4.96	7.31±0.01
	$\mathcal{GWE}$	100	2	2803.58±5.26	2375.26±4.84	7.30±0.01
B	$\mathcal{CF}$	27	0	2798.81±4.85	2371.95±4.51	7.29±0.01
	$\mathcal{WF}$	550	10	2751.40±4.38*	2330.54±3.95*	7.16±0.01*
	$\mathcal{SE}$	18	1	2788.45±4.61**	2360.75±4.33**	7.26±0.01**
	$\mathcal{TD}$	69	2	2747.51±5.60**	2323.50±5.36**	7.12±0.02**
	$\mathcal{GWE}$	100	3	2749.97±2.31**	2327.66±2.43**	7.16±0.01**

the counting numbers, * better than $\mathcal{D}+\mathcal{C}+\mathcal{T}$, ** better than $\mathcal{D}+\mathcal{C}+\mathcal{T}$ and best in ***T***, ***D***, and ***B***.

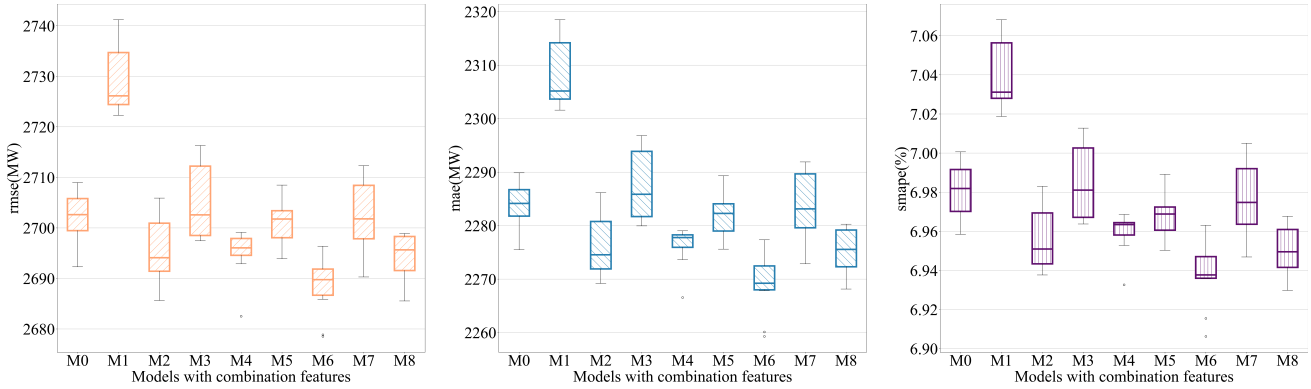

Fig. 2. Errors box plot for the models with textual features combination. The left subplot is for the $\overline{rmse}$(MW), the middle one for the $\overline{mae}$(MW), and the right one for the $\overline{smape}$(%).

TABLE III
$\overline{rmse}$, $\overline{mae}$, AND $\overline{smape}$ ON WEEKDAYS AND WEEKENDS.

Day type	$\overline{rmse}$(MW)	$\overline{rmse}$(MW)-T	$\overline{mae}(MW)$	$\overline{mae}$(MW)-T	$\overline{smape}$(%)	$\overline{smape}$(%)-T
Weekdays	2846.73	2762.44	2392.79	2313.22	7.23	6.95
Weekends	2623.65	2431.61	2268.61	2089.53	7.29	6.70

TABLE IV
P-VALUES OF DM-TEST RESULTS

	$\mathcal{D}$	$\mathcal{D}+\mathcal{C}+\mathcal{T}$	$\mathcal{M}_0$	$\mathcal{M}_6$
$\mathcal{D}$	1.0000	**0.0000****	**0.0000****	**0.0000****
$\mathcal{D}+\mathcal{C}+\mathcal{T}$	0.9999	1.0000	**0.0404***	**0.0181***
$\mathcal{M}_0$	0.9999	0.9596	1.0000	**0.0322***
$\mathcal{M}_6$	0.9999	0.9818	0.9678	1.0000

* for $0.01 < p < 0.05$, and ** for $0 < p \leq 0.01$

correlation is more pronounced for mps-D in winter. The three words with negative correlations in Figure 4 are all related to coronavirus. They show strong negative correlations with load in the spring and summer, with decreasing correlations in the subsequent seasons. The coronavirus-related Topic-18 shows similar regularity. We also noticed that the peaks differ in spring (daytime) and summer (evening). In addition, social sentiment correlated higher in winter. Among the dimensional features of GloVe, only Dim-51, train transportation in the UK, presents a positive correlation with the load in spring and summer and is more noticeable on weekends. This correlation is relatively strong in the daytime in spring, and there is a peak from hour 19h to 22h in summer.

TABLE V
TEXTUAL FEATURES DESCRIPTION

Group	Features
$\mathcal{WF}_T$	Andy, Murray, newspaper, headlines, rules, Ireland, year, NHS, staff, Glasgow, family, home, European, mark, updates, Paris, say, elections, premier, hit, bomb, second, funeral, talks, Spain, budget, driver, care, sorry, Scotland, job, coronavirus
$\mathcal{WF}_D$	following, around, Ireland, MPS, least, away, reach, schools, wife, shows, weeks, help, figures, days, lead, Wales, security, hit, outside, Scotland, Monday, leaders, restrictions, pandemic, coronavirus
$\mathcal{WF}_B$	coming, power, city, inside, job, ahead, social, strong, return, war
$\mathcal{SE}_B$	the minimum subjectivity value
$\mathcal{TD}_B$	**Topic-5 (Politics related to Ireland)**: Ireland, Northern, Irish, Belfast, DUP, Foster, republic, border, Sinn, Neill **Topic-18 (Coronavirus)**: covid, coronavirus, pandemic, cummings, Downing, street, question, adviser, Johnson, questions
$\mathcal{GWE}_B$	**Dim-9 (Weapons)**: Hossein, warhead, Gangnam, interceptor, missiles, bomb, enriched, clerical, Quds, Ballistic **Dim-51 (transportation)**: Persia, Ibn, Arriva, Transpennine, Merseyrail, fax, Mesopotamia, BBBofC, Crosscountry,Daren **Dim-69 (Military)**: ang, corps, Muhammadu, commandant, army, commander, graduated, ante, military, Buhari

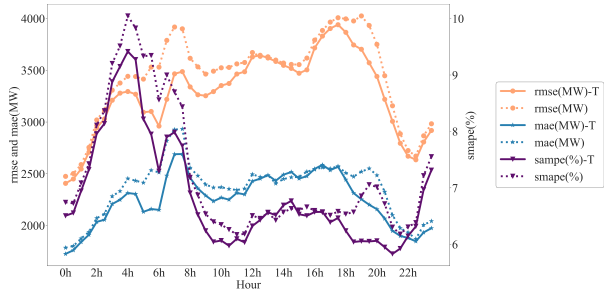

Fig. 3. $\overline{rmse}$, $\overline{mae}$, and $\overline{smape}$ on different hours. The dashed and solid lines are forecasting without and with textual features.

B. Local correlations

Figure 5 lists two days with the most negative (2021-02-20) and positive (2021-01-22) coefficients of driver-T, for example. Each subplot contains the textual features (y-axis) and their coefficients (x-axis) of the LIME model. The word frequencies of coronavirus-T, coronavirus-D, and pandemic-D serve negative roles on both days. At the same time, the coefficients of the other features vary, which is reasonable from a local view.

C. Causality effects

We tested the causality effect of the text features from the subsection IV-A and IV-B separately concerning each of the 48 half-hours of the coming day's load. We kept the treatment affects τs in Formula 7 corresponding to $p < 0.05$, set the rest to 0 and plotted the τ distribution of each feature over the day, as shown in Figure 6.

According to [48], the causality effect around 10% to 20% is significant. In Figure 6, the coefficients are mainly gathered around 0, indicating a weak causality effect. We also notice that the causality distributions for some features are flat. For example, pandemic-D, which shows a negative causality effect stretched to -30%, is evidence that the word frequency of 'pandemic' in news descriptions causes loads forecasting negatively. Topic-18, also related to coronavirus, shows a relatively strong negative correlation with load in Figure 4. However, in Figure 6, the coefficients for Topic-18 all stack up around 0. We, therefore, cannot conclude that there is a robust causal relationship between Topic-18 and load forecasting. This fact again confirms that correlation and causality are different aspects.

D. Considerations

This work proceeds by starting with essential forecasting of the electricity load from the UK with the calendar and temperature features, continuing to explore the textual features that are helpful for forecasting. Followed by some statistical tests to show the significant improvement and analysis of the explanations in key findings.

There is still room for some discussion. The causality effect distributions in Figure 6 show that several textual features have multiple modes, where some modes are centered on coefficients with absolute value $\geq 10\%$. For instance, a negative mode is identified in Dim-9, showing that there exist conditions that this dimension (related to tensions in the Middle East and assumed consequences on the oil market) is predicted to have a negative causal impact on electricity demand. This illustrates the potential of such text-based features to enrich predictive analysis applied to forecasting total load at the national or regional level. Another example is the Dim-69: at the beginning, it was included among the features selected through the Granger test. However, a second verification, related to the military, suggests a weak relation with electric load in the UK and is difficult to observe on the day ahead time scale. It is then considered that: i) this feature has, in general, a weak correlation with the load in every hour of the day, and every season or day category (as seen in Figure 4, and ii) its causality score in Figure 6 is dense around zero. These facts suggest that Dim-69 has been identified initially because of spurious correlation and that further analysis and human evaluation are necessary to understand the importance of every feature to be fed to the forecast model.

Generally, a low Pearson correlation corresponds to low causality, but some cases show high correlation and low causality, such as Topic-18. This may be due to coincidence, and we have no evidence that Topic-18 caused the change in electricity load.

The study depends on the dataset used, containing the period of the COVID-19 pandemic in the UK. This rare event influences the results. In the other case, it is good to verify that the method proposed identified the keywords related to the COVID-19 pandemic, which show the most relevance for load variation.

It is possible now to provide answers to the three main questions listed in Section I-C. The first research question

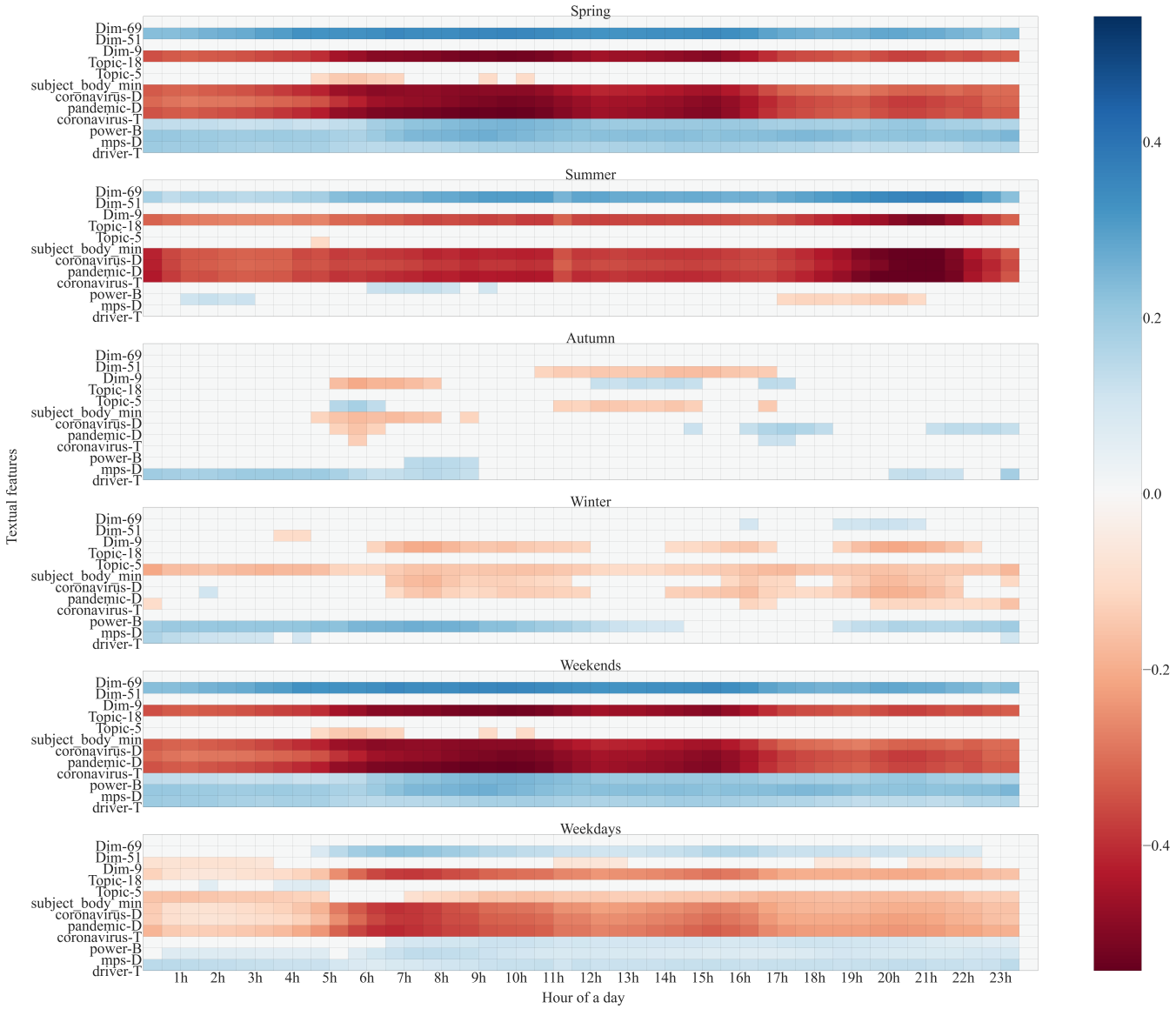

Fig. 4. Pearson coefficients for textual features and load. Colored grids with $p < 0.05$ stand for significant Pearson correlations. Blue ones are with a positive correlation between the feature and the hour. The red ones stand for the negative correlations.

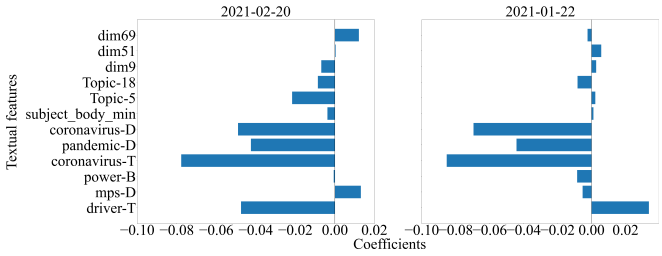

Fig. 5. Feature coefficients from LIME model

asked *IF* it is possible to extract valuable information from news in order to improve electricity demand prediction. The answer is yes, and this can be seen in several results of this work, such as in Table II, which shows the improvements of the forecasting algorithm designed on regression, temperature, and calendar information when adding features extracted from textual contents. Also, Figure 4 shows the correlation at different hours, seasons, and day types of the most relevant textual features. Moreover, the text-based model improves the official standard by around 4%, 11%, and 10% in terms of RMSE, MAE, and SMAPE.

The second question was *HOW* to treat textual information in order to extract valuable features to improve demand prediction. We would answer the question with the best performance textual features presented in the experiments: word frequency counting, sentiment scores, and global word embeddings. In particular, the last method is expected to be more robust to new keywords relative to new concepts or events that the public has not yet experienced.

The third research question posed was to understand *WHY* the improved performance was observed and to explain the phenomena identified. We would explain some key findings in the word frequency analysis. Firstly, the effect of the recent

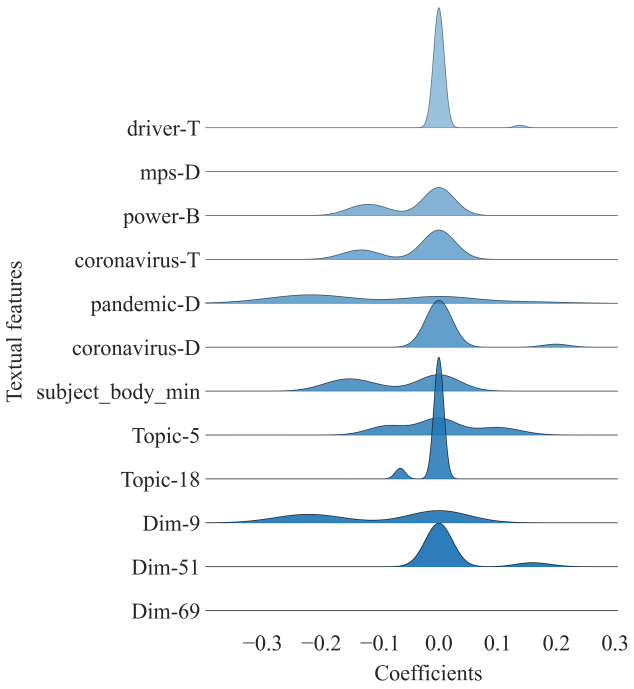

Fig. 6. Causality effects from the Double ML in half-hours. The x-axis is the causality treatment effects, and the y-axis is the textual features.

COVID-19 pandemic appears as keywords. It is thought that the mechanism identified by the algorithm is related to the reduced demand due to lockdowns enforced in the years 2020-21. These keywords show obvious negative correlations and causality effects. Secondly, the news related to Northern Ireland politics impacts the electricity demand in Topic-5. The interpretation is that these are a symptom of more generic political instability or may result from economic problems that cause a demand reduction. Thirdly, the tension in the Middle East, identified by Dim-9 mainly related to weapons. It is hypothesized that international tension in the oil-rich region may impact the economy and hence the electric load.

The work also has some limitations: Firstly, this work avoids exploring further consequences of social sciences, economics, or energy policy findings. The focus is kept on testing the main hypotheses and producing results that can be explored further. Secondly, the work is carried out on national aggregated electric demand, which is a parameter with a relatively stable pattern. Therefore day-ahead forecast errors are low, and it is more difficult to identify the effect of the additional data tested.

V. Conclusions

This paper studied the link between unstructured textual information in news and electricity demand. The overall methodology can be summarised as follows: firstly, textual information in the news is converted into numerical time series, including count features, word frequencies, sentiment scores, topic distributions, and word embeddings according to different methods of TextBlob, LDA model, and GloVe. Secondly, after the Granger test aimed at removing spurious correlations, the rest features are fed to an existing load forecasting algorithm working with known predictors such as calendar information and temperature values. Finally, the performance is compared, and the inputs are analysed to understand the mechanism of the news affecting electricity load.

The study was carried out on the datasets of news and electricity demand related to the United Kingdom for 2016-2021. In general, reduced performance improvements in the region of 4%, 11%, and 10% in RMSE, MAE, and SMAPE are observed. The best-performing method is a feature combination model with word frequency from news titles, sentiment scores and GloVe word embeddings from news text bodies. These features identified keywords relative to the COVID-19 pandemic, the minimum subjectivity of public sentiments, and international conflicts.

This study, far from closing the subject, opens a new series of questions to be treated in further research. For example, the results must be replicated on other datasets (news and load), possibly in different countries and with different spatial resolutions, to reduce the average effect present in national electricity demand. Since the effect of some social events has a more prolonged impact, it is better to replicate the study on longer horizons. Other methods for data analysis and NLP are worth trying; examples are testing n-grams instead of single keywords and using more complicated deep networks. Probabilistic forecasting is another research scenario to benchmark the performance of our proposed approach against other metrics, such as sharpness, CRPS, and reliability, and to test the method against other challenges, such as forecasting extreme loads. It is interesting to understand if the relationship discovered between the textual data and the electricity demand can also be explained through other variables, such as economic or criminal activity; Finally, more fine-grained experiments should focus only on the situations where the existing methods produce higher errors to verify if social aspects help reduce significant or potentially more influential forecast errors.

Acknowledgement

The author Yun BAI was supported by the program of the China Scholarship Council (CSC Nos. 202106020064).

REFERENCES

[1] M. Joshi, D. Das, K. Gimpel, and N. A. Smith, "Movie reviews and revenues: An experiment in text regression," in *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, 2010, pp. 293–296.

[2] M. P. Clements and U. Fritsche, "Text-based data and forecasting: Editor's introduction," pp. 1476–1477, 2020.

[3] D. Obst, "Textual data and transfer learning for time series forecasting," Ph.D. dissertation, Aix-Marseille, 2021.

[4] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decision support systems*, vol. 50, no. 1, pp. 164–175, 2010.

[5] R. P. Schumaker and H. Chen, "A discrete stock price prediction engine based on financial news," *Computer*, vol. 43, no. 1, pp. 51–56, 2010.

[6] P. Resnik, A. Garron, and R. Resnik, "Using topic modeling to improve prediction of neuroticism and depression in college students," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1348–1353.

[7] F. Rodrigues, I. Markou, and F. C. Pereira, "Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach," *Information Fusion*, vol. 49, pp. 120–129, 2019.

[8] N. Peng, K. Li, and Y. Qin, "Leveraging multi-modality data to airbnb price prediction," in *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*. IEEE, 2020, pp. 1066–1071.

[9] Y. Zhang, P. Siriaraya, Y. Kawai, and A. Jatowt, "Analysis of street crime predictors in web open data," *Journal of Intelligent Information Systems*, vol. 55, no. 3, pp. 535–559, 2020.

[10] X. Zhang, H. Saleh, E. M. Younis, R. Sahal, and A. A. Ali, "Predicting coronavirus pandemic in real-time using machine learning and big data streaming system," *Complexity*, vol. 2020, 2020.

[11] Y. Huang, W. Huang, S. Yan, H. Wang, and J. Yan, "The research on the forecast of tourism demand based on baidu search index–taking beijing as an example," in *2021 13th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. IEEE, 2021, pp. 91–94.

[12] Y. Bai, X. Li, H. Yu, and S. Jia, "Crude oil price forecasting incorporating news text," *International Journal of Forecasting*, vol. 38, no. 1, pp. 367–383, 2022.

[13] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting twitter sentiment analysis," in *Proceedings of the second international workshop on issues of sentiment discovery and opinion mining*, 2013, pp. 1–9.

[14] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from twitter text," *Journal of Computational Science*, vol. 36, p. 101003, 2019.

[15] B. Xue, C. Fu, and Z. Shaobin, "A study on sentiment computing and classification of sina weibo with word2vec," in *2014 IEEE International Congress on Big Data*. IEEE, 2014, pp. 358–363.

[16] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective lstms for target-dependent sentiment classification," *arXiv preprint arXiv:1512.01100*, 2015.

[17] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1422–1432.

[18] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[19] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," in *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, 1998, pp. 159–168.

[20] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[22] T. Joachims, "A probabilistic analysis of the rocchio algorithm with tfidf for text categorization." Carnegie-mellon univ pittsburgh pa dept of computer science, Tech. Rep., 1996.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[24] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] Z. Zhou, J. Zhao, and K. Xu, "Can online emotions predict the stock market in china?" in *International conference on web information systems engineering*. Springer, 2016, pp. 328–342.

[27] A. Tom, N. Joel, B. Steven *et al.*, "Nltk documentation," https://www.nltk.org.

[28] C. Kaur and A. Sharma, "Twitter sentiment analysis on coronavirus using textblob," EasyChair, Tech. Rep., 2020.

[29] R. Bose, P. Aithal, and S. Roy, "Sentiment analysis on the basis of tweeter comments of application of drugs by customary language toolkit and textblob opinions of distinct countries," *Int. J*, vol. 8, 2020.

[30] A. A. Chaudhri, S. Saranya, and S. Dubey, "Implementation paper on analyzing covid-19 vaccines on twitter dataset using tweepy and text blob," *Annals of the Romanian Society for Cell Biology*, pp. 8393–8396, 2021.

[31] X. Li, Y. Bai, and Y. Kang, "Exploring the social influence of the kaggle virtual community on the m5 competition," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1507–1518, 2022.

[32] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[33] X. Li, W. Shang, and S. Wang, "Text-based crude oil price forecasting: A deep learning approach," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1548–1560, 2019.

[34] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[35] D.-H. Kim, E.-K. Lee, and N. B. S. Qureshi, "Peak-load forecasting for small industries: A machine learning approach," *Sustainability*, vol. 12, no. 16, p. 6539, 2020.

[36] R. Porteiro, S. Nesmachnow, and L. Hernández-Callejo, "Short term load forecasting of industrial electricity using machine learning," in *Ibero-American Congress of Smart Cities*. Springer, 2019, pp. 146–161.

[37] W. Xiang, P. Xu, J. Fang, Q. Zhao, Z. Gu, and Q. Zhang, "Multi-dimensional data-based medium-and long-term power-load forecasting using double-layer catboost," *Energy Reports*, vol. 8, pp. 8511–8522, 2022.

[38] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[39] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[40] J. Aldrich, "Correlations genuine and spurious in pearson and yule," *Statistical science*, pp. 364–376, 1995.

[41] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," 2018.

[42] V. Chernozhukov, M. Goldman, V. Semenova, and M. Taddy, "Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels," *arXiv*, pp. arXiv–1712, 2017.

[43] "Entso-e. transparency platform," https://transparency.entsoe.eu.

[44] D. Obst, J. De Vilmarest, and Y. Goude, "Adaptive methods for short-term electricity load forecasting during covid-19 lockdown in france," *IEEE transactions on power systems*, vol. 36, no. 5, pp. 4754–4763, 2021.

[45] B. Wu, L. Wang, S.-X. Lv, and Y.-R. Zeng, "Effective crude oil price forecasting using new text-based and big-data-driven model," *Measurement*, vol. 168, p. 108468, 2021.

[46] L. Chang, D. Pedrno, and M. Ambrosio, "News-crawler," https://github.com/LuChang-CS/news-crawler.

[47] S. Matthew, "Bbc news front page archive," https://dracos.co.uk/made/bbc-news-archive/archive.php.

[48] V. Chernozhukov, H. Kasahara, and P. Schrimpf, "Causal impact of masks, policies, behavior on early covid-19 pandemic in the us," *Journal of econometrics*, vol. 220, no. 1, pp. 23–62, 2021.