IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 16 November 2022, accepted 30 November 2022, date of publication 1 December 2022, date of current version 7 December 2022.

*Digital Object Identifier 10.1109/ACCESS.2022.3226324*

RESEARCH ARTICLE

# Explainable Automatic Industrial Carbon Footprint Estimation From Bank Transaction Classification Using Natural Language Processing

**JAIME GONZÁLEZ-GONZÁLEZ[1], SILVIA GARCÍA-MÉNDEZ[1], FRANCISCO DE ARRIBA-PÉREZ[1], FRANCISCO J. GONZÁLEZ-CASTAÑO[1], AND ÓSCAR BARBA-SEARA[2]**

[1]Information Technologies Group, atlanTTic, Telecommunication Engineering School, University of Vigo, 36310 Vigo, Spain
[2]CoinScrap Finance S. L., 36002 Pontevedra, Spain

Corresponding author: Silvia García-Méndez (sgarcia@gti.uvigo.es)

This work was supported in part by Xunta de Galicia, Spain, under Grant ED481B-2021-118 and Grant ED481B-2022-093; and in part by the Spanish National Funds through Centro Para el Desarrollo Tecnológico Industrial (Center for Industrial Technological Development) under Project EXP00146826/IDI-20220298.

**ABSTRACT** Concerns about the effect of greenhouse gases have motivated the development of certification protocols to quantify the industrial carbon footprint (cf). These protocols are manual, work-intensive, and expensive. All of the above have led to a shift towards automatic data-driven approaches to estimate the cf, including Machine Learning (ml) solutions. Unfortunately, as in other sectors of interest, the decision-making processes involved in these solutions lack transparency from the end user's point of view, who must blindly trust their outcomes compared to intelligible traditional manual approaches. In this research, manual and automatic methodologies for cf estimation were reviewed, taking into account their transparency limitations. This analysis led to the proposal of a new explainable ml solution for automatic cf calculations through bank transaction classification. Consideration should be given to the fact that no previous research has considered the explainability of bank transaction classification for this purpose. For classification, different ml models have been employed based on their promising performance in similar problems in the literature, such as Support Vector Machine, Random Forest, and Recursive Neural Networks. The results obtained were in the 90 % range for accuracy, precision, and recall evaluation metrics. From their decision paths, the proposed solution estimates the $co_2$ emissions associated with bank transactions. The explainability methodology is based on an agnostic evaluation of the influence of the input terms extracted from the descriptions of transactions using locally interpretable models. The explainability terms were automatically validated using a similarity metric over the descriptions of the target categories. Conclusively, the explanation performance is satisfactory in terms of the proximity of the explanations to the associated activity sector descriptions, endorsing the trustworthiness of the process for a human operator and end users.

**INDEX TERMS** Explainable artificial intelligence, machine learning, natural language processing, carbon footprint, banking.

## I. INTRODUCTION

### A. RESEARCH GAP AND MOTIVATION

Concerns about climatic change [1], [2] related to the increasing emission of greenhouse gases (ghg) led 187 countries to sign the Paris Agreement[1] in 2015. This accord expressed the need for policies and regulations on ghg emissions such as carbon dioxide ($co_2$). The so-called carbon footprint (cf) can be defined as the amount of ghg released to the atmosphere

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

[1]Available at https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement, November 2022

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

throughout the life cycle of a product or human activity [3]. Over the years, there have been many proposals to estimate the cf of different entities [4], including individuals, families, industries, and geographical bodies such as cities [5].

The motivations for the calculation of cf are diverse, with compliance with environmental legislation and the certification of industrial sustainability (iso 14064[2]) being two of the most relevant reasons. Another relevant inducement is self-checking to avoid environmental taxes [6] and attract funding from ecologically-minded investors [7]. Moreover, individuals, especially young people, have pressing concerns regarding the effects of climate change [8], [9]. Consequently, diverse tracking applications allow end users to estimate and reduce their cf [10].

cf estimation solutions can be divided into manual and automatic approaches:

**Manual solutions.** For individuals, manual calculator applications require estimates of consumption habits, travel, *etc.*, as input data. These applications employ predefined formulae [11]. For industrial certifications, there exist consulting companies, such as aecom[3] and kpmg[4] whose environmental services include cf estimation.

**Automatic solutions.** Some examples are the *DO*,[5] *Enfuce*[6] and *Joro*[7] apps. Supervised approaches rely on the Classification of Individual Consumption by Purpose (coicop[8]) by the United Nations or other categories of consumption habits. Bank transactions are useful for individuals useful [12]. For industries, little Enterprise Resource Planning (erp) includes cf estimation [13].

As far as we know, although automatic estimation of cf from bank transaction descriptions has already been considered for end users, it is a novel problem in the industry. In fact, the explainability of industrial cf estimation based on the automatic classification of bank transactions has not yet been considered in previous research, as supported by the state-of-the-art discussion in Section II.

### B. CONTRIBUTION

In this paper, we propose an explainable automatic solution for industrial cf estimation based on a supervised bank transaction classification model. The training set was labeled as coicop classes.

Regrettably, classification tasks performed by Machine Learning (ml) models are often opaque [14], which may affect customer trust, especially in industrial contexts; hence, there is a growing interest in Explainable Artificial Intelligence (xai). Explainability methodologies allow for the extraction of intrinsic knowledge about the models' decisions.[9]

Departing from a categorization model combining ml with Natural Language Processing (nlp) techniques, the main contribution of this study lies in the proposal of the automatic explainability of cf estimation decisions. As previously mentioned, no authors have considered this aspect despite its relevance, for example, to examine consultancy analytics. The methodology extracts a set of relevant words for the classifier, and this word set is then validated with a similarity metric by comparing it with descriptions of the corresponding activity sectors.

### C. PAPER ORGANIZATION

The remainder of this paper is organized as follows. Section II reviews the state of the art in bank transaction classification applied to cf calculation using ml models and focuses on the explainability feature. Section III describes the proposed architecture for explainable automatic industrial cf estimation. Section IV presents the experimental data-set and implementations used, along with the results obtained in terms of classification and explainability. Finally, Section V summarizes the conclusions and proposes future research.

## II. RELATED WORK

Many previous studies have applied ml in fields such as E-commerce [15], incident management in information systems [16], and medical record analysis [17]. In finance [18], [19], ml models have been considered for detecting financial opportunities in social networks [20], fraud [21], [22], market sentiment [23], risk [24], accounting [25], and financial transaction classification [26].

In particular, bank transaction classification is a type of short-text classification that was already covered in our previous work [27]. The latter topic has been applied to problems among those in which intelligent budget management deserves attention [28], [29], [30].

Nevertheless, no previous work on bank transaction classification had an xai perspective (with the sole exception of Kotios et al. [31], although it did not involve any nlp methodology) nor considered industrial cf estimation.

The base classification methodologies involved are numerous and include simple Naive Bayes classifiers [32], supervised learning models such as Random Forests (rf) [33], [34], and Support Vector Machine (svm) [35], [36], along with more complex approaches based on Deep Learning (dl) and Neural Networks (nn) [37], [38].

[2] Available at `https://www.iso.org/standard/66453.html`, November 2022

[3] Available at `https://aecom.com/services/environmental-services`, November 2022

[4] Available at `https://home.kpmg/xx/en/home/insights/2020/12/environmental-social-governance-esg-and-sustainability.html`, November 2022

[5] Available at `https://www.diva-portal.org/smash/get/diva2:1604075/FULLTEXT01.pdf`, November 2022

[6] Available at `https://enfuce.com`, November 2022

[7] Available at `https://www.joro.app`, November 2022

[8] Available at `https://unstats.un.org/unsd/class/revisions/coicop_revision.asp`, November 2022

[9] Available at `https://www.darpa.mil/program/explainable-artificial-intelligence`, November 2022

The first solutions for cf estimation typically follow official protocols and practices[10] and rely on manual calculations [11], [39]. These protocols are time consuming and expensive to apply at the industrial level. More recent solutions oriented to end users have performed automatic cf estimation from bank transactions [12] and employed social networking [40] to foster user engagement [41].

End users are mainly motivated by environmental awareness and may be less concerned about the decision transparency of solutions. Conversely, industrial users may obtain important advantages from the application of automatic methodologies based on enterprise data, but solution transparency must be provided. In this regard, the incorporation of ai in Industry 4.0 has boosted the application of xai strategies in recent years [42], [43] to shed some light on the decisions of automatically supervised [44] and unsupervised [45], [46] learning models. Furthermore, explainability allows the prediction of behavior of these algorithms [47].

The existence of different explainability approaches is motivated by the variety of learning algorithms:

- **Model-agnostic explainability**. It considers ml models as black boxes and applies reverse engineering to infer their behavior.
  - **Model induction**. It consists of a counterfactual study of feature changes [48] or a correlation analysis of features and outputs [49], [50].
  - **Local explanation**. It exploits local linear interpretable models that match the results of those under analysis [51], [52]. These local explanations can be enhanced using additional contextual or semantic information [53].
- **Model-dependent explainability**. It is based on the inherent structure of ml models.
  - **Interpretable models**. Certain learning techniques are easily understandable to humans, as in the case of Decision Tree (dt) [54], [55], [56], rf [57], [58], and svm [59].
  - **Deep explanations**. Variations in dl models allow the determination of explainable features by decomposing the decision into the contributions of the input features [60] or by inferring the transfer function between layers [61].

To the best of our knowledge, this study represents the first attempt to explain industrial cf estimations from the automatic classification of enterprise bank transactions. The proposed approach takes advantage of the different ml models. Therefore, it follows a model-agnostic explainability strategy. Finally, an automatic validation of the explanation quality is provided.

## III. METHODOLOGY

Figure 1 illustrates the modular scheme of the proposed solution. The frames in white represent the elements within the processing pipeline, while the frames in blue represent external sources. Gray blocks correspond to higher-level tasks, as described in the independent subsections. This section aims to provide a conceptual perspective. Detailed implementations are described in Section IV.

In summary, the classification module labels the bank transactions used to estimate the cf. Then, the explainability module automatically generates and validates the descriptions associated with the classifier decisions.

### A. PRE-PROCESSING

The features used as input data for the classification task were engineered from textual bank transaction data. For this purpose, the text was processed using the following nlp techniques:

- **Numbers' removal**. Bank textual data usually contain quantitative information such as the receiver's bank account, receipt number, and product codes. These numbers are typically irrelevant for classification purposes because they are transaction specific.
- **Terms' reconstruction**. As bank descriptions are rather limited in length, relevant terms may be abbreviated or replaced with acronyms. Thus, these terms need to be expanded into natural language.
- **Removal of symbols and diacritic marks**. All symbols (*e.g.*, asterisks, hyphens, *etc.*), accents, diacritic marks, and diaereses were removed prior to text lemmatization.
- **Stop words and code removal**. Words with little semantic load, such as determiners, prepositions, general-usage verbs, and alphanumeric codes (*e.g.*, customer identifiers), are removed.
- **Text lemmatization**. Finally, the remaining terms are split into tokens and converted into lemmas.

### B. CLASSIFICATION MODULE

Once the processed bank transaction descriptions contain mostly semantically meaningful terms, the classification task is performed.

#### 1) FEATURE ENGINEERING

Before training the ml models, the outcome of the pre-processing module is converted into vectors. Specifically, the resulting terms of each bank transaction description are transformed into wordgram elements, *i.e.*, complete words, provided that our final goal is explainability.

#### 2) CLASSIFICATION

Transactions are classified using learning models that fulfill two requirements: (*i*) high classification performance of the target labels used for cf estimation (see Section IV), and (*ii*) straightforward extraction of self-explainable features from the trained estimators to fill the explainability templates. Based on their suitability in explainability research in the literature, svm, rf, and Recursive Neural Networks (rnn) were selected.

[10] Available at `https://ghgprotocol.org/standards`, November 2022

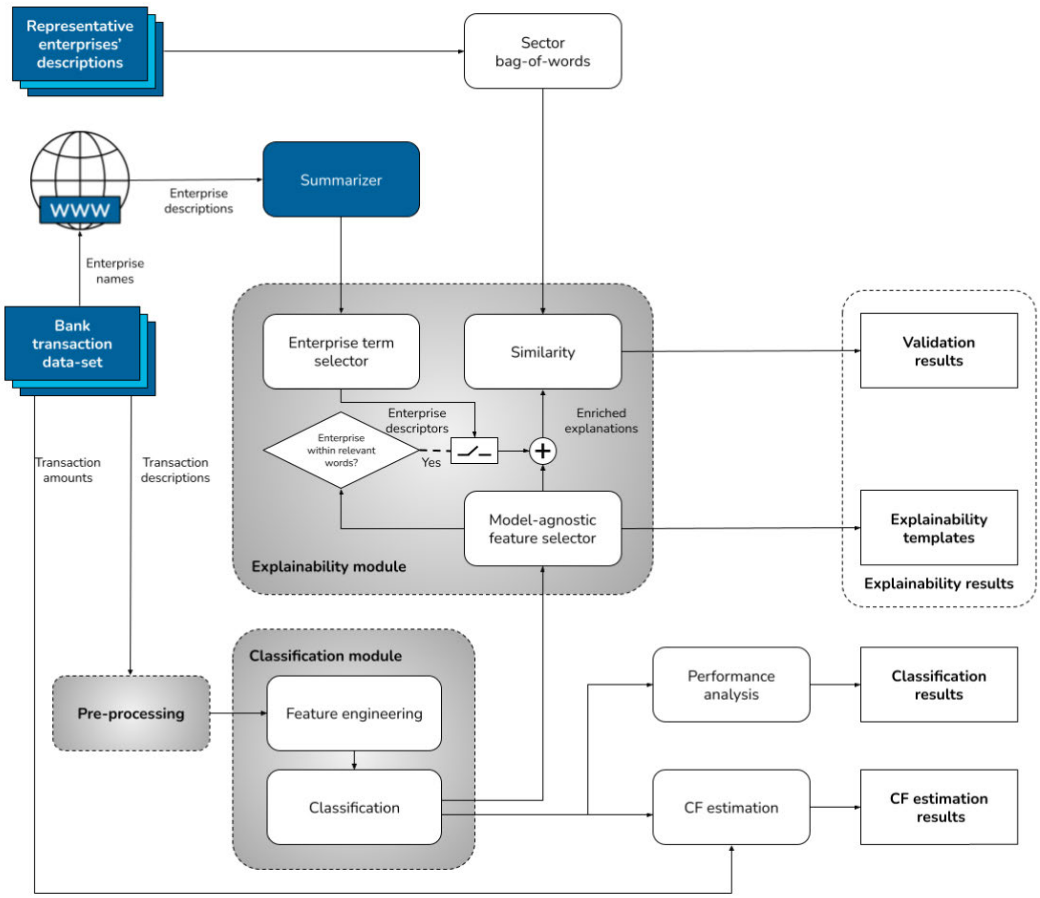



**FIGURE 1.** System architecture.

### C. EXPLAINABILITY MODULE

The goal of the explainability methodology is to provide a human operator with an in-depth understanding of the classification process and validate the corresponding relevant explanation terms with a metric of proximity between these terms and the descriptions of the cf categories. In principle, the explanatory terms are those that are considered relevant during training by the model.

However, due to the combined effect of pre-processing and feature engineering on the short bank transaction textual data, the explanations are enriched as follows:

#### 1) ENTERPRISE TERM SELECTION

Sometimes, the descriptions of transactions include explicit references to particular enterprises. By identifying these enterprises (see Section IV), it is possible to retrieve their descriptions from the Internet, which are likely to be representative of their activity sector. These descriptions were pre-processed using the same method described in Section III-A. The *summarizer* extracts all the nouns in the processed descriptions, and from these, the *enterprise term selector* takes the most representative ones, as detailed in Section IV-B4.

Additionally, the *similarity* calculation requires the collection of representative terms for each sector. Therefore, a bag-of-words is generated per sector using descriptions of the most representative enterprises (see Section IV).

#### 2) MODEL-AGNOSTIC FEATURE SELECTION

Because the classification module employs different ml models with particular internal structures, the system follows a model-agnostic approach. The latter method creates a local surrogate [44] model to select the explanatory terms for each bank transaction. The *model-agnostic feature selector* recursively analyzes the feature relevance by removing particular features (the deeper the impact, the higher the relevance). If any enterprise name is present in the initial batch of explanation terms for a bank transaction, those explanation terms are expanded using the descriptors of the enterprise, thanks to the *enterprise term selector*.

#### 3) SIMILARITY

Given the expanded explanation sets, the explainability module computes a similarity metric between the explanation set for each bank transaction and the bag-of-words of the economic sector, as selected by the classifier. Previous

*La clasificación del movimiento* <transaction_id> *en la categoría* <output_category> *puede explicarse en order decreciente por los términos relevantes*: <term$_1$> ... <term$_n$>.

**Listing 1. Original template in Spanish.**

The classification of transaction <transaction_id> into the category <output_category> can be explained by relevant terms: (in decreasing order) <term$_1$> ... <term$_n$>.

**Listing 2. Template translated to English.**

authors have also used contextual and semantic information to enhance explainability [53], [62].

### D. CARBON FOOTPRINT ESTIMATION

Once the transactions are classified, the proposed system automatically obtains their estimated cf from the formulae of the sectors to which they are predicted to belong and the bank transaction amount, as described in Section IV-B5.

## IV. EXPERIMENTAL EVALUATION AND DISCUSSION

In this section, we present the experimental data-set and technical implementations.

### A. EXPERIMENTAL DATA-SET

The data-set is composed of 25,853 bank transactions issued by Spanish banks compiled by CoinScrap Finance SL.[11] Note that this data-set is comparable in size to that in our previous study on bank transaction classification [27].

It was downsampled using the `FuzzyWuzzy` Python library[12] to keep only those entries sufficiently representative and distinguishable. Those samples with descriptions with a similarity greater than 90 % were discarded. The downsampling process resulted in 2,619 transaction archetypes, with an average length of 10 words/73 characters.

The transactions are divided into three main categories: car and transport (*automóvil y yransporte*), enterprise expenditures (*gastos de empresa*), and commodities (*suministros*), and several subcategories. Thus, a multi-class transformation [63], [64] process is applied to combine the main categories with their respective subcategories to map the following coicop categories:

- **Car and transport - gas stations (*gasolineras*)** (coicop 7.2). Payments in gas stations.
- **Car and transport - private transport (*transporte privado*)** (coicop 7.3). Payments in private transport services.
- **Car and transport - public transport (*transporte público*)** (coicop 7.3). Purchase of public transportation tickets (buses and trains).

[11]Available at `https://coinscrapfinance.com`, November 2022

[12]Available at `https://pypi.org/project/fuzzywuzzy`, November 2022

**TABLE 1. Distribution of samples in the data-set.**

| Category | Percentage |
|---|---|
| Car and transport - gas stations | 23.18 % |
| Car and transport - private transport | 10.84 % |
| Car and transport - public transport | 9.00 % |
| Car and transport - flights | 11.34 % |
| Enterprise expenditures - parcel and courier | 7.25 % |
| Commodities - water bill | 16.80 % |
| Commodities - electricity bill | 16.15 % |
| Commodities - gas bill | 5.38 % |

- **Car and transport - flights (*vuelos*)** (coicop 7.3). Purchase of airline tickets.
- **Enterprise expenditures - parcel and courier (*paquetería y mensajería*)** (coicop 8.1). Payment of public and private postal services.
- **Commodities - water bill (*agua*)** (coicop 4.4). Water supply receipts.
- **Commodities - electricity bill (*electricidad*)** (coicop 4.5). Receipt of energy supply.
- **Commodities - gas bill (*gas*)** (coicop 4.5). Gas supply receipts.

Table 1 shows the distribution of transactions by economic sector. Regarding description lengths, for instance, the category commodities - electricity bill has, on average, 16 words per description, while car and transport - private transport has only 6 words per description. Bank transaction pre-processing reduces the overall average description size to 7 words/50 characters.

### B. IMPLEMENTATIONS

Experiments were performed on a computer with the following specifications:

- **Operating System**. Ubuntu 20.04.3 LTS 64 bits
- **Processor**. IntelXeon Platinum 8375C 2.9 GHz
- **RAM**. 64 gb DDR4
- **Disk**. 500 gb SSD

For clarity, the corresponding architecture in Section III is indicated for each implementation description.

#### 1) PRE-PROCESSING MODULE (IMPLEMENTATION OF SECTION III-A)

Diacritic marks, numbers, identifiers, and codes were removed using regular expressions. The same technique is used to reconstruct common acronyms, such as s.l. (*Sociedad Limitada*, Limited Company) or e.s. (*estación de servicio*, gas station). Stop word removal (including general-usage verbs such as "to be" and "do") is based on the Spanish stop word list from the `NLTK` Python library.[13] For tokenizing purposes, the same `NLTK` Python library[13] was used and the resulting

[13]Available at `https://www.nltk.org`, November 2022

tokens were lemmatized with the `spaCy` Python library[14] using the `es_core_news_sm` model.[15]

### 2) FEATURE ENGINEERING MODULE (IMPLEMENTATION OF SECTION III-B1)

The selected classification models require different vectorization processes. For the svc and rf models, the `CountVectorizer`[16] function from the `scikit-learn` Python library was used for wordgram extraction. After the preliminary tests, wordgrams (one word) and biwordgrams (two words) were extracted. The features were downsampled using `SelectPercentile`[17] function from the `scikit-learn` Python library to keep those with the highest correlation with the target variable. Prior knowledge led us to select the chi-squared score function [20].

For the lstm model, the `Tokenizer`[18] function from the `Keras` Python library was used. It converts text into sequences of token identifiers embedded in the network.

### 3) CLASSIFICATION MODULE (IMPLEMENTATION OF SECTION III-B2)

The following models were used:

- **Linear Support Vector Classification** (svc). `LinearSVC`[19] implementation from `scikit-learn` Python library.
- **Random Forest** (rf). `RandomForestClassifier`[20] implementation from the `scikit-learn` Python library.
- **Long Short-Term Memory** (lstm). The `Sequential`[21] structure and `LSTMLayer`[22] were implemented from the `Keras` Python library.

Hyperparameter selection for the svc and rf models was performed using the `GridSearchCV`[23] function of the `scikit-learn` Python library. Listings 3 and 4 detail the hyperparameter ranges and the final choices, respectively.

class_weight = [None, **balanced**],
loss = [**hinge**, squared_hinge],
max_iter = [50, **100**, 250, 500, 1000],
multi_class = [ovr, **crammer_singer**],
tol = [**1e$^{-10}$**, 1e$^{-9}$, 1e$^{-8}$, 1e$^{-7}$, 1e$^{-6}$, 1e$^{-5}$, 1e$^{-4}$, 1e$^{-3}$, 1e$^{-2}$],
penalty = [**l2**],
C = [1e$^{-4}$, 5e$^{-3}$, 1e$^{-3}$, 5e$^{-2}$, 1e$^{-2}$, 5e$^{-1}$, **1e$^{-1}$**, 1]]

**Listing 3. Hyperparameter selection for svc (best values in bold).**

n_estimators = [50, 100, 250, **500**, 1000, 2000],
max_depth = [10, 25, 50, **100**, 200],
max_leaf_nodes = [50, 100, **250**, 500],
criterion = [**gini**, entropy]

**Listing 4. Hyperparameter selection for rf (best values in bold).**

The configuration used for the lstm model included the `SpatialDropout`[24] layer (equivalent to `SelectPercentile`[17]). The final dropout percentage applied prior to `LSTMLayer`[22] was 20 % of the tensors. The drop percentage of `LSTMLayer`[22] was also 20 %.

### 4) EXPLAINABILITY MODULE (IMPLEMENTATION OF SECTION III-C)

The explainability methodology comprises two complementary processes: (*i*) the generation of explanations based on explainability templates and (*ii*) the validation of these explanations in terms of their consistency compared with human knowledge about the target sectors.

The similarity metric of the validation process uses the bag-of-words from the target sectors and descriptors of the enterprises in the experimental data-set as input. For the former element, CoinScrap Finance s.l. provided a corporate lexicon[25] created from the descriptions of the top six enterprises of each target sector, with ten representative nouns and five representative verbs each. Conversely, Spanish companies' names[26] and their descriptions[27] were extracted from the Internet.

The *enterprise term selector* chooses up to ten terms from each summarized description of an enterprise in the data-set. The *summarizer* followed the same steps as the pre-processing module by removing stop words, common verbs, numbers, and codes. The system creates a list of words for each enterprise from the resulting list of lemmas. If the

[14] Available at `https://spacy.io`, November 2022
[15] Available at `https://spacy.io/models/es#es_core_news_sm`, November 2022
[16] Available at `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html`, November 2022
[17] Available at `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selectio.SelectPercentile.html`, November 2022
[18] Available at `https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer`, November 2022
[19] Available at `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html`, November 2022
[20] Available at `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`, November 2022
[21] Available at `https://keras.io/api/models/sequential`, November 2022
[22] Available at `https://keras.io/api/layers/recurrent_layers/lstm`, November 2022
[23] Available at `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`, November 2022
[24] Available at `https://keras.io/api/layers/regularization_layers/spatial_dropout1d`, November 2022
[25] Available at `https://docs.google.com/spreadsheets/d/1Tq2l9An6DybVTHig_5O_5_KN-0VucgjSFT8wGQluahc/edit?usp=sharing`, November 2022
[26] Available at `https://guiaempresas.universia.es/localidad/MADRID`, November 2022
[27] Available at `https://docs.google.com/spreadsheets/d/1SNT4avp9ki4beD6tYCH27zE6FQpsQUXTdH5vuLVF0yc/edit?usp=sharing`, November 2022

list contains more than ten terms, only the ten most frequent terms are retained.

The *model-agnostic feature selector* performs recursive feature selection tests using the lime[28] Python library given its wide acceptance in the literature [44]. As previously explained, the previous features are enriched with the enterprise descriptors obtained by the *enterprise term selector* in case the bank description contains the name of a company to generate the explanation sets.

For similarity metrics between groups of terms, we considered two different approaches: *(i)* Jaccard similarity as a baseline [65], [66], and *(ii)* our own sophisticated metric based on lexical and semantic proximity [67]. The cosine distance was discarded provided that the terms in the descriptions had no logical ordering. Using a similarity metric, the system calculates the similarities between enriched bank transaction explanations and the bag-of-words of the target sectors so that the highest similarity can be expected between an enriched explanation of a bank transaction and its target sector according to the classification module.

#### 5) CARBON FOOTPRINT MODULE (IMPLEMENTATION OF SECTION IV-B5)

The cf, that is, the ghg emissions associated with a transaction, is directly related to the transaction amount. The conversion estimate depends on the sector:

- **Car and transport - gas stations**. $co_2$ emissions $CF_{gs}$ depend on fuel volume and the emission factor of the fuel $\epsilon_f$. As bank transactions do not include the type of fuel, the emission factor averages the emissions of gasoline and diesel. The volume is derived from the payment amount $p$ and the average price per liter $avp_f$ of fuel at transaction time.

$$CF_{gs} = \frac{p}{avp_f} \cdot \epsilon_f \tag{1}$$

- **Car and transport - private/public transport**. For private transport, it is necessary to first distinguish between taxi payments and other private services. We applied these keywords for this purpose. Each of these alternatives has its own emission factor, $\epsilon_t$ and $\epsilon_c$, respectively. Distances are calculated from the average prices per kilometer of the region $avp_t$ (for taxis) and $avp_c$ (for private companies) and the amount paid $p$. Prices are obtained from official and private pricing lists.

  For taxis:

$$CF_{taxi} = \frac{p}{avp_t} \cdot \epsilon_t \tag{2}$$

  For private companies, $CF_{comp}$ is defined similarly as $avp_c$ and $\epsilon_c$. For public transport, the Spanish Transport Ministry publishes average references for the price per kilometer[29] $avp_{pt}$. $co_2$ emissions depend on the travel distance. The emission factor for the transportation means considered is $\epsilon_{pt}$.

$$CF_{pt} = \frac{p}{avp_{pt}} \cdot \epsilon_{pt} \tag{3}$$

- **Car and transport - flights**. In this sector, the price per kilometer $avp_{fl}$ must be averaged, as it varies depending on the airline and plane model. Given this estimate and the payment amount $p$, we calculate the $co_2$ emission from the corresponding travel distance and the emission factor for a commercial aircraft $\epsilon_{fl}$.

$$CF_{fl} = \frac{p}{avp_{fl}} \cdot \epsilon_{fl} \tag{4}$$

- **Enterprise expenditures - parcel and courier**. Both private companies and public entities record parcel transport costs per kilometer $avp_{pc}$ on a yearly basis. From these data and the amount $p$, it is possible to estimate the shipment distance and, therefore, its cf from the emission factor $\epsilon_{pc}$.

$$CF_{pc} = \frac{p}{avp_{pc}} \cdot \epsilon_{pc} \tag{5}$$

- **Commodities - water bill**. Unlike the rest of the bank transactions, for water bills, we do not calculate ghg emissions but the total consumption of water $TWC$, which depends on the average price of the service $avp_w$ and the amount paid $p$.

$$TWC = \frac{p}{avp_w} \tag{6}$$

- **Commodities - electricity/gas bill**. The daily price per kWh $kwp_i$ for the $i$-day of the last month is publicly available. The consumption of electricity from a receipt with amount $p$ is estimated from the average price in the previous month. From the emission factor $\epsilon_e$ for electricity:

$$CF_e = \frac{p}{avp_e} \cdot \epsilon_e = \frac{p}{\frac{1}{m}\sum_{i=1}^{m} kwp_i} \cdot \epsilon_e \tag{7}$$

  where $m$ denotes the number of days in the previous month. $CF_g$ is defined similarly from $\epsilon_g$.

From the predicted class of transactions and their amounts, the system presents the users with the estimated volume of ghg associated with each transaction. Table 2 presents examples of transactions and their corresponding $co_2$ emissions in kilograms. Table 3 presents an example of water consumption estimated from a water bill transaction.

### C. CLASSIFICATION RESULTS

$K$-fold cross-validation is a common strategy for proper validation of prediction results [68]. In particular, we applied a 10-fold cross-validation, as implemented with the `StratifiedKFold`[30] function from the `scikit-learn`

[28] Available at `https://github.com/marcotcr/lime`, November 2022

[29] Available at `https://www.mitma.gob.es/transporte-terrestre/observatorios/observatorios-y-estudios` (Spanish), November 2022

[30] Available at `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectPercentile.html`, November 2022

**TABLE 2.** Samples of cf estimation results.

| Description | Amount (€) | Predicted sector | Parameters | $CO_2$ emission (kg) |
|---|---|---|---|---|
| BALLENOIL ALBAL | 80.0 | Car and transport<br>Gas stations | $avp_f$=1.83 €/L<br>$\varepsilon_f$=2.35 kg $CO_2$/L | 102.733 |
| LIC [NUM] TAXI MADRID | 10.1 | Car and transport<br>Private transport | $avp_t$=2.02 €/km<br>$\varepsilon_t$=0.17 kg $CO_2$/km | 0.855 |
| Tj-renfe virtual internet | 142.6 | Car and transport<br>Public transport | $avp_{pt}$=0.15 €/km<br>$\varepsilon_{pt}$=0.035 kg $CO_2$/km | 33.273 |
| COMPRA TARJ. [NUM] Ryanair-Madrid | 34.68 | Car and transport<br>Flights | $avp_{fl}$=0.05 €/km<br>$\varepsilon_{fl}$=0.192 kg $CO_2$/km | 133.171 |
| SE CORREOS Y TELEGRAFOS S (VILLENA) | 29.0 | Enterprise expenditures<br>Parcel and courier | $avp_{pc}$=1.3 €/km<br>$\varepsilon_{pc}$=0.158 kg $CO_2$/km | 3.525 |
| FACTURA DE GAS PM [NUM] [NUM] | 48.04 | Commodities<br>Gas bill | $avp_g$=0.1398 €/kWh<br>$\varepsilon_g$=0.203 kg $CO_2$/kWh | 69.757 |
| RECIBO IBERDROLA CLIENTES, S.A.U RECIBO [NUM] | 23.0 | Commodities<br>Electricity bill | $avp_e$=0.098 €/kWh<br>$\varepsilon_e$=0.25 kg $CO_2$/kWh | 58.673 |

**TABLE 3.** Sample of water consumption results.

| Description | Amount (€) | Predicted sector | Parameters | Water consumption (L) |
|---|---|---|---|---|
| RECIBO AGUA-[NUM]-BO. | 50.11 | Commodities - water bill | $avp_w$=0.0017 €/L | 29304.094 |

**TABLE 4.** Classification results.

| Model | Accuracy | Precision | Recall | Training time (s) |
|---|---|---|---|---|
| SVC | 93.72 % | 94.36 % | 92.34 % | 0.532 |
| RF | 89.18 % | 90.24 % | 86.36 % | 20.33 |
| LSTM | 92.34 % | 92.37 % | 90.97 % | 399.09 |

**TABLE 5.** Directly validated explanations for the rf classifier.

| Metric | Validated |
|---|---|
| Jaccard | 34.85 % |
| Linguistic proximity metric [67] | 46.89 % |

Python library, to calculate average accuracy, precision, recall, and training times. Table 4 presents the results for the svc, rf, and lstm models.

svc and lstm achieved over 90 % of accuracy. svc is the most time-efficient model. Regarding the training time, lstm was the most time-consuming. rf is an intermediate alternative with slightly lower performance. Consequently, the best model, considering the performance-time tradeoff, was svc, but the classification performance of the three models selected was similar.

### D. EXPLAINABILITY PERFORMANCE RESULTS

The rf model was used as the baseline because of its inferior performance, whereas svc was selected as the target classifier.

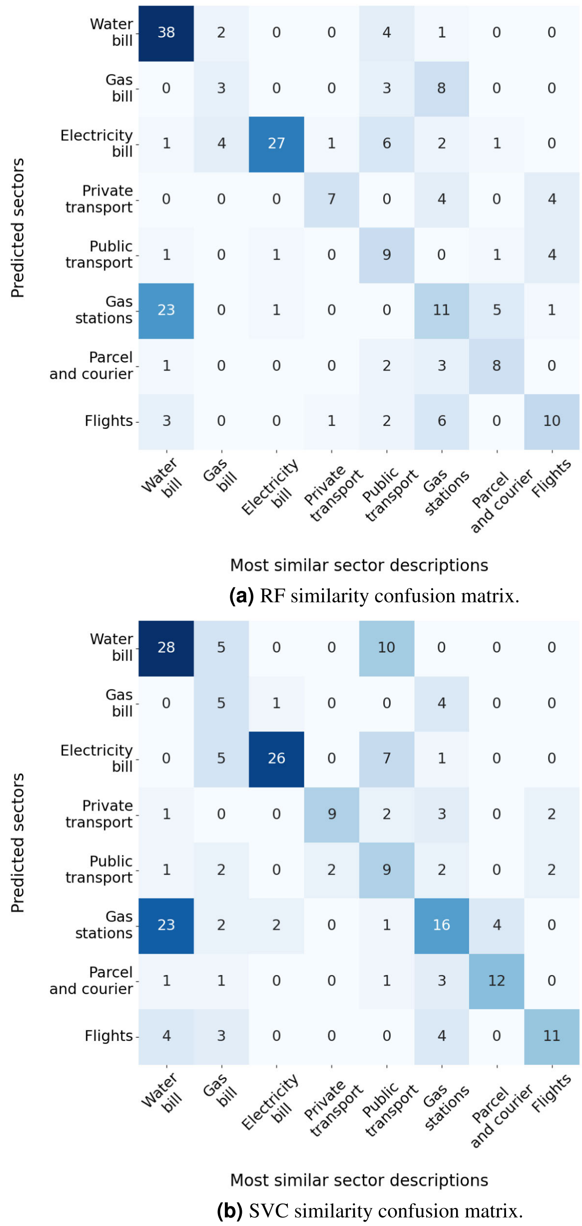
Table 5 shows the percentage of explanations for the rf baseline model that could be validated directly. An explanation is considered to be directly "validated" when the sector closest to a bank transaction explanation is predicted by the classifier.

As shown in Table 5, Jaccard similarity results in a lower percentage of directly validated explanations. Therefore, in the rest of the experiments, our linguistic metric [67] was used. This metric is well-suited to our goal, as it requires fewer terms per explanation than the Jaccard distance to detect similarity, and unlike the cosine distance, it does not rely on term ordering. The differences between the lists of terms in the bank transaction explanation sets generated for both models were analyzed. For rf, each explanation set contained 7.67 words on average, while 8.19 words on average in the case of svc. Similarities between pairs of explanation sets were then computed, resulting in an overall average similarity of 0.79. Note that the similar performances of both classification methods seem to be related to the similarity of their explanation sets. Thus, the explanation potential is consistent with the classification performance.

Table 6 shows the explanatory performance of both the models. For explanations that were neither obvious nor directly validated, we performed a second in-depth analysis to check their trustworthiness in a human operator. We divided them by manual inspection into "coherent" (when the human operator considered that the explanation was correct given the predicted sector) and "ambiguous" (when the human operator could not determine from the explanation itself the sector that was predicted by the classifier). Those "coherent" explanations that contained the name of a company of the target sector are obviously satisfactory and, as such, they were marked as "obvious". Finally, we considered "empty" those explanations whose similarity to all sectors is zero. This may

**TABLE 6.** Explanation performance.

| Model | Satisfactory | | | Unsatisfactory | |
|---|---|---|---|---|---|
| | Validated | Obvious | Coherent | Empty | Ambiguous |
| SVC | 48.13 % | 9.96 % | 15.77 % | 12.79 % | 13.35 % |
| RF | 46.89 % | 9.54 % | 15.35 % | 13.28 % | 14.94 % |




**(a)** RF similarity confusion matrix.

**(b)** SVC similarity confusion matrix.

**FIGURE 2.** Confusion matrices, predicted sectors versus most similar sector descriptions.

be due to the fact that lime fails to detect any representative term or no selected explanation term is representative enough (*i.e.*, simple alphanumerical codes). Therefore, the lower the percentage of ambiguous and empty explanations, the higher the trustworthiness.

The explanation performances were similar, which resulted from the use of a model-agnostic feature selector and the similar classification performance of both models. Satisfactory explanations exceeded 70 %, of which approximately 60 % could be automatically "validated". Regarding unsatisfactory explanations ("empty" and "ambiguous"), only over 12 % are "empty" and offer no information to a human operator.

Figure 2 illustrates the confusion matrices of the predicted sectors versus the most similar sector descriptions for the direct validation. The main deviation occurred for gas stations, the category with the shortest explanations, with only four words on average. These are frequently closest to the water bill sector description. Other common deviations exist between gas stations and the gas bill, and between the three categories of commodities.

Most of these deviations do not correspond to unsatisfactory results from the perspective of a human operator. For example, let us consider an explanation of a bank transaction that was predicted to belong to the commodities - electricity bill and was closer to car and transport - public transport. The explanation set contained the relevant terms 'energia', 'referencia', 'recibo referencia', 'recibo', 'energy', 'nxs', 'nexus', 'mandato nxs', 'nexus energia' and 'referencia mandato'. It includes several instances of the meaningful terms *energía* and *energy* that, in the Spanish context, are directly related to the electricity sector from the perspective of the human operator.

The system finally presents explanations by following the template in Listing 2. Some examples are:

- **Car and transport - gas stations.** *The classification of transaction 423 into the category car and transport - gas stations can be explained by relevant terms (in decreasing order)*: cedipsa, service (*servicio*), station (*estacion*), gas station (*estacion servicio*), payment (*pago*), cedipsa payment (*pago* cedipsa).
- **Car and transport - public transport.** *The classification of transaction 895 into category car and transport - public transport can be explained by relevant terms (in decreasing order)*: renfe, madrid, travelers (*viajeros*), renfe card (*tarjeta* renfe), renfe travelers (*viajeros* renfe), purchase (*compra*), travelers app (*viajeros* app), app, dev, card (*tarjeta*).
- **Enterprise expenditures - parcel and courier.** *The classification of transaction 1269 into category enterprise expenditures - parcel and courier can be explained by relevant terms (in decreasing order)*: mail (*correos*), mail payment (*pago correos*), purchase (*compra*), purchase payment (*pago de compra*), mail leganes (*correos leganes*), card (*tarjeta*).
- **Commodities - water bill.** *The classification of transaction 1514 into category commodities - water bill can be explained by relevant terms (in decreasing order)*: water (*agua*), water receipt (*recibo agua*), receipt (*recibo*), reference order (*referencia mandato*), reference (*referencia*), order (*mandato*), receipt reference (*referencia recibo*).

In these examples, the lists of relevant terms in the explanations are highly related to the respective sectors. These include *electricidad* (electricity), *gas* (gas), *agua* (water), and *estacion servicio* (gas station). Two of the explanations are obvious because they contain the names of companies offering the services (Cedipsa and Renfe), but they also contain other highly informative words. There are other generalist terms, such as *recibo* (receipt), *compra* (purchase), and *tarjeta* (card). Although they do not occupy the first positions in their respective lists, they are less relevant than sector-specific terms.

### E. COMPARISON WITH PRIOR WORK

The cf estimation has recently attracted significant commercial interest. However, there are few automatic solutions based on bank transaction classification in the literature owing to its novelty.

Although a few previous studies have applied bank transaction classification to industrial use cases, the classification performance achieved by other researchers on different finance-related problems is illustrative.

E. Folkestad et al. (2017) [28] exploited data from DBpedia[31] and Wikidata[32] for bank transaction classification. They reported 83.48 % accuracy using Logistic Regression (lr) (10.24 % less than with our approach). Moreover, E. Vollset et al. (2017) [29] augmented corporate data with external semantic resources to improve bank transaction classification. They obtained 92.97 % accuracy also with lr (0.75 % less than with our approach).

The nlp-based budget management solution by S. Allegue et al. (2021) [30] obtained similar results to our approach with an Adaptive Random Forest model, with a difference of only 1 % in precision.

The non nlp-based svm solution for cash flow prediction for small & medium enterprises by D. Kotios et al. (2022) [31] attained a precision that was only 0.2 % higher than ours, after trying many other algorithms.

Given the similar performance of the existing solutions, some contributions directly focus on the problem description. This is the case of the *Svalna* app by Andersson [12], an automatic carbon footprint estimation application based on users' transactions and environmental data from governmental agencies.

This apparent intrinsic high separability of the problem is consistent with our own results with the three different classification methodologies. Because the focus of our contribution is on classification explainability and given the small gap between methodologies in this and other works, and despite the advantages of rf for self-explainability [69], we have applied a model-agnostic explanation methodology.

[31]Available at `https://es.dbpedia.org`, November 2022

[32]Available at `https://www.wikidata.org`, November 2022

## V. CONCLUSION

In this study, a novel explainable solution for automatic industrial cf estimation from bank transactions is proposed, addressing the lack of transparent decision explanation methodologies for this problem. The explanation is especially important to trust the outcome of automatic processes, for them to replace more expensive alternatives, such as consultancy analytics. Indeed, even though automatic explainability has not been tackled in this domain, the study of the state of the art has also revealed that there are no previous works or existing commercial solutions for automatic industrial cf estimation based on bank transactions.

The original data source includes more than 25,000 bank transactions. It was annotated for classification using coicop categories.

The classification methodology for bank transactions followed a supervised learning strategy by combining ml with nlp techniques based on our approach in [27]. The widely used svm, rf, and lstm models achieve satisfactory performance levels of 90 % for all metrics, which is consistent with the results reported by other authors in the literature.

The agnostic explanation methodology extracts a set of relevant words for the classifier, and this explanation set is then validated with a similarity metric by comparing the set with the descriptions of carbon-intensive activity sectors. Despite the scarcity of content in industrial bank transaction descriptions, over 70 % of the explanations are satisfactory to a human operator, and 60 % have been automatically validated from company descriptions of the target sectors. Only 15 % of the explanations were ambiguous, and there is a margin for improvement in the rest (which we tag as ''empty'') if side information on alphanumeric codes of industrial activity is provided. We consider these results encouraging for further study on the automatic explainability of cf estimation in industrial sectors.

In summary, the highlights of this study are as follows:

- The main contribution of this study is a novel solution for automatic industrial cf estimation from bank transactions based on supervised ml and nlp techniques.
- The performance of the underlying bank transaction classification methodology is comparable to that of other researchers [30], [31].
- An experimental data-set composed of more than 25,000 bank transactions was used.
- Over 70 % of the natural language explanations automatically generated with a model-agnostic approach are satisfactory for end users. Of these, 60 % have been automatically validated. Less than 15 % are ambiguous.

Regarding the limitations of this study, the supervised classification methodology requires manual annotation of bank transactions for training purposes. In addition, the categories for cf estimation could change, depending on the activity sector. We chose a well-established reference, but finer details may be required to account for business-specific expenses.

In future work, we plan to extend this research to other main languages, enrich explanations with complementary enterprise information, and study the effect of hierarchical methodologies on categorization by leveraging the relations between target classes. We also plan to move towards a semi-supervised approach by combining the current solution with a rule scheme, such as those proposed by other authors [31]. Another possible line of research is the comparison of the model-agnostic approach to explainability with model-specific methodologies.

## ACKNOWLEDGMENT

The authors would like to thank the indebted to CoinScrap Finance S. L. for providing them with the experimental dataset.

## REFERENCES

[1] "The sustainable development goals report 2021," United Nations Dept. Econ. Social Affairs, New York, NY, USA, Tech. Rep., 6, 2021.

[2] V. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P. R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J. B. R. Matthews, Y. Chen, X. Zhou, M. I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield, "Global warming of 1.5°C. An IPCC special report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty," Intergovernmental Panel Climate Change (IPCC), Geneva, Switzerland, Tech. Rep., 15, 2018.

[3] T. Wiedmann and J. Minx, "A definition of 'carbon footprint,'" in *Ecological Economics Research Trends*. Hauppauge, NY, USA: Nova Science Publishers, 2008, ch. 1, pp. 1–11.

[4] D. Pandey, M. Agrawal, and J. S. Pandey, "Carbon footprint: Current methods of estimation," *Environ. Monitor. Assessment*, vol. 178, nos. 1–4, pp. 135–160, Jul. 2011.

[5] L. Ionescu, "Urban greenhouse gas accounting for net-zero carbon cities: Sustainable development, renewable energy, and climate change," *Geopolitics, Hist., Int. Relations*, vol. 14, no. 1, pp. 155–171, 2022.

[6] N. Zhu, Y. Bu, M. Jin, and N. Mbroh, "Green financial behavior and green development strategy of Chinese power companies in the context of carbon tax," *J. Cleaner Prod.*, vol. 245, Feb. 2020, Art. no. 118908.

[7] L. Ionescu, "Transitioning to a low-carbon economy: Green financial behavior, climate change mitigation, and environmental energy sustainability," *Geopolitics, Hist., Int. Relations*, vol. 13, no. 1, pp. 86–96, 2021.

[8] T. L. Milfont, "The interplay between knowledge, perceived efficacy, and concern about global warming and climate change: A one-year longitudinal study," *Risk Anal.*, vol. 32, no. 6, pp. 1003–1020, Jun. 2012.

[9] S. Luís, C.-M. Vauclair, and M. L. Lima, "Raising awareness of climate change causes? Cross-national evidence for the normalization of societal risk perception of climate change," *Environ. Sci. Policy*, vol. 80, pp. 74–81, Feb. 2018.

[10] S. Hoffmann, W. Lasarov, and H. Reimers, "Carbon footprint tracking apps. What drives consumers adoption intention?" *Technol. Soc.*, vol. 69, May 2022, Art. no. 101956.

[11] J. Mulrow, K. Machaj, J. Deanes, and S. Derrible, "The state of carbon footprint calculators: An evaluation of calculator design and user interaction features," *Sustain. Prod. Consumption*, vol. 18, pp. 33–40, Apr. 2019.

[12] D. Andersson, "A novel approach to calculate individuals carbon footprints using financial transaction data—App development and design," *J. Cleaner Prod.*, vol. 256, May 2020, Art. no. 120396.

[13] D. Zvezdov and S. Hack, "Carbon footprinting of large product portfolios. Extending the use of enterprise resource planning systems to carbon information management," *J. Cleaner Prod.*, vol. 135, pp. 1267–1275, Nov. 2016.

[14] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics*. Singapore: Springer, 2020, pp. 99–111.

[15] L. Tan, M. Y. Li, and S. Kok, "E-commerce product categorization via machine translation," *ACM Trans. Manage. Inf. Syst.*, vol. 11, no. 3, pp. 1–14, Sep. 2020.

[16] S. Silva, R. Pereira, and R. Ribeiro, "Machine learning in incident categorization automation," in *Proc. 13th Iberian Conf. Inf. Syst. Technol. (CISTI)*, Jun. 2018, pp. 1–6.

[17] G. T. Berge, O.-C. Granmo, T. O. Tveit, M. Goodwin, L. Jiao, and B. V. Matheussen, "Using the Tsetlin machine to learn human-interpretable rules for high-accuracy text categorization with medical applications," *IEEE Access*, vol. 7, pp. 115134–115146, 2019.

[18] J. Huang, J. Chai, and S. Cho, "Deep learning in finance and banking: A literature review and classification," *Frontiers Bus. Res. China*, vol. 14, no. 1, p. 13, Dec. 2020.

[19] S. Rakshit, N. Clement, and N. R. Vajjhala, "Exploratory review of applications of machine learning in finance sector," in *Advances in Data Science and Management* (Lecture Notes on Data Engineering and Communications Technologies), vol. 86. 2022, Singapore: Springer, pp. 119–126.

[20] F. De Arriba-Perez, S. Garcia-Mendez, J. A. Regueiro-Janeiro, and F. J. Gonzalez-Castano, "Detection of financial opportunities in micro-blogging data with a stacked classification system," *IEEE Access*, vol. 8, pp. 215679–215690, 2020.

[21] M. N. Ashtiani and B. Raahemi, "Intelligent fraud detection in financial statements using machine learning and data mining: A systematic literature review," *IEEE Access*, vol. 10, pp. 72504–72525, 2022.

[22] C. S. Kolli and U. D. Tatavarthi, "Fraud detection in bank transaction with wrapper model and Harris water optimization-based deep recurrent neural network," *Kybernetes*, vol. 50, no. 6, pp. 1731–1750, Jul. 2021.

[23] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, pp. 131662–131682, 2020.

[24] S. Bhatore, L. Mohan, and Y. R. Reddy, "Machine learning techniques for credit risk evaluation: A systematic literature review," *J. Banking Financial Technol.*, vol. 4, no. 1, pp. 111–138, Apr. 2020.

[25] C. Bardelli, A. Rondinelli, R. Vecchio, and S. Figini, "Automatic electronic invoice classification using machine learning models," *Mach. Learn. Knowl. Extraction*, vol. 2, no. 4, pp. 617–629, Nov. 2020.

[26] R. K. Jørgensen and C. Igel, "Machine learning for financial transaction classification across companies using character-level word embeddings of text fields," *Intell. Syst. Accounting, Finance Manag.*, vol. 28, no. 3, pp. 159–172, Jul. 2021.

[27] S. Garcia-Mendez, M. Fernandez-Gavilanes, J. Juncal-Martinez, F. J. Gonzalez-Castano, and O. B. Seara, "Identifying banking transaction descriptions via support vector machine short-text classification based on a specialized labelled corpus," *IEEE Access*, vol. 8, pp. 61642–61655, 2020.

[28] E. Folkestad, E. Vollset, M. R. Gallala, and J. A. Gulla, "Why enriching bus. transactions with linked open data may be problematic in classification tasks," in *Knowledge Engineering and Semantic Web* (Communications in Computer and Information Science), vol. 786. Cham, Switzerland: Springer, 2017, pp. 347–362.

[29] E. Vollset, E. Folkestad, M. R. Gallala, and J. A. Gulla, "Making use of external company data to improve the classification of bank transactions," in *Advanced Data Mining and Applications* (Lecture Notes in Computer Science). vol. 10604. Cham, Switzerland: Springer, 2017, pp. 767–780.

[30] S. Allegue, T. Abdellatif, and H. El Abed, "SBM: A smart budget manager in banking using machine learning, NLP, and NLU," *Concurrency Comput., Pract. Exper.*, p. e6673, Oct. 2021.

[31] D. Kotios, G. Makridis, G. Fatouros, and D. Kyriazis, "Deep learning enhancing banking services: A hybrid transaction classification and cash flow prediction approach," *J. Big Data*, vol. 9, no. 1, pp. 1–29, Oct. 2022.

[32] H. Gao, X. Zeng, and C. Yao, "Application of improved distributed naive Bayesian algorithms in text classification," *J. Supercomput.*, vol. 75, no. 9, pp. 5831–5847, Sep. 2019.

[33] N. N. A. Sjarif, N. F. M. Azmi, S. Chuprat, H. M. Sarkan, Y. Yahya, and S. M. Sam, "SMS spam message detection using term frequency-inverse document frequency and random forest algorithm," *Proc. Comput. Sci.*, vol. 161, pp. 509–515, Jan. 2019.

[34] Z. Taşkin and U. Al, "A content-based citation analysis study based on text categorization," *Scientometrics*, vol. 114, no. 1, pp. 335–357, 2018.

[35] M. Goudjil, M. Koudil, M. Bedda, and N. Ghoggali, "A novel active learning method using SVM for text classification," *Int. J. Automat. Comput.*, vol. 15, no. 3, pp. 290–298, 2018.

[36] K. Kim and S. Y. Zzang, "Trigonometric comparison measure: A feature selection method for text categorization," *Data Knowl. Eng.*, vol. 119, pp. 1–21, Jan. 2019.
[37] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, "Convolutional recurrent neural networks for text classification," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2019, pp. 1–6.
[38] H. A. Almuzaini and A. M. Azmi, "Impact of stemming and word embedding on deep learning-based Arabic text categorization," *IEEE Access*, vol. 8, pp. 127913–127928, 2020.
[39] C. Adewale, J. P. Reganold, S. Higgins, R. D. Evans, and L. Carpenter-Boggs, "Agricultural carbon footprint is farm specific: Case study of two organic farms," *J. Cleaner Prod.*, vol. 229, pp. 795–805, Aug. 2019.
[40] A. Biørn-Hansen, W. Barendregt, and D. Andersson, "Introducing financial data and groups in a carbon calculator: Issues with trust and opportunities for social interaction," in *Proc. 7th Int. Conf. (ICT)*, Jun. 2020, pp. 11–17.
[41] W. Barendregt, A. Biørn-Hansen, and D. Andersson, "Users experiences with the use of transaction data to estimate consumption-based emissions in a carbon calculator," *Sustainability*, vol. 12, no. 18, p. 7777, Sep. 2020.
[42] F. Emmert-Streib, O. Yli-Harja, and M. Dehmer, "Explainable artificial intelligence and machine learning: A reality rooted perspective," *WIREs Data Mining Knowl. Discovery*, vol. 10, no. 6, p. e1368, Nov. 2020.
[43] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where," *IEEE Trans. Ind. Informat.*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022.
[44] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *J. Artif. Intell. Res.*, vol. 70, pp. 245–317, Jan. 2021.
[45] G. Montavon, J. Kauffmann, W. Samek, and K.-R. Müller, "Explaining the predictions of unsupervised learning models," in *Proc. Int. Workshop Extending Explainable AI Beyond Deep Models Classifiers*. Vienna, Austria: Springer, 2022, pp. 117–138.
[46] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," *Knowl.-Based Syst.*, vol. 214, Feb. 2021, Art. no. 106685.
[47] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.
[48] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.
[49] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, 2015.
[50] D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," *J. Roy. Stat. Society, Ser. B*, vol. 82, no. 4, pp. 1059–1086, Sep. 2020.
[51] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.
[52] G. Plumb, D. Molitor, and A. Talwalkar, "Model agnostic supervised local explanations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2515–2524.
[53] S. Kiefer, "CaSE: Explaining text classifications by fusion of local surrogate explanation models with contextual and semantic knowledge," *Inf. Fusion*, vol. 77, pp. 184–195, Jan. 2022.
[54] O. Sagi and L. Rokach, "Explainable decision forest: Transforming a decision forest into an interpretable tree," *Inf. Fusion*, vol. 61, pp. 124–138, Sep. 2020.
[55] C. Cousins and M. Riondato, "CaDET: Interpretable parametric conditional density estimation with decision trees and forests," *Mach. Learn.*, vol. 108, nos. 8–9, pp. 1613–1634, Sep. 2019.
[56] Ö. Sürer, D. W. Apley, and E. C. Malthouse, "Coefficient tree regression: Fast, accurate and interpretable predictive modeling," *Mach. Learn.*, pp. 1–37, Nov. 2021.
[57] M. P. Neto and F. V. Paulovich, "Explainable matrix–visualization for global and local interpretability of random forest classification ensembles," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 2, pp. 1427–1437, Feb. 2021.
[58] S. Tandra and A. Manashty, "Probabilistic feature selection for interpretable random forest model," in *Advances in Information and Communication* (Advances in Intelligent Systems and Computing), vol. 1364. Cham, Switzerland: Springer, 2021, pp. 707–718.
[59] P. Ponte and R. G. Melko, "Kernel methods for interpretable machine learning of order parameters," *Phys. Rev. B, Condens. Matter*, vol. 96, no 20, 2017, Art. no. 205146.
[60] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.
[61] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.
[62] J. Rožanec, E. Trajkova, I. Novalija, P. Zajec, K. Kenda, B. Fortuna, and D. Mladenic, "Enriching artificial intelligence explanations with knowledge fragments," *Future Internet*, vol. 14, no. 5, p. 134, Apr. 2022.
[63] R. B. Pereira, A. Plastino, B. Zadrozny, and L. H. C. Merschmann, "Categorizing feature selection methods for multi-label classification," *Artif. Intell. Rev.*, vol. 49, no. 1, pp. 57–78, 2016.
[64] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognit.*, vol. 118, Oct. 2021, Art. no. 107965.
[65] A. Jain, A. Jain, N. Chauhan, V. Singh, and N. Thakur, "Information retrieval using cosine and Jaccard similarity measures in vector space model," *Int. J. Comput. Appl.*, vol. 164, no. 6, pp. 28–30, Apr. 2017.
[66] R. Singh and S. Singh, "Text similarity measures in news articles by vector space model using NLP," *J. Inst. Eng., Ser. B*, vol. 102, no. 2, pp. 329–338, Apr. 2021.
[67] F. De Arriba-Pérez, S. García-Méndez, F. J. González-Castaño, and E. Costa-Montenegro, "Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with natural language processing capabilities," *J. Ambient Intell. Humanized Comput.*, pp. 1–16, Apr. 2022.
[68] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised machine learning: A brief primer," *Behav. Therapy*, vol. 51, no. 5, pp. 675–687, Sep. 2020.
[69] J. Wanner, L.-V. Herm, K. Heinrich, and C. Janiesch, "Stop ordering machine learning algorithms by their explainability! An empirical investigation of the tradeoff between performance and explainability," in *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2021, pp. 245–258.

**JAIME GONZÁLEZ-GONZÁLEZ** received the B.S. degree in telecommunication technologies engineering and the M.S. degree in telecommunication engineering from the University of Vigo, Spain, in 2020 and 2022, respectively, where he is currently pursuing the Ph.D. degree with the Information Technologies Group. He is also a Researcher with the Information Technologies Group, University of Vigo. His research interests include the development of machine learning solutions for automatic text classification and augmentative and alternative communication.

**SILVIA GARCÍA-MÉNDEZ** received the Ph.D. degree in information and communication technologies from the University of Vigo, in 2021. Since 2015, she has been working as a Researcher with the Information Technologies Group, University of Vigo. She is currently collaborating with foreign research centers as part of her postdoctoral stage. Her research interests include natural language processing techniques and machine learning algorithms.

**FRANCISCO DE ARRIBA-PÉREZ** received the B.S. degree in telecommunication technologies engineering, the M.S. degree in telecommunication engineering, and the Ph.D. degree from the University of Vigo, Spain, in 2013, 2014, and 2019, respectively. He is currently a Researcher with the Information Technologies Group, University of Vigo. His research interests include the development of machine learning solutions for different domains, such as finance and health.

**FRANCISCO J. GONZÁLEZ-CASTAÑO** received the B.S. degree from the University of Santiago de Compostela, Spain, in 1990, and the Ph.D. degree from the University of Vigo, Spain, in 1998. He is currently a Full Professor at the University of Vigo, where he leads the Information Technologies Group. He has authored over 100 papers in international journals in the fields of telecommunications and computer science and has participated in several relevant national and international projects. He holds three U.S. patents.

**ÓSCAR BARBA-SEARA** received the B.S. degree in computer science and the M.S. degree in superior computer engineering from the University of Vigo and the M.S. degree in e-commerce from the Pontifical University of Salamanca. He is currently pursuing the Ph.D. degree with the University of Vigo. He has worked as the CTO or the Technical Manager in several IT projects with more than 14 years of experience in the public and private sectors and in IT integrations with international corporations, such as Mapfre, Caser, Abanca, Evo Banco, Vodafone, and AON. He is currently the CTO of Two Initiatives in the fintech sector involving machine learning research for the analysis of financial and market-related texts.