Natural Language Engineering (2024), **30**, pp. 1255–1276
doi:10.1017/S135132492300030X

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

How you describe procurement calls matters: Predicting outcome of public procurement using call descriptions

Utku Umur Acikalin[1], Mustafa Kaan Gorgun[1], Mucahid Kutlu[1] and Bedri Kamil Onur Tas[2]

[1]TOBB University of Economics and Technology, Ankara, Turkey and [2]Department of Economics and Finance, Sultan Qaboos University, Muscat, Oman

Corresponding author: Utku Umur Acikalin; Email: u.acikalin@etu.edu.tr

(Received 17 April 2022; revised 8 March 2023; accepted 27 May 2023; first published online 10 August 2023)

Abstract

A competitive and cost-effective public procurement (PP) process is essential for the effective use of public resources. In this work, we explore whether descriptions of procurement calls can be used to predict their outcomes. In particular, we focus on predicting four well-known economic metrics: (i) the number of offers, (ii) whether only a single offer is received, (iii) whether a foreign firm is awarded the contract, and (iv) whether the contract price exceeds the expected price. We extract the European Union's multilingual PP notices, covering 22 different languages. We investigate fine-tuning multilingual transformer models and propose two approaches: (1) multilayer perceptron (MLP) models with transformer embeddings for each business sector in which the training data are filtered based on the procurement category and (2) a k-nearest neighbor (KNN)-based approach fine-tuned using triplet networks. The fine-tuned MBERT model outperforms all other models in predicting calls with a single offer and foreign contract awards, whereas our MLP-based filtering approach yields state-of-the-art results in predicting contracts in which the contract price exceeds the expected price. Furthermore, our KNN-based approach outperforms all the baselines in all tasks and our other proposed models in predicting the number of offers. Moreover, we investigate cross-lingual and multilingual training for our tasks and observe that multilingual training improves prediction accuracy in all our tasks. Overall, our experiments suggest that notice descriptions play an important role in the outcomes of PP calls.

Keywords: Multilinguality; Text classification

1. Introduction

The World Bank estimates that 12% of global gross domestic product (GDP) is spent on public procurement (PP) (Bosio and Djankov 2020). In the European Union (EU), over 250,000 public authorities outlay more than € 2 trillion every year, accounting for 14% of the EU GDP.[a] Lower competition in PP results in to higher contract prices and significantly decreases cost-effectiveness. The World Bank states that "poor governance of PP turns public investments into major political and economic liabilities, hinders development goals, and results in added costs and the waste of public funds."[b]

To ensure the effective use of public funds, the European Commission directives promote the PP market to be open, competitive, and cost-effective. To foster transparency and increase firm participation, EU public officials are required to publish contract notices openly. Tender Electronic Daily (TED)[c] publishes more than 746K procurement award notices in a year.

[a]https://simap.ted.europa.eu/en_GB/web/simap/european-public-procurement

[b]https://www.worldbank.org/en/news/feature/2016/08/17/reforming-and-modernizing-public-procurement-in-mena

[c]https://ted.europa.eu/TED/

Approximately 2400 PP notices are published every weekday. Firms can browse and search for available PP calls of 27 EU member states. While PP notices are released regularly and are publicly available, their information content differs extensively. *We hypothesize that descriptions of PP notices affect the outcome of procurement processes.* For instance, poorly explained descriptions with limited details are likely to cause additional challenges for firms seeking business opportunities in foreign countries, decreasing competition and, thereby, increasing procurement costs. Accordingly, authorities can manipulate the integrity of the bidding process by limiting the information contents of notices.

Considering the massive number of calls, PP regulators cannot manually assess whether a call for procurement is adequate for a fair, competitive, and cost-effective PP process. Therefore, automatic systems are required to assist PP regulators and the effective use of public funds.

In this paper, we investigate how to predict competition and cost-effectiveness in EU PP using call descriptions. We adopt well-known economic metrics to measure competition and cost-effectiveness. Specifically, we employ the number of offers (Onur, Özcan, and Taş 2012a), whether a single offer is received (Fazekas and Kocsis 2020), and whether a foreign firm is awarded a contract (Kutlina-Dimitrova and Lakatos 2016), as measures of competition. We gauge cost-effectiveness by assessing if the contract price exceeds the estimated cost (Taş *et al.* 2019). As EU PP call descriptions cover many different languages, we investigate fine-tuning multilingual transformer models for each prediction task. In addition, we propose two different methods: (1) using a separate multilayer perceptron (MLP) model with transformer embeddings for each business sector in which training data are filtered based on business sector and (2) a k-nearest neighbor (KNN)-based approach fine-tuned using triplet networks (TNs) (Hoffer and Ailon 2015).

In our extensive experiments, we show that the fine-tuned MBERT model outperforms all other models in predicting PP calls that receive a single offer and calls in which a foreign firm is awarded the contract. In addition, our MLP and KNN-based models outperform all other models in predicting cost-effectiveness and the number of offers, respectively. Furthermore, our KNN-based model outperforms all baselines in all tasks. Moreover, we investigate the impact of multilingual training and cross-lingual training using two different multilingual language models. In particular, we seek answers to the following research questions.

RQ-1: Do descriptions of PP calls affect the number of offers, whether a foreign firm can be awarded, and the contract price? We rely on contract descriptions in our proposed models and do not use any information on the macroeconomic status of the countries or business sectors. We show that our proposed methods outperform baseline methods that use macroeconomic information, suggesting that descriptions of PP calls affect bidding process outcomes. Therefore, we believe that text generation models can be employed to improve economic outcomes of PP calls. We leave this exciting research problem as future work.

RQ-2: Which multilingual model is effective for these prediction tasks? We use XLMR and MBERT separately to compare their performance. The fine-tuned MBERT outperforms the fine-tuned XLMR in all four tasks. However, the results for our MLP-based methods are mixed such that the best-performing model changes across tasks.

RQ-3: Is multilingual training effective in analyzing EU PP calls? We show that the performance of a single MBERT model fine-tuned with descriptions spanning 22 different languages is higher than the overall performance of 22 MBERT models, each fine-tuned with a different language. This suggests that using data in different languages improves the performance of our models.

RQ-4: Which source-target language pairs are the most effective for cross-lingual training for our prediction tasks? For each language pair in our dataset, we fine-tune MBERT model with descriptions written in one language and test the model with descriptions written in the same or another language. Generally, as expected, using the same language for both fine-tuning and

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

testing yields the highest performance for all languages. However, in some cases we also observe that cross-lingual training yields higher performance than training with descriptions in the same language. This might be due to the varying label distribution and number of PP calls for each language. In addition, our results suggest that the performance of cross-lingual training is not only affected by linguistic similarities of languages; economic similarities of countries also play an important role in our prediction tasks.

Our contribution in this work is fourfold. (1) To the best of our knowledge, this is the first study using EU's extremely multilingual PP dataset to predict the outcomes of procurement calls using their descriptions. (2) Our models outperform the baseline models. In addition, we investigate the impact of multilingual and cross-lingual training in 22 different languages. (3) Our results show that the information content of PP calls is essential to predicting the level of competition and cost-effectiveness in PP, suggesting that regulators should monitor and regulate the information content of notices. (4) We release our code for extracting data from the TED dataset and conducting our experiments to maintain the reproducibility of our results.[d]

The rest of this paper is organized as follows. Section 2 discusses the related literature. We explain the PP dataset in Section 3 and define the prediction tasks we focus on in Section 4. In Section 5, we describe the proposed methods. We present our experimental results in Section 6 and conclude the paper in Section 7.

2. Related work

In this section, we present studies related to our work from different aspects. In particular, we discuss (i) studies about economic analysis of PP calls (Section 2.1), (ii) prediction models for various tasks related to PP (Section 2.2), (iii) studies using the TED dataset for other NLP tasks (Section 2.3), and (iv) multilingual document classification methods (Section 2.4).

2.1 Economic analysis of public procurement

The size and economic importance of PP have attracted the interest of many researchers, particularly in the field of economics. For instance, Iimi (2006) examines procurement auctions for Japanese official development assistance projects and concludes that a 1% increase in the number of offers decreases the price of contracts by 0.2%. Similarly, Onur, Ozcan & Taş (2012b) show that procurement prices in Turkish PP auctions are negatively related to the number of offers. Taş (2022) provides empirical evidence on the significance of PP regulation quality on EU PP outcomes. Fazekas and Blum (2021) state that governments can achieve 5–10% price savings by improving PP practices. Gorgun *et al.* (2022) examine the impact of descriptions on PP notices. They find that competition is considerably higher when descriptions contain more information. Furthermore, they report that EU governments could save up to € 80 billion if PP descriptions were to contain detailed information. In our work, we focus on developing prediction models that may be beneficial for authorities and economists working on PP, rather than analyzing the economic impact of PP calls.

2.2 Prediction tasks for public procurement calls

Several researchers have worked on various prediction tasks about PP such as cost increase with respect to low bids (Williams and Gong 2014), uncertainty risks of projects in bidding processes (Lee and Yi 2017; Jang *et al.* 2019), contracts likely to result in corruption investigations (Gallego, Rivero, and Martínez 2021), the number of bidders (Demiray, Ozbayoglu, and Taş 2015;

[d] https://github.com/utkuumur/HowYouDescribeProcurementCallsMatters

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

Rabuzin and Modrusan 2019), and award price (García Rodríguez *et al.* 2019). We now focus on studies on prediction tasks similar to ours.

Demiray *et al.* (2015) predict the number of offers in Turkish PP auctions using support vector machine (SVM) with nine features such as estimated cost, whether it involves goods or services, and whether it is open to foreign firms. Rabuzin and Modrusan (2019) focus on Croatian PP calls and predict whether a PP call will receive a single offer or more, considering that a single offer might be an indicator of corruption. They use Naive Bayes, logistic regression (LR), and SVM with term frequency-inverse document frequency (TF-IDF) vectors. In order to predict the number of offers in the TED dataset, Gorgun, Kutlu, and Taş (2020) investigate several features including named entities, description lengths, TF-IDF vectors, business sector, country, and language codes. They report that KNN with *n*-gram vectors yields the highest performance. In their work, they use only 59,397 PP notices and translate all non-English descriptions to English using automatic translation tools. In our work, we use a much larger dataset and utilize multilingual language models, instead of machine translation.

García Rodríguez *et al.* (2019) predict the award price of Spanish PP calls using random forest (RF) regressor with 14 features such as date, duration, business sector, province, and municipality. Lima *et al.* (2020) extract the texts of 15 million Brazilian PP calls and annotate 1907 of them as "risky" in terms of corruption and fraud. They use a bidirectional long short-term memory model to detect risky calls. In our work, we use a dataset covering multiple countries and languages and work on four different prediction tasks.

Imhof and Wallimann (2021) propose a supervised method that uses the distribution of bids to detect collusive groups of firms in procurement auctions. Aldana *et al.* (2022) propose an ensemble of RF to detect corrupt contracts in Mexico's PP. They use a wide range of features such as descriptors of the buyer/supplier relation, type of contract, and others.

Modrusan *et al.* (2020) focus on determining which parts of the PP calls are important to predict suspicious procurement procedures. They determine 11 terms such as "minimum," "maximum," "requirement," "technical and expert" based on their interviews with experts in the field of the PP process. Subsequently, they show that using paragraphs that contain one of the selected terms and the following 10 paragraphs increases performance in predicting suspicious procurement procedures.

In our study, we work on four prediction tasks including the number of offers, whether a contract will receive a single offer, whether a foreign firm will be awarded, and effectiveness of contracts. In all our prediction tasks, we use only descriptions of calls to investigate their impact on the outcome of the procurement processes.

2.3 Studies using TED dataset

In our work, we use TED dataset to develop and test our models. Because it is a large, publicly available multilingual dataset, several researchers have utilized it as a resource for various NLP tasks. For instance, Ahmia *et al.* (2018) prepare a corpus of aligned sentences extracted from the TED dataset, which may be useful for machine translation studies. Simperl *et al.* (2018) focus on constructing a knowledge graph from TED data. They indicate multilinguality as a challenge in linking and comparing documents for analysis. They propose mapping word representations into a common semantic vector space which enables the comparison of document similarities across languages. Similarly, Zhang *et al.* (2023) focus on constructing a structured procurement contract database using TED dataset to ease the tendering process in healthcare. The authors mention the incompleteness of documents and differences in XML format as challenges they encountered. Mehrbod and Grilo (2018) develop a named entity recognizer to empower semantic search in the TED dataset.

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

2.4 Multilingual document classification

As we propose classification and regression models for multilingual documents, now we discuss studies on multilingual document classification. Pappas and Popescu-Belis (2017) propose multilingual hierarchical attention networks for document classification and evaluate their approach on a dataset covering news articles written in eight different languages. Van der Heijden *et al.* (2021) propose a meta-learning approach for few-shot multilingual and cross-lingual adaptation.

A number of studies assess the performance of MBERT in various conditions. For instance, Wu and Dredze (2019) investigate the performance of MBERT in a cross-lingual zero-shot transfer setting for various NLP tasks including document classification, part-of-speech tagging, named entity recognition, natural language inference, and dependency parsing. They show that MBERT yields high performance in all tasks. Pires *et al.* (2019) show that MBERT's performance decreases for the languages with different sentence structures from English. Karthikeyan *et al.* (2019) study the cross-lingual ability of MBERT, analyzing its architecture, linguistic properties of languages, and the learning objectives. They find that the lexical overlap between languages plays a negligible role in the cross-lingual performance of MBERT.

Many researchers also investigate whether we should use multilingual models trained with large datasets covering various languages or monolingual models trained with relatively smaller datasets focusing on a single language. For instance, Catelli *et al.* (2022) explore the effects of cross-lingual learning for Italian sentiment analysis for hotel reviews. They show that MBERT outperforms the Italian-specific BERT model when trained with a mixed dataset composed of English-Italian reviews while Italian BERT outperforms MBERT when trained with only Italian reviews.

Piscitelli *et al.* (2021) develop a CNN architecture for tweet classification using multilingual MUSE model. They evaluate its performance for zero-shot learning in English, Spanish, and Italian languages and show that multilingual training increases the performance for the languages with limited training data whereas it decreases the performance for English.

Aluru *et al.* (2020) compare the performances of several approaches for multilingual hate speech detection for different data sizes and monolingual and multilingual settings. For the monolingual setting, they report that LASER with LR model outperforms others when the training data are limited. On the other hand, BERT outperforms others when there are more training data. They also compare the performances of LASER and MBERT in a cross-lingual setting and report that LASER achieves the highest score in six languages, whereas MBERT outperforms LASER in three languages. However, when these models are further trained with data in the target language, MBERT performs LASER in eight languages.

To our knowledge, our work is the first study working on these four prediction tasks on a dataset covering 22 European languages. We compare the performance of MBERT and XLMR for our prediction tasks. In addition, we compare cross-lingual performance of our models and explore which language pairs are suitable for cross-lingual training. Furthermore, we investigate whether we should use multilingual or monolingual models for our prediction tasks.

3. Dataset

We use the TED dataset, which has been published by the Publications Office of the EU since 2006 through the European Union Open Data Portal website.[e] The published dataset contains many different types of information on bidding processes and their award results, including date, contract type, procuring authority, common procurement vocabulary (CPV) codes (i.e., nine-digit codes indicating the specific line of business), contract descriptions, lot numbers (if procurement is divided into lots), number of offers, awarded and estimated prices, and awarded firm. Table 1 shows samples of descriptions and other information we use in our work.

[e] https://data.europa.eu/euodp

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

Table 1. Sample public procurement calls in English, German, and Greek

CPV	Description	Authority	Estimated price	Actual price	Awarded firm	Num. offers
45,232,470	Providing secure site or sites for a registered social housing landlord in North Cheshire to transport non-hazardous waste for recycling and disposal of the waste. Provision of various sized skips at individual sites. Provision of sealed asbestos skips to receive Chrysotile asbestos. Waste removal and recycling services for approximately 1 500 tonnes of waste per annum	Weaver Vale Housing Trust (UK)	523,396.84	523,396.84	Nick Brookes Group (UK)	1
3,220,000	Rahmenvertrag zur Lieferung von Frischobst und Frischgemüse zur Direktlieferung an die o.a. Truppenküchen im Einsatzkontingent ISAF (Afghanistan) für den Zeitraum von sieben Monaten (Mazar-E-Sharif für vier Monate). Die Lieferung erfolgt grundsätzlich wöchentlich an den jeweilig o.a. Empfänger sowie auf Abruf bei Bedarf. Wird detailliert in den Verdingungsunterlagen dargestellt Umfang siehe auch Ziffer II 1 2	Verpflegungsamt der Bundeswehr (DE)	1,500,000	1,663,669	Supreme Foodservice GmbH (CH)	6
31,650,000	Προμήϑεια συνϑετιϰών μονωτήρων αναρτήσεως ϰαι στήριξης. — 50 000 τεμάχια συνϑετιϰών μονωτήρων αναρτήσεως, ϰαι — 150 000 τεμάχια συνϑετιϰών μονωτήρων στήριξης	Δημόσια Επιχείρηση Ηλεϰτρισμού ΑΕ (ΓΡ)	1,312,500	978,243.01	TYCO Electronics Raychem GMBH (DE)	11

PP contracts divided into lots have the main description about its general definition and a specific description for each lot division. In this work, we use PP contracts published between 2011 and 2018. In addition, we filter out PP contracts divided into lots because using lots would result in oversampling of main descriptions of contracts divided into lots. Furthermore, we encountered many problems in the extraction of lot descriptions due to changing and complex XML format. Thus, we leave the analysis of contracts with lots as future work.

3.1 Topics

PP notices are categorized hierarchically such that notices are first divided into divisions, then groups, then classes, then categories, and finally subcategories. The dataset we use has 46 divisions, 322 groups, 1304 classes, 3403 categories, and 7841 subcategories in total, covering a wide range of topics, such as the delivery of different food products for armed forces in Berlin, Germany (PP notice ID: 2012385044), and the purchase of construction materials for the University Hospital in Vilnius, Lithuania (PP notice ID: 201286594).

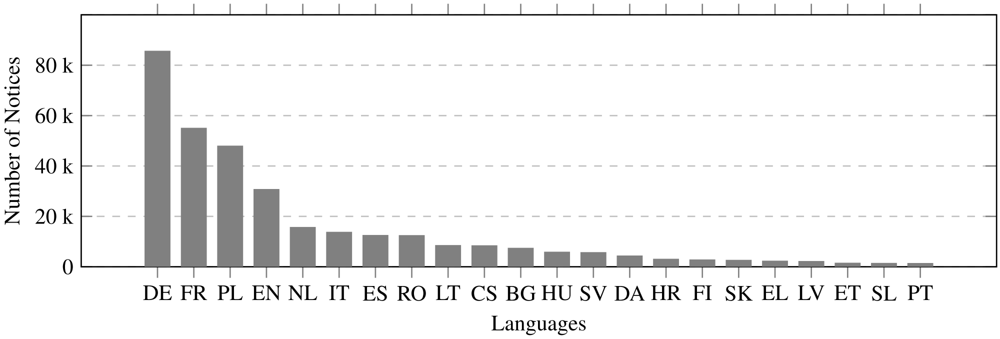


Figure 1. The number of all notices for each language in our dataset.

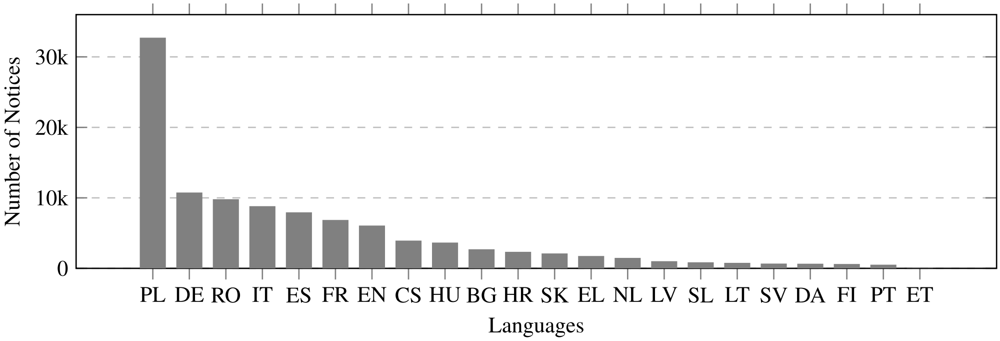


Figure 2. Number of notices that provide the estimated price information for each language in our dataset.

3.2 Data extraction

The TED dataset is publicly available in complex but structured XML files. However, there are still many challenges in extracting the data due to the multilingual nature of the data, changes in the format of XML files throughout the years, and human errors. For instance, we observed that there is no common XML tag name for descriptions. While the majority of the descriptions are provided in three description-specific XML tags (short_descr, short_description_contract, and short_contract_description), a portion of the contracts provides description under general XML tags which might also contain other types of information. Therefore, regarding the general XML tags, we used the ones which contain the word “description” or its translation in the respective document’s original language (e.g., “Beskrivelse” for Danish and “Opis” for Polish) in the XML tag description. Furthermore, in contracts with multiple original languages, we select English descriptions if available. Otherwise, we randomly choose one of the respective original languages. In total, we created a dataset that consists of 333.832 notices. However, the estimated price value was provided in only 106.122 notices.

3.3 Language

Contract documents are written in 24 different European languages: Bulgarian (BG), Czech (CS), Danish (DA), German (DE), Greek (EL), English (EN), Spanish (ES), Estonian (ET), Finnish (FI), French (FR), Irish (GA), Croatian (HR), Hungarian (HU), Italian (IT), Lithuanian (LT), Latvian (LV), Maltese (MT), Dutch (NL), Polish (PL), Portuguese (PT), Romanian (RO), Slovak (SK), Slovenian (SL), and Swedish (SV). Figures 1 and 2 show the distribution of contracts’

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

original language in our dataset for all notices we extracted and for notices in which the estimated price information is provided, respectively. The distribution of languages is unbalanced, such that notices in German, French, Polish, English, and Dutch (i.e., the five most frequent languages) constitute 70.58% of all notices. Considering only notices with estimated price information, the distribution of languages is again unbalanced, such that notices written in Polish, German, Romanian, Italian, and Spanish (i.e., the five most frequent languages) constitute 66% of the data. In our work, we do not use Maltese and Irish due to the excessively limited number of calls written in these languages. Overall, our dataset covers descriptions written in 22 different languages with various alphabets (e.g., Latin, Cyrillic, and Greek).

4. Problem definition

The main goal of this paper is to develop a model to detect whether a given PP description is suitable for reaching EU goals, that is, conducting a cost-effective and competitive PP process. However, labeling descriptions whether they are suitable or not is a challenging task, even for the experts specialized in PP. In addition, it significantly decreases data size due to its massive cost. Therefore, instead of directly predicting how suitable a description is, we adopt metrics used by economists to assess the quality of a procuring process. In particular, we predict (i) the number of offers, (ii) whether it will receive a single offer, (iii) whether a foreign firm will be awarded, and (iv) whether the awarded price exceeds the estimated price using the descriptions of PP notices. Now, we explain each prediction task in detail.

4.1 Predicting number of offers ($P_{NumOffers}$)

EU authorities require a competitive open market to effectively utilize public resources. The procurement price is likely to decrease if the awarding process is competitive. Therefore, the number of offers submitted for PP notices is an effective metric to assess the level of competition (Onur, Özcan, and Taş 2012a). In this task, we predict the number of offers for a given PP notice using its description.

4.2 Predicting whether a contract will receive a single offer ($P_{SingleOffer}$)

PP calls with a single offer are used as an indicator for potential corruption (Rabuzin and Modrusan 2019). Wachs, Fazekas, and Kertész (2021) investigate corruption risk in EU PP using single-bidder contracts since there is no competition. Charron *et al.* (2017) show that single-bidding is a significant red flag for corruption in PP. In contrast to the $\mathrm{P}_{\mathrm{NumOffers}}$ task, we focus on finding PP notices that received exactly one offer. In this task, we consider calls with a single offer as positive class and apply a binary classification of PP calls using their descriptions.

4.3 Predicting whether a foreign firm will be awarded ($P_{FFAward}$)

Firms in any EU member country can apply for PP calls of authorities in any member country. However, it is still challenging for a firm to be awarded by a foreign authority due to being less familiar with the respective country than local firms. Thus, if a foreign firm can win a contract, it indicates market transparency and highlights the competitiveness of the bidding process (Kutlina-Dimitrova and Lakatos 2016). In this task, we consider PP calls in which a foreign firm is awarded as positive class and predict whether a foreign firm wins a contract for a given PP notice using its description.

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

4.4 Predicting effectiveness of contracts ($P_{Effectiveness}$)

The EU authorities calculate the estimated prices for PP calls by conducting extensive market analyses. The calculation of the estimated price is an essential part of the EU PP process. On average, actual contract prices are 90–95% of estimated prices, according to studies that analyze PP markets in Japan (Ishii 2009), Italy (Coviello, Guglielmo, and Spagnolo 2018), Turkiye (Onur, Ozcan, and Taş 2012b), and the EU (Fazekas and Tóth 2018). Accordingly, awarded prices higher than estimated prices suggest an ineffective bidding process (Taş *et al.* 2019). These contracts have poor value-for-money performance. In this task, we predict whether a given PP notice will cause an ineffective contract (i.e., the awarded price is higher than the estimated price), using its description. We consider ineffective contracts as the positive class.

5. Proposed methods

We propose three different approaches for our prediction tasks. In our first approach, we investigate fine-tuning multilingual transformer models (Section 5.1). In our second approach, we propose training machine learning algorithms with multilingual vectors on data filtered based on CPV codes (Section 5.2). In our third approach, we propose a fine-tuning procedure using TNs (Section 5.3).

5.1 Fine-tuning multilingual language models

The dataset we use covers 22 different languages and requires multilingual language models for our prediction tasks. Therefore, we fine-tune a transformer model for each prediction task. We investigate MBERT (Devlin *et al.* 2019) and XLMR (Conneau *et al.* 2020) separately to compare their performances. We only use extracted description parts of the PP notices for fine-tuning. We use truncation and padding for the overflowing and the underflowing cases, respectively. In the fine-tuning process, an additional linear layer is appended to the transformer model with random weights. The additional layer is task-specific; therefore, a classification layer is added for $P_{SingleOffer}$, $P_{FFAward}$, and $P_{Effectiveness}$ tasks, and a regression layer is added for $P_{NumOffers}$. Subsequently, these models are trained with cross-entropy loss and mean square error loss for the classification and regression tasks, respectively.

5.2 Filtering the training data based on CPV codes

As mentioned in Section 3, the TED dataset covers a wide range of procurement services and products in different countries. Therefore, the values we predict are also likely to be affected by the sector of PP calls and the country of the procuring authorities. For instance, the number of offers for the purchase of construction materials might be significantly smaller than the number of offers for the delivery of food products. In addition, if a country has no local firm eligible to provide a procured item (e.g., a particular construction material), then a foreign company will be awarded. In order to consider the differences across sectors, we divide the training dataset into multiple subsets, where each subset has PP notices for a separate business division (i.e., the first two digits of CPV codes). Subsequently, we train a separate model for each business division and language using the respective data. Additionally, economic conditions play a significant role in PP outcomes. Gugler, Weichselbaumer, and Zulehner (2015) show that competition in PP increases during economic crises. The contract prices in the Austrian PP decreased by 3.3% during the financial crisis of 2009. We use sentence embeddings returned by pre-trained multilingual language models to represent the notice descriptions. We investigated various machine learning algorithms such as KNN, RF, MLP, and logistic/linear regression.

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

5.3 Fine-tuning using Triplet Networks

We apply TNs (Hoffer and Ailon 2015) for our prediction tasks. Our goal is to make the embeddings of PP notices that have the same label *closer* (i.e., lower cosine distance) compared to an arbitrary PP notice. In particular, in TNs, the input consists of three examples: anchor a, positive p, and negative n. The loss function is as follows:

$$\max(\text{CosineDistance}(v_a, v_p) - \text{CosineDistance}(v_a, v_n) + \epsilon, 0) \qquad (1)$$

where ϵ is the margin that ensures that v_p is at least ϵ closer to v_a than v_n and v_a, v_p, and v_n are the embeddings for a, p, and n, respectively.

We select triplets for each task using the training dataset as follows. For each anchor PP a, we select a positive PP p such that (i) a and p are written in the same language, (ii) they share the same CPV code, and (iii) their labels are equal. Selecting negatives is more difficult because if the selected negative PP is not related to the positive PP, the model can easily distinguish it from the positive PP, which would make this triplet fruitless for training. Therefore, instead of selecting negatives uniformly at random, we select three different negatives for each pair *(a,p)* based on their CPV codes. In particular, we randomly select (i) one negative sample among those with the same CPV code, (ii) another negative sample among those sharing the first five digits of the CPV, and (iii) another one with any CPV code. All selected negatives are written in the same language as a.

Because the label distributions of tasks and the sizes of the dataset differ from each other, we select a similar number of triplets for each task by adjusting the number of positive examples selected for each anchor notice a. We train a single model for both $\text{P}_{\text{NumOffers}}$ and $\text{P}_{\text{SingleOffer}}$ tasks based on the number of offers, because using the actual number of offers is more informative than using binary labels in this kind of training.

The number of triplets for the tasks $\text{P}_{\text{NumOffers}}$, $\text{P}_{\text{SingleOffer}}$, $\text{P}_{\text{FFAward}}$, and $\text{P}_{\text{Effectiveness}}$ is 1,268,827, 1,268,827, 1,048,603, and 1,127,083, respectively. We first fine-tune MBERT model using the loss function described above. Subsequently, we use the fine-tuned model to obtain an embedding for each description. Then, these embeddings are fed into a KNN model that returns the median label value of the k-closest neighbor for each task. We refer to this model as $\text{TN}_{\text{KNN-MBERT}}$. As in our filtering method in Section 5.2, we also apply filtering methods to our KNN model. In particular, we filter based on CPV codes ($\text{TN}_{\text{KNN-MBERT-CPV}}$) and language ($\text{TN}_{\text{KNN-MBERT-LANG}}$). In the $\text{TN}_{\text{KNN-MBERT-CPV}}$ and $\text{TN}_{\text{KNN-MBERT-LANG}}$ methods, we ensure that all the k neighbors belong to the same CPV code and are written in the same language, respectively.

6. Experiments

In this section, we first explain our experimental setup including the train and test datasets, evaluation metrics, baseline methods, and implementation details (Section 6.1). Next, we present our experimental results, including a comparison against baselines and the impact of multilingual and cross-lingual training (Section 6.2).

6.1 Experimental setup

Train and test datasets. The details of our dataset are provided in Table 2. We observe that a portion of PP documents does not contain information about the predicted outcome, yielding different dataset sizes for each task. For all four tasks, we use 80% of the datasets for training and the rest for testing. The label distributions for both the $\text{P}_{\text{FFAward}}$ and $\text{P}_{\text{Effectiveness}}$ tasks are highly imbalanced. Only 14.5% of notices are found to be ineffective, and in 3.3% of calls a foreign firm

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

Table 2. Statistics of our dataset

	Training set	Test set
Years covered	2011–2018	2011–2018
The number of notices for $\mathrm{P}_{\mathrm{NumOffers}}$	267064	66768
The number of notices for $\mathrm{P}_{\mathrm{SingleOffer}}$	267064	66768
The number of notices for $\mathrm{P}_{\mathrm{FFAward}}$	267064	66768
The number of notices for $\mathrm{P}_{\mathrm{Effectiveness}}$	84896	21226
The ratio of single offer PP notices	21.0%	21.1%
The ratio of ineffective PP notices	14.5	14.5
The ratio of contracts awarded foreign firm	3.4 %	3.5%
The minimum number of offers	1	1
The maximum number of offers	18	18
The average number of offers	3.911	3.939
The median number of offers	3	3
The standard deviation in number of offers	3.073	3.106

is awarded. The label distribution for the $\mathrm{P}_{\mathrm{SingleOffer}}$ tasks is slightly more balanced compared to the other classification tasks. 21.0% of the notices received a single offer. For the regression task, that is, $\mathrm{P}_{\mathrm{NumOffers}}$, the standard deviation and the average number of offers are 3.079 and 3.918, respectively.

Evaluation metrics. In order to evaluate the performances of models, we report mean absolute error (MAE) for $\mathrm{P}_{\mathrm{NumOffers}}$ task and F_1 score for $\mathrm{P}_{\mathrm{SingleOffer}}$, $\mathrm{P}_{\mathrm{FFAward}}$, and $\mathrm{P}_{\mathrm{Effectiveness}}$ tasks. When calculating the F_1 score, we consider contracts with a single offer, awarded foreign firms, and ineffective contracts (i.e., minority classes) as positive classes.

Baseline models. We compare our models against the five baselines listed below.

- **Median of the Sector** ($\mathrm{Med}_{\mathrm{Sec}}$)**.** As discussed before, the sector might have a significant impact on the results of PP calls. For instance, the number of eligible firms in a sector places an upper bound on the number of offers a PP call can receive. This method, $\mathrm{Med}_{\mathrm{Sec}}$, returns the median value of all contracts in the respective sector. If that sector has no contract, we use the median value of all contracts.
- **Median of the Sector in the Procuring Country** ($\mathrm{Med}_{\mathrm{SecCnt}}$). The economic status of the procuring country may also affect the outcome of procurement calls. Therefore, $\mathrm{Med}_{\mathrm{SecCnt}}$ returns the median value of the contracts in the respective sectors of the procuring country. If no contract in the train set is suitable, we use the median value of all contracts.

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

- **KNN w/uni-grams** ($\mathrm{KNN}_{\mathrm{unigram}}$). Gorgun *et al.* (2020) report that KNN with n-grams is effective in predicting the number of offers on the TED dataset. Therefore, we extract 10K uni-grams based on their TF-IDF scores and use a KNN model.
- **Linear/Logistic Regression w/uni-grams** (LRunigram)**.** Similar to $\mathrm{KNN}_{\mathrm{unigram}}$, we use linear regression for the $\mathrm{P}_{\mathrm{NumOffers}}$ task and LR for the others with 10K uni-grams.
- **Linear/Logistic Regression with Macroeconomic Variables** ($\mathrm{LR}_{\mathrm{MEV}}$)**.** To assess the determinants of the number of offers, Taş *et al.* (2019) use a generalized linear model with procurement-specific variables such as the type of contracting authority, country of authority, the first two digits of CPV codes (i.e., division of the sector), whether the contract was covered by the Government Procurement Agreement, and the type of procedure. We implement their model for our prediction tasks and use linear and LR models for our regression and classification tasks, respectively.

Implementation. We use NLTK[f] for tokenization and stemming. In stemming, we use SnowBallStemmer for the languages it supports. Otherwise, we use the stemmer by selecting the language attribute as "porter." We use scikit library[g] for TF-IDF vectors. We set the minimum document frequency to 0.8. For the baseline methods, we use scikit library with default parameters.

Hyper-parameter tuning. In order to set the hyper-parameters, we further split the initial training set into training and validation sets. We use 25% of the initial training set as the validation set, that is, 60% and 20% of all data are used as the training and validation sets, respectively. Once parameters are tuned, we use the whole training data, that is, 80% of all data, to train models. The hyper-parameters we tune are as follows.

For fine-tuned models, MBERT and XLMR models, we test the learning rates: 1e-5, 2e-5, 3e-5 for each task and choose the best-performing parameters for the remaining experiments. Accordingly, we set the learning rate for fine-tuning models to 1e-5 for $\mathrm{P}_{\mathrm{NumOffers}}$, $\mathrm{P}_{\mathrm{SingleOffer}}$, and $\mathrm{P}_{\mathrm{FFAward}}$ tasks, and 2e-5 for the $\mathrm{P}_{\mathrm{Effectiveness}}$ task. We fine-tune MBERT and XLMR models for 3 epochs with a batch size of 32.

For the KNN models, we examine various values for k in $\{1, 3, 5, \ldots, 51\}$ for each task in the validation set. Accordingly, we set k to 39 for the $\mathrm{P}_{\mathrm{NumOffers}}$ task and 3 for the $\mathrm{P}_{\mathrm{FFAward}}$, $\mathrm{P}_{\mathrm{SingleOffer}}$, and $\mathrm{P}_{\mathrm{Effectiveness}}$ tasks. For the TN-based training procedure, we set the learning rate to 8e-6 and ϵ to 0.1 in all tasks and used a batch size of 16.

6.2 Experimental results

In this section, we first compare our proposed models and the baselines (RQ-1 and RQ-2). Next, we report the impact of multilingual training (RQ-3). Finally, we present our analysis for cross-lingual training (RQ-4).

6.2.1 Impact of filtering, machine learning algorithms, and text representation

In this experiment, we train various machine learning algorithms with different multilingual embeddings with and without filtering of the train data based on CPV codes. In particular, we

[f] https://www.nltk.org/
[g] https://scikit-learn.org/stable/index.html

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

Table 3. Experimental results for various machine learning algorithms using embedding of MBERT and XLMR models. The values in the second column indicate whether filtering based on CPV code is applied or not. The best result for each task is written in bold.

		$\mathrm{P_{NumOffers}}$	$\mathrm{P_{SingleOffer}}$	$\mathrm{P_{FFAward}}$	$\mathrm{P_{Effectiveness}}$
Method	Filtering	MAE	F_1	F_1	F_1
$\mathrm{KNN_{MBERT}}$	✗	2.127	0.397	0.300	0.268
	√	2.092	0.412	0.293	0.267
$\mathrm{KNN_{XLMR}}$	✗	2.129	0.400	0.294	0.277
	√	2.085	0.417	0.288	0.281
$\mathrm{LR_{MBERT}}$	✗	2.024	0.215	0.135	0.109
	√	2.070	0.303	0.115	0.045
$\mathrm{LR_{XLMR}}$	✗	1.999	0.058	0.000	0.000
	√	2.046	0.143	0.000	0.000
$\mathrm{RF_{MBERT}}$	✗	2.110	0.215	0.059	0.105
	√	1.997	0.341	0.114	0.097
$\mathrm{RF_{XLMR}}$	✗	2.082	0.260	0.116	0.123
	√	1.981	0.361	0.142	0.124
$\mathrm{MLP_{MBERT}}$	✗	2.040	0.412	0.285	**0.286**
	√	1.998	**0.430**	**0.305**	0.236
$\mathrm{MLP_{XLMR}}$	✗	1.915	0.374	0.274	0.131
	√	**1.853**	0.397	0.261	0.124

use MBERT and XLMR to represent descriptions and train separate models using RF, MLP, KNN, and logistic/linear regression (LR). Table 3 shows the results.

For the $\mathrm{P_{NumOffers}}$ task, filtering has a slightly positive impact in most of the cases. In addition, for the $\mathrm{P_{SingleOffer}}$ task, filtering improved the performance in all cases. However, in the $\mathrm{P_{Effectiveness}}$ task, filtering has a negative impact in general, and the best-performing model is $\mathrm{MLP_{MBERT}}$ without filtering. Regarding the results for $\mathrm{P_{FFAward}}$ task, the impact of filtering is mixed without any clear positive or negative impact.

However, considering the impact of embeddings, we do not have a clear conclusion. In the $\mathrm{P_{NumOffers}}$ and $\mathrm{P_{SingleOffer}}$ tasks, XLMR yields a higher performance in most of the cases whereas MBERT is preferable for the $\mathrm{P_{FFAward}}$ task. In the $\mathrm{P_{Effectiveness}}$ task, the average performances of MBERT-based methods are higher than the average performances of XLMR-based methods. It is also noteworthy that LR with XLMR yields an F1 score of 0 in the $\mathrm{P_{FFAward}}$ and $\mathrm{P_{Effectiveness}}$ tasks. This may be owing to the highly limited number of positive cases in these tasks. From the perspective of the machine learning algorithms, we observe that MLP yields the best results in all four tasks.

Overall, MLP with BERT embeddings yields the best results in classification tasks (i.e., $\mathrm{P_{SingleOffer}}$, $\mathrm{P_{FFAward}}$, and $\mathrm{P_{Effectiveness}}$), whereas MLP with XLMR embeddings outperforms other models in the $\mathrm{P_{NumOffers}}$ task. In addition, the best-performing models for the three tasks use our filtering approach.

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

Table 4. Experimental results for the fine-tuning procedure based on triplet networks for each task. The best results are written in bold.

Method	$\mathrm{P}_{\mathrm{NumOffers}}$	$\mathrm{P}_{\mathrm{SingleOffer}}$	$\mathrm{P}_{\mathrm{FFAward}}$	$\mathrm{P}_{\mathrm{Effectiveness}}$
	MAE	F_1	F_1	F_1
$\mathrm{TN}_{\text{KNN-MBERT-LANG}}$	**1.789**	**0.428**	**0.352**	**0.246**
$\mathrm{TN}_{\text{KNN-MBERT-CPV}}$	1.820	0.414	0.339	0.201
$\mathrm{TN}_{\text{KNN-MBERT}}$	1.831	0.405	0.334	0.210

Table 5. Experimental results for baselines, fine-tuned MBERT and XLMR models, and the best-performing models based on filtering and TN for each task. The best results are written in bold. The standard deviations across multiple runs of XLMR and MBERT models are shown. We calculated p-values between the best-performing models (shown in bold) and all baselines for each task and found that all p-values are less than 0.01.

Method	$\mathrm{P}_{\mathrm{NumOffers}}$	$\mathrm{P}_{\mathrm{SingleOffer}}$	$\mathrm{P}_{\mathrm{FFAward}}$	$\mathrm{P}_{\mathrm{Effectiveness}}$
	MAE	F_1	F_1	F_1
$\mathrm{Med}_{\mathrm{Sec}}$	1.912	0.235	0.042	0.036
$\mathrm{Med}_{\mathrm{SecCnt}}$	1.846	0.378	0.282	0.141
$\mathrm{KNN}_{\mathrm{unigram}}$	1.975	0.366	0.223	0.224
$\mathrm{LR}_{\mathrm{unigram}}$	1.950	0.308	0.205	0.125
$\mathrm{LR}_{\mathrm{MEV}}$	2.037	0.246	0.190	0.119
Fine-tuned MBERT	1.812 ± 0.003	**0.436** ± 0.001	**0.394** ± **0.002**	0.175 ± 0.006
Fine-tuned XLMR	1.877 ± 0.029	0.399 ± 0.016	0.379 ± 0.003	0.132 ± 0.005
Best MLP Configuration	1.853	0.430	0.305	**0.286**
$\mathrm{TN}_{\text{KNN-MBERT-LANG}}$	**1.789**	0.428	0.352	0.246

In our next experiment, we evaluate the impact of filtering in our TN-based approach. For all tasks, we compare the performances of $\mathrm{TN}_{\text{KNN-MBERT}}$ (i.e., no filtering), $\mathrm{TN}_{\text{KNN-MBERT-LANG}}$ (i.e., language based filtering), and $\mathrm{TN}_{\text{KNN-MBERT-CPV}}$ (i.e., CPV based filtering). The results are shown in Table 4.

We observe that filtering based on CPV slightly improves the performance in all tasks except for the $\mathrm{P}_{\mathrm{Effectiveness}}$ task. In addition, filtering based on language has a positive impact on all tasks, yielding the best results for all tasks. This might be because almost all European countries have different official languages. Therefore, filtering based on language also reduces the negative impact of macroeconomic differences across countries in our models.

6.2.2 Comparison against baselines

Now, we compare our models against the baseline models. In particular, we report the performance of the baseline methods explained in Section 6.1, fine-tuned MBERT and XLMR models, and the best-performing models in our previous experiments (i.e., $\mathrm{TN}_{\text{KNN-MBERT-LANG}}$ and MLP configurations yielding the highest score for each task).

We run our transformer-based methods three times (due to random seed value) and report the average score here. Table 5 presents the results.

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

Med$_{\text{SecCnt}}$ seems to be a highly strong baseline method, outperforming fine-tuned XLMR model, our MLP-based methods, and all other baselines in P$_{\text{NumOffers}}$ task. This suggests that the number of offers highly depends on the sector and country, without a high variance across the notices.

However, the fine-tuned MBERT model and TN$_{\text{KNN-MBERT-LANG}}$ outperform all baseline models in the P$_{\text{NumOffers}}$ task. All our methods outperform LRMEV, which utilizes macroeconomic variables, providing empirical evidence that the description of notices impacts the number of offers in PP.

In the P$_{\text{SingleOffer}}$ task, Med$_{\text{SecCnt}}$ achieves the best results compared with the other baselines. However, all of our models, except MLP$_{\text{MBERT}}$, outperform Med$_{\text{SecCnt}}$.

The fine-tuned MBERT model yields the highest classification accuracy. In the P$_{\text{FFAward}}$ task, all our methods outperform all baselines. Med$_{\text{SecCnt}}$ is again a strong baseline and outperforms all other baseline models. While the awarding of a foreign firm can be affected by many macroeconomic issues, it is noteworthy that our models are able to catch this macroeconomic situation based on the descriptions.

For P$_{\text{Effectiveness}}$, interestingly, KNN$_{\text{unigram}}$ outperforms all other baselines and our fine-tuned MBERT and XLMR models. This suggests that notices with particular phrases have similar effectiveness results. MLP with MBERT yields the highest score, and TN$_{\text{KNN-MBERT-LANG}}$ outperforms all baselines.

Regarding MBERT versus XLMR, we observe that fine-tuned MBERT outperforms fine-tuned XLMR in all tasks. In comparison with our models, our findings are mixed as no method consistently outperforms others. However, TN$_{\text{KNN-MBERT-LANG}}$ consistently outperforms all baselines in all tasks, whereas the fine-tuned models and MLP-based models outperform the baselines in three tasks.

In order to better understand the results, we calculate the precision and recall scores for each classification task. The precision score of the best-performing model for P$_{\text{SingleOffer}}$, P$_{\text{FFAward}}$, and P$_{\text{Effectiveness}}$ is 0.606, 0.511, and 0.434, respectively. Similarly, the recall score of the best-performing model for P$_{\text{SingleOffer}}$, P$_{\text{FFAward}}$, and P$_{\text{Effectiveness}}$ is 0.340, 0.321, and 0.159, respectively. We note that precision scores are higher than recall scores in all cases.

While our proposed models outperform baselines, their performance might be considered low. There might be many reasons for these scores. Firstly, we focus on predicting the outcome of PP processes by only using notice descriptions in order to explore their economic impact. However, obviously there are other factors affecting PP outcomes. For instance, if a country does not have any firm working on nuclear energy, a procurement call for building a nuclear energy plant cannot be won by a local firm regardless of the PP description. Therefore, we believe that the performance of our models can be improved by incorporating macroeconomic features of the countries. However, utilizing this kind of features is out of scope of this study. Furthermore, the label distribution is highly imbalanced with very few examples for positive classes. For instance, only 3.5% of the contracts in the train set are won by foreign firms. This kind of problems might be potentially solved by applying data augmentation methods. We leave this issue as a future work.

6.2.3 Prediction performance for each language

In order to further analyze multilingual aspect of our dataset, we calculate the performance of fine-tuned MBERT for each language and task. Figure 3 shows the results for the P$_{\text{NumOffers}}$ task. The lowest MAE scores of the fine-tuned MBERT are for Croatian, Estonian, Slovak, Danish, and Czech, whereas the highest MAE scores are for Spanish, German, Portuguese, Italian, and Greek. Generally, we observe that the MAE score is higher for the languages of economically developed countries than for those of economically developing countries. This might be because the number of offers can be generally high in economically developed countries due to having more firms suitable to participate in bidding processes. Therefore, we divide the MAE score for each language

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

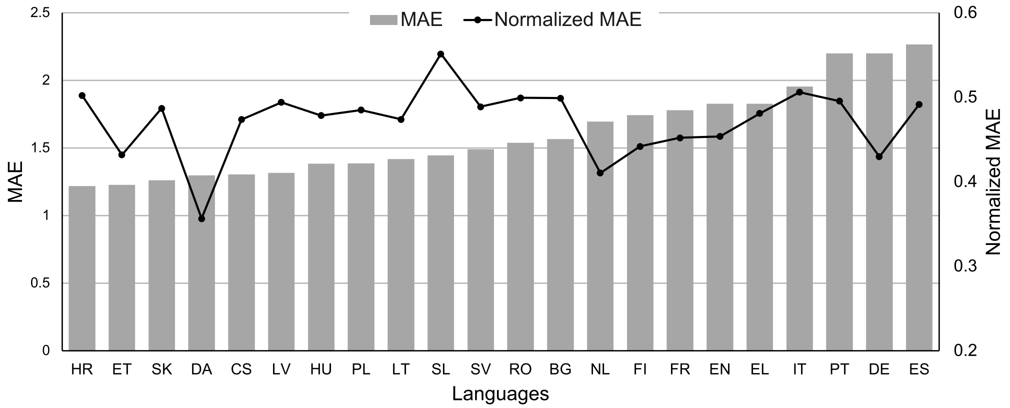


Figure 3. Performance of the fine-tuned MBERT in $P_{NumOffers}$ task for each language. In the normalization of the MAE scores, we divide the MAE scores by the average number of offers of the respective language.

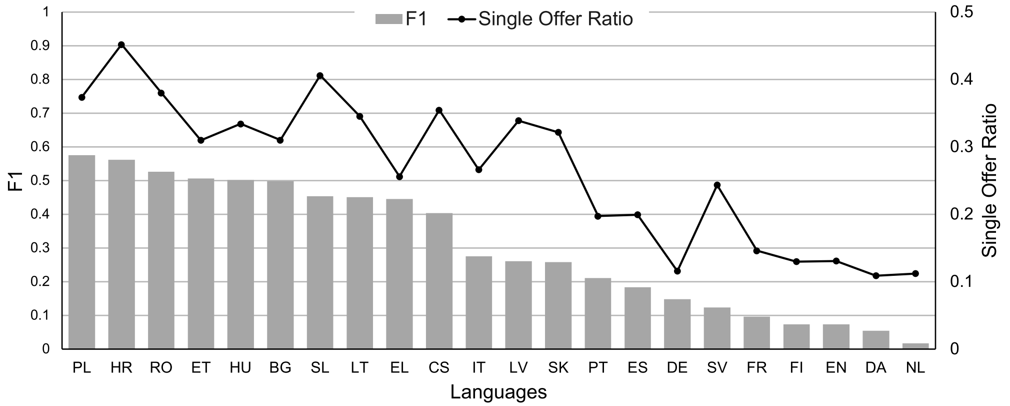


Figure 4. Performance of the fine-tuned MBERT in $P_{SingleOffer}$ task and the ratio of procurement notices with a single offer in the train set for each language.

by the average number of offers for each respective language. The results are shown as a line graph in Figure 3. Normalized MAE scores for languages are usually similar and have a lower variance than MAE scores. We achieve the lowest normalized MAE scores for Danish, Dutch, and German, which are Germanic languages. Their linguistic similarity to English, which is also a Germanic language, might be one of the factors that resulted in a higher performance than others.

Figure 4 presents F_1 score of the fine-tuned MBERT model and ratio of procurement notices with a single offer in the train set for each language. Generally, we observe a correlation between single offer ratio and F_1 score. For instance, we achieve the highest F_1 scores for Polish, Hungarian, and Romanian, which are also languages with the highest single offer ratio in the train set. In addition, we have the lowest F_1 scores for English, Danish, and Dutch, which are the languages with lowest single offer ratio. Overall, our results suggest that the performance of the model is affected by the number of labels in the train set instead of the linguistic features of languages. Therefore, the performance of the models might be also improved using data augmentation methods.

We present the F_1 score of the fine-tuned MBERT model and the ratio of procurement notices in which the award price exceeds the estimated price (i.e., ineffective procurement calls) in the train set for each language in Figure 5. We observe that while the fine-tuned MBERT model

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

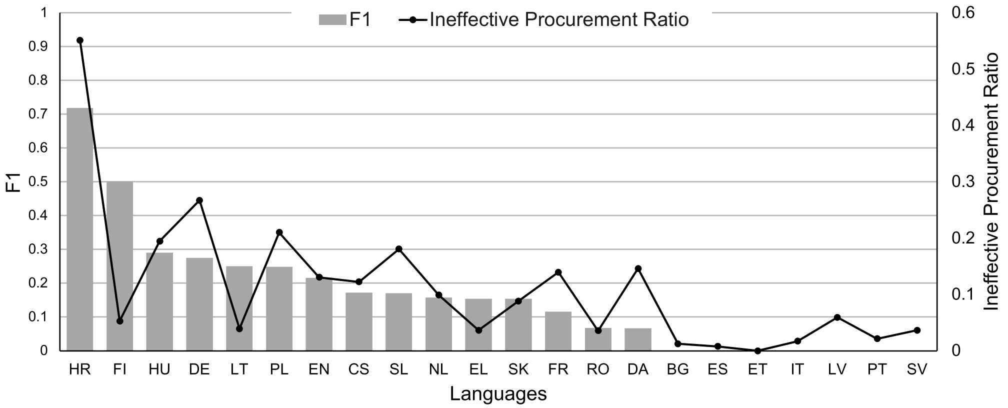


Figure 5. Performance of the fine-tuned MBERT in $\mathrm{P}_{\mathrm{Effectiveness}}$ task and the ratio of effective procurement notices in the train set for each language.

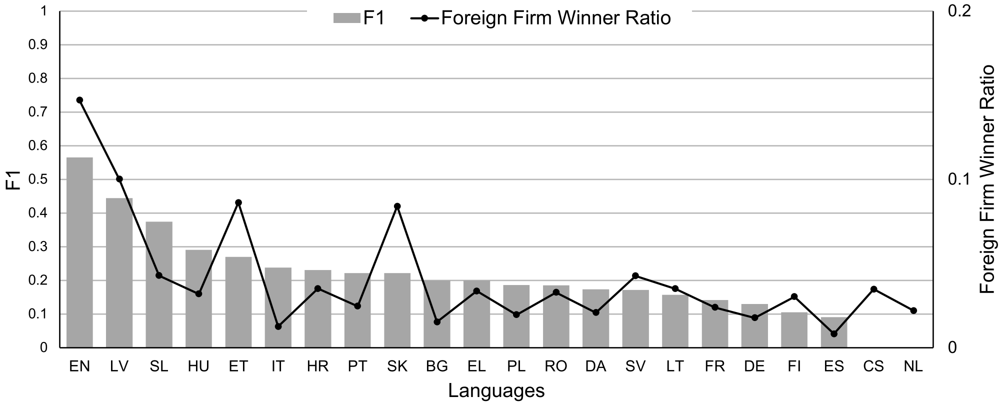


Figure 6. Performance of the fine-tuned MBERT in $\mathrm{P}_{\mathrm{FFAward}}$ task and the ratio of procurement notices in which a foreign company is awarded.

achieves a high prediction accuracy for particular languages (e.g., Croatian and Finnish), it cannot detect any ineffective PP call correctly for seven languages. Among these seven languages, most of them have very low ineffective procurement ratio (e.g., Spanish and Bulgarian). Furthermore, in our dataset, there is no cost-ineffective Estonian call. However, while the ineffective procurement ratio for Latvian PP calls is not highly limited, the fine-tuned MBERT model again achieves a F_1 score of zero. This might be because the total number of calls in Latvian is also highly limited (see Figure 2). Interestingly, the model performance is low for all Romance languages (i.e., Italian, Spanish, Portuguese, and French, and Romanian), while we do not see any clear pattern for other language families.

Finally, we present the results for the $\mathrm{P}_{\mathrm{FFAward}}$ task in Figure 6. Our observations are as follows. First, we do not observe any pattern regarding any language family. Second, the performance of the model seems mostly affected by the positive class ratio. Finally, the fine-tuned MBERT model interestingly cannot detect any contract in which a foreign firm is awarded for the Czech and Dutch languages.

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

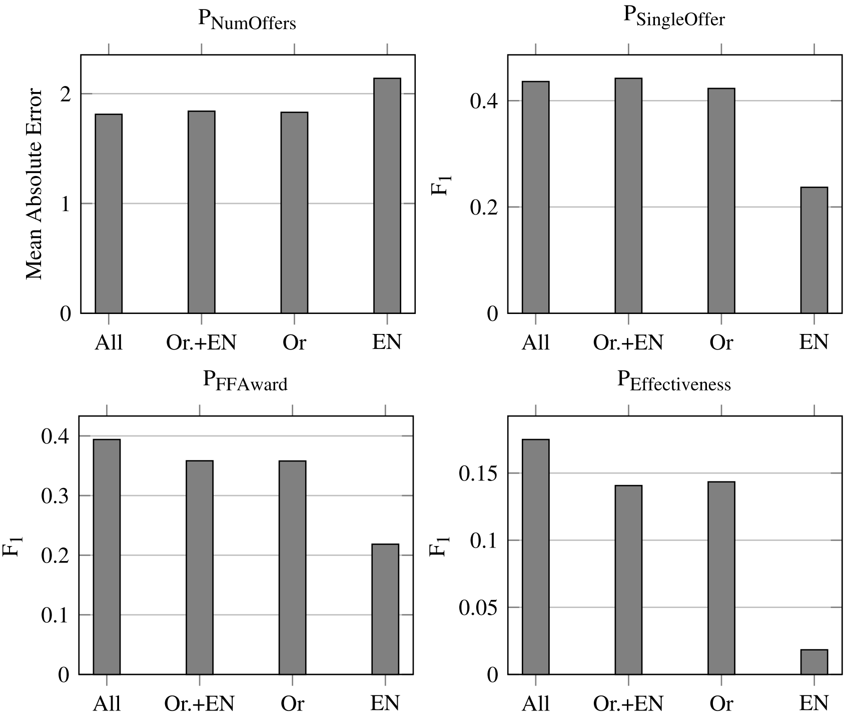


Figure 7. Performance of the fine-tuned MBERT model using various train data for all four tasks. “Or” stands for “Original”. The columns that correspond to the models trained using all data (All) show the average of the three runs.

6.2.4 Impact of multilingual training

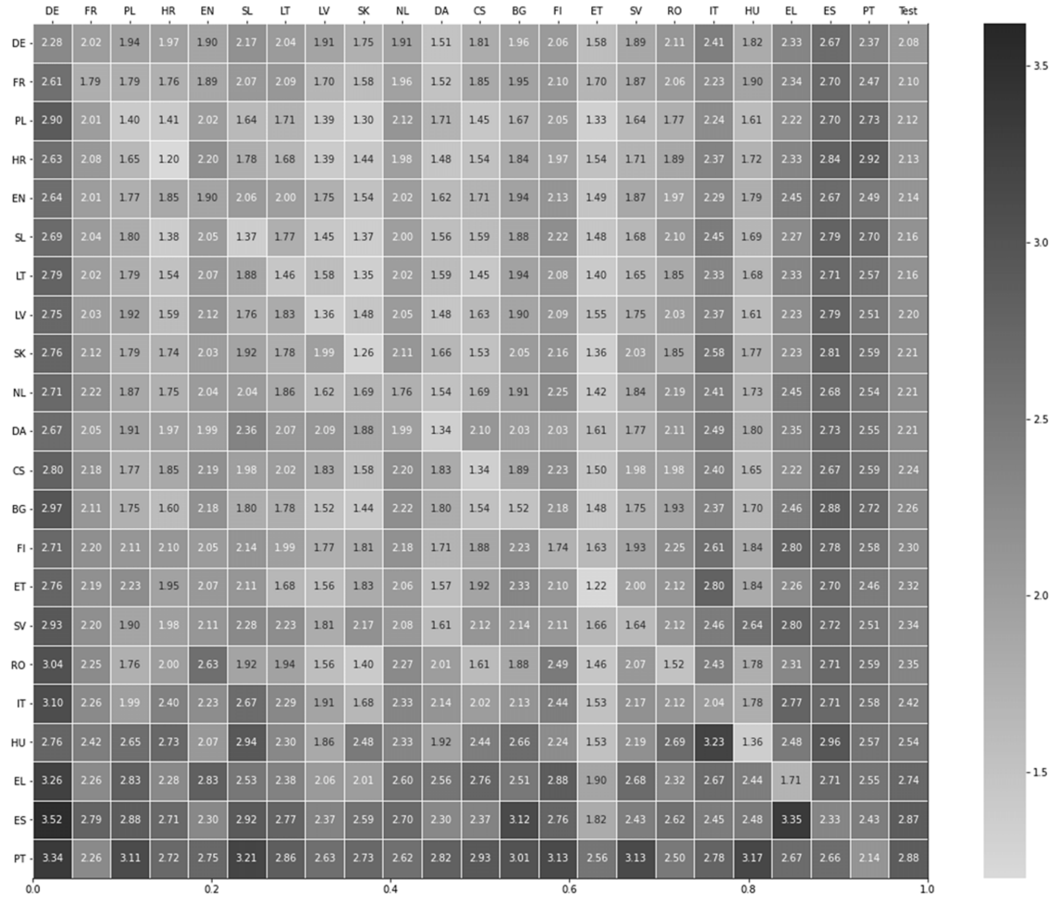
Covering all 22 languages might introduce noise in training our models. Therefore, in this experiment we investigate the impact of multilingual training. In particular, we fine-tune MBERT model for each task using the following data configurations: (1) fine-tuning a single model using all data as in previous experiments (*All*), (2) fine-tuning a separate MBERT model for each language using contract descriptions written in that language (*Original*), (3) fine-tuning a single MBERT model using contracts written in English (*EN*), and (4) fine-tuning a separate MBERT model for each language using contracts written in that language or in English (*Original+EN*). In the *Original* and *Original+EN* cases, we actually have 22 different models. To test the performance of these approaches, we separate the test data based on its language and make predictions using the MBERT model fine-tuned on that language and then gather the predictions to calculate the overall MAE and F_1 score. The results are shown in Figure 7.

We observe that training with the contracts in the original language yields much higher performance than cross-lingual training with English contracts in all four tasks. In addition, using English data in addition to the data in the original language has a very slight impact on the performance. However, training with all languages yields the best performance in $P_{NumOffers}$, $P_{FFAward}$, and $P_{Effectivness}$ tasks, respectively. This shows that multilingual training improves our prediction models even though there are huge macroeconomic differences across countries. In the $P_{SingleOffer}$ task, *Or+EN* has a slightly better performance than *All*.

6.2.5 Cross-lingual training with all languages

In this experiment, we investigate the performance of cross-lingual training for each language pair. In order to increase the readability of this paper, we not only present results only for the $P_{NumOffers}$

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press



	DE	FR	PL	HR	EN	SL	LT	LV	SK	NL	DA	CS	BG	FI	ET	SV	RO	IT	HU	EL	ES	PT	Test
DE	2.28	2.02	1.94	1.97	1.90	2.17	2.04	1.91	1.75	1.91	1.51	1.81	1.96	2.06	1.58	1.89	2.11	2.41	1.82	2.33	2.67	2.37	2.08
FR	2.61	1.79	1.79	1.76	1.89	2.07	2.09	1.70	1.58	1.96	1.52	1.85	1.95	2.10	1.70	1.87	2.06	2.23	1.90	2.34	2.70	2.47	2.10
PL	2.90	2.01	1.40	1.41	2.02	1.64	1.71	1.39	1.30	2.12	1.71	1.45	1.67	2.05	1.33	1.64	1.77	2.24	1.61	2.22	2.70	2.73	2.12
HR	2.63	2.08	1.65	1.20	2.20	1.78	1.68	1.39	1.44	1.98	1.48	1.54	1.84	1.97	1.54	1.71	1.89	2.37	1.72	2.33	2.84	2.92	2.13
EN	2.64	2.01	1.77	1.85	1.90	2.06	2.00	1.75	1.54	2.02	1.62	1.71	1.94	2.13	1.49	1.87	1.97	2.29	1.79	2.45	2.67	2.49	2.14
SL	2.69	2.04	1.80	1.38	2.05	1.37	1.77	1.45	1.37	2.00	1.56	1.59	1.88	2.22	1.48	1.68	2.10	2.45	1.69	2.27	2.79	2.70	2.16
LT	2.79	2.02	1.79	1.54	2.07	1.88	1.46	1.58	1.35	2.02	1.59	1.45	1.94	2.08	1.40	1.65	1.85	2.33	1.68	2.33	2.71	2.57	2.16
LV	2.75	2.03	1.92	1.59	2.12	1.76	1.83	1.36	1.48	2.05	1.48	1.63	1.90	2.09	1.55	1.75	2.03	2.37	1.61	2.23	2.79	2.51	2.20
SK	2.76	2.12	1.79	1.74	2.03	1.92	1.78	1.99	1.26	2.11	1.66	1.53	2.05	2.16	1.36	2.03	1.85	2.58	1.77	2.23	2.81	2.59	2.21
NL	2.71	2.22	1.87	1.75	2.04	2.04	1.86	1.62	1.69	1.76	1.54	1.69	1.91	2.25	1.42	1.84	2.19	2.41	1.73	2.45	2.68	2.54	2.21
DA	2.67	2.05	1.91	1.97	1.99	2.36	2.07	2.09	1.88	1.99	1.34	2.10	2.03	2.03	1.61	1.77	2.11	2.49	1.80	2.35	2.73	2.55	2.21
CS	2.80	2.18	1.77	1.85	2.19	1.98	2.02	1.83	1.58	2.20	1.83	1.34	1.89	2.23	1.50	1.98	1.98	2.40	1.65	2.22	2.67	2.59	2.24
BG	2.97	2.11	1.75	1.60	2.18	1.80	1.78	1.52	1.44	2.22	1.80	1.54	1.52	2.18	1.48	1.75	1.93	2.37	1.70	2.46	2.88	2.72	2.26
FI	2.71	2.20	2.11	2.10	2.05	2.14	1.99	1.77	1.81	2.18	1.71	1.88	2.23	1.74	1.63	1.93	2.25	2.61	1.84	2.80	2.78	2.58	2.30
ET	2.76	2.19	2.23	1.95	2.07	2.11	1.68	1.56	1.83	2.06	1.57	1.92	2.33	2.10	1.22	2.00	2.12	2.80	1.84	2.26	2.70	2.46	2.32
SV	2.93	2.20	1.90	1.98	2.11	2.28	2.23	1.81	2.17	2.08	1.61	2.12	2.14	2.11	1.66	1.64	2.12	2.46	2.64	2.80	2.72	2.51	2.34
RO	3.04	2.25	1.76	2.00	2.63	1.92	1.94	1.56	1.40	2.27	2.01	1.61	1.88	2.49	1.46	2.07	1.52	2.43	1.78	2.31	2.71	2.59	2.35
IT	3.10	2.26	1.99	2.40	2.23	2.67	2.29	1.91	1.68	2.33	2.14	2.02	2.13	2.44	1.53	2.17	2.12	2.04	1.78	2.77	2.71	2.58	2.42
HU	2.76	2.42	2.65	2.73	2.07	2.94	2.30	1.86	2.48	2.33	1.92	2.44	2.66	2.24	1.53	2.19	2.69	3.23	1.36	2.48	2.96	2.57	2.54
EL	3.26	2.26	2.83	2.28	2.83	2.53	2.38	2.06	2.01	2.60	2.56	2.76	2.51	2.88	1.90	2.68	2.32	2.67	2.44	1.71	2.71	2.55	2.74
ES	3.52	2.79	2.88	2.71	2.30	2.92	2.77	2.37	2.59	2.70	2.30	2.37	3.12	2.76	1.82	2.43	2.62	2.45	2.48	3.35	2.33	2.43	2.87
PT	3.34	2.26	3.11	2.72	2.75	3.21	2.86	2.63	2.73	2.62	2.82	2.93	3.01	3.13	2.56	3.13	2.50	2.78	3.17	2.67	2.66	2.14	2.88

Figure 8. Comparison of cross-lingual MAE scores for the $\mathrm{P}_{\mathrm{NumOffers}}$ task. Lighter colors indicate better MAE scores. The overall MAE score of each language is given in the right-most column (Test). Languages are sorted in ascending order based on their overall performance.

task but also discuss the results for other tasks. In particular, for each language, we fine-tune an MBERT model using notices in that language and then calculate its performance for all languages separately. The results are shown in Figure 8.

Our observations are as follows. First, as expected, we achieve the best result for each language when the test set and train set have documents in the same language. However, in other tasks, we also observe that in some cases cross-lingual training yields higher performance than training with descriptions in the same language. This might be due to the varying data sizes and label distributions for each language. Second, using only Portuguese documents causes the worst overall performance in the $\mathrm{P}_{\mathrm{NumOffers}}$ task. This might be because the notices in Portuguese are highly limited in our dataset (see Figure 1). However, fine-tuning with Spanish, Italian, and Romanian notices causes the second, fifth, and sixth worst performances, respectively, although they are among the most frequent languages in our dataset. As Spanish, Italian, Romanian, and Portuguese are linguistically similar (i.e., all of them are members of the Romance language family), linguistic features might be one of the factors impacting the performance. As a counterargument, French is also another Romance language but yields the second-best performance. Therefore, we think that macroeconomic similarities between countries might be another factor affecting cross-lingual training performance. In addition, we note that the language yielding the worst overall

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

performance varies across tasks. Finally, the best overall performance we achieve with cross-lingual training is 2.08, which is higher than the MAE scores of all baselines (see Table 5), suggesting that multilingual training is essential.

Overall, the performance of cross-lingual training varies across tasks, and there is no language that consistently outperforms the others. This suggests that we should be cautious when using PP data for a country to predict the outcome of other countries' PP calls. However, note that our experiments presented in Figure 7 suggest that using data from other countries as additional training data improves prediction performance.

7. Conclusion

In this work, we explore how to predict competition and cost-effectiveness of the EU's PP calls using notice descriptions. We formulate four prediction tasks: (1) the number of offers, (2) whether there will be a single offer, (3) whether the awarded firm is foreign, and (4) whether the contract price exceeds the estimated cost. The massive dataset shared by EU imposes several challenges such as spanning 22 different languages and highly imbalanced label distribution, and others. We investigate fine-tuning multilingual transformer models and also propose applying MLP and KNN methods with filtering training data and applying TN approach.

In our extensive experiments, we make the following observations. Firstly, the best-performing model changes across prediction tasks. In particular, fine-tuned MBERT achieves the highest performance in predicting whether the awarded firm is foreign and whether there will be a single offer. However, using TN method with language-based filtering and MLP model with MBERT embeddings outperform all other models in predicting the number of offers and ineffectiveness, respectively. Secondly, our TN-based method outperforms all baseline methods in all tasks. Thirdly, the performance of models is likely to change significantly across languages, suggesting that linguistic and macroeconomic issues of countries impact the prediction performance. Fourthly, multilingual training yields higher prediction performance than monolingual training. Finally, while macroeconomic conditions of countries seem to affect the outcome of procurement processes, our results suggest that description of notices also impacts the outcome such that we are able to predict the outcome based on descriptions more accurately than using macroeconomic knowledge about countries.

There are several research directions we plan to seek in the future. Firstly, we will investigate the impact of language models we use. In particular, we will explore language-agnostic models [e.g., LaBSE (Feng *et al.* 2020) and LASER (Artetxe and Schwenk 2019)]. In addition, we will investigate whether multilingual models reduce the prediction performance due to their multilingual aspect. Thus, we will pre-train a separate BERT model for each language using corresponding procurement notices and use these models in our prediction tasks. Secondly, our results show that the training data size and ratio of the positive class highly impact the performance of the models. Therefore, we will work on how to reduce the negative impact of imbalanced data distribution across languages. For instance, we can use data augmentation methods to increase the positive class ratio. Thirdly, we plan to adapt explainable artificial intelligence methods to help users to have insight about how our models work. We think that this is particularly important because as users understand the reasons for a predicted value, they can assess predictions better and act accordingly. Fourthly, we believe that text generation models such as GPT-3 can be beneficial to help authorities to prepare effective description calls. In particular, we envision a system which generates PP descriptions that will help to increase competitiveness and cost-effectiveness. Lastly, as we show that multilingual training improves prediction accuracy, we also plan to use the TED dataset for PP calls from non-EU countries.

Acknowledgments. This research was supported by the Scientific and Technological Research Council of Turkiye (TUBITAK) ARDEB 1001 (Grant No 119K986). Our opinions are our own.

References

Ahmia O., Béchet N. and Marteau P.-F. (2018). Two multilingual corpora extracted from the tenders electronic daily for machine learning and machine translation applications. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Aldana A., Falcón-Cortés A. and Larralde H. (2022). A machine learning model to identify corruption in México's public procurement contracts. *arXiv preprint arXiv:2211.01478*.

Aluru S. S., Mathew B., Saha P. and Mukherjee A. (2020). Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Artetxe M. and Schwenk H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* **7**, 597–610.

Bosio E. and Djankov S. (2020). *How Large Is Public Procurement?* World Bank Blogs.

Catelli R., Bevilacqua L., Mariniello N., di Carlo V. S., Magaldi M., Fujita H., De Pietro G. and Esposito M. (2022). Cross lingual transfer learning for sentiment analysis of Italian tripadvisor reviews. *Expert Systems with Applications* **209**, 118246.

Charron N., Dahlström C., Fazekas M. and Lapuente V. (2017). Careers, connections, and corruption risks: Investigating the impact of bureaucratic meritocracy on public procurement processes. *The Journal of Politics* **79**(1), 89–104.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L. and Stoyanov V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky D., Chai J., Schluter N. and Tetreault J. R. (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Association for Computational Linguistics, pp. 8440–8451.

Coviello D., Guglielmo A. and Spagnolo G. (2018). The effect of discretion on procurement performance. *Management Science* **64**(2), 715–738.

Demiray Y., Ozbayoglu A. and Tas B. (2015). Estimation of the number of participants in government tenders with computational intelligence. In *3rd Workshop on Social and Algorithmic Issues in Business Support (SAIBS)*, pp. 433–437.

Devlin J., Chang M., Lee K. and Toutanova K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein J., Doran C. and Solorio T. (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, pp. 4171–4186.

Fazekas M. and Blum J. R. (2021). Improving public procurement outcomes: Review of tools and the state of the evidence base. In *Policy Research Working Paper; No. 9690*. World Bank.

Fazekas M. and Kocsis G. (2020). Uncovering high-level corruption: Cross-national objective corruption risk indicators using public procurement data. *British Journal of Political Science* **50**(1), 155–164.

Fazekas M. and Tóth B. (2018). The extent and cost of corruption in transport infrastructure. New evidence from Europe. *Transportation Research Part A: Policy and Practice* **113**, 35–54.

Feng F., Yang Y., Cer D., Arivazhagan N. and Wang W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Gallego J., Rivero G. and Martínez J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting* **37**(1), 360–377.

García Rodríguez M. J., Rodríguez Montequín V., Ortega Fernández F. and Villanueva Balsera J. M. (2019). Public procurement announcements in Spain: Regulations, data analysis, and award price estimator using machine learning. *Complexity* **2019**, 1–20.

Gorgun M. K., Kutlu M. and Taş B. K. O. (2020). Predicting the number of bidders in public procurement. In *2020 5th International Conference on Computer Science and Engineering (UBMK)*. IEEE, pp. 360–365.

Gorgun M. K., Kutlu M. and Tas B. K. O. (2022). Information is essential for competitive and cost-effective public procurement. *Journal of Information Science*, 01655515221141042.

Gugler K., Weichselbaumer M. and Zulehner C. (2015). Competition in the economic crisis: Analysis of procurement auctions. *European Economic Review* **73**, 35–57.

Hoffer E. and Ailon N. (2015). Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. Cham: Springer, pp. 84–92.

Iimi A. (2006). Auction reforms for effective official development assistance. *Review of Industrial Organization* **28**(2), 109–128.

Imhof D. and Wallimann H. (2021). Detecting bid-rigging coalitions in different countries and auction formats. *International Review of Law and Economics* **68**, 106016.

Ishii R. (2009). Favor exchange in collusion: Empirical study of repeated procurement auctions in japan. *International Journal of Industrial Organization* **27**(2), 137–144.

Jang Y., Yi J., Son J., Kang H. and Lee J. (2019). Development of classification model for the level of bid price volatility of public construction project focused on analysis of pre-bid clarification document. In *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, vol. 36. IAARC Publications, pp. 1245–1253.

Karthikeyan K., Wang Z., Mayhew S. and Roth D. (2019). Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

https://doi.org/10.1017/S135132492300030X Published online by Cambridge University Press

Kutlina-Dimitrova Z. and Lakatos C. (2016). Determinants of direct cross-border public procurement in EU member states. *Review of World Economics* **152**(3), 501–528.

Lee J. and Yi J.-S. (2017). Predicting project's uncertainty risk in the bidding process by integrating unstructured text data and structured numerical data using text mining. *Applied Sciences* **7**(11), 1141.

Lima M., Silva R., de Souza Mendes F. L., de Carvalho L. R., Araujo A. and de Barros Vidal F. (2020). Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a bi-LSTM approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 1580–1588.

Mehrbod A. and Grilo A. (2018). Tender calls search using a procurement product named entity recogniser. *Advanced Engineering Informatics* **36**, 216–228.

Modrusan N., Rabuzin K. and Mrsic L. (2020). Improving public sector efficiency using advanced text mining in the procurement process. In *DATA*, pp. 200–206.

Onur İ., Özcan R. and Taş B. K. O. (2012a). Public procurement auctions and competition in Turkey. *Review of Industrial Organization* **40**(3), 207–223.

Onur I., Ozcan R. and Tas B. K. O. (2012b). Public procurement auctions and competition in Turkey. *Review of Industrial Organization* **40**(3), 207–223.

Pappas N. and Popescu-Belis A. (2017). Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1015–1025.

Pires T., Schlinger E. and Garrette D. (2019). How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001.

Piscitelli S., Arnaudo E. and Rossi C. (2021). Multilingual text classification from twitter during emergencies. In *2021 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, pp. 1–6.

Rabuzin K. and Modrusan N. (2019). Prediction of public procurement corruption indices using machine learning methods. In *KMIS*, pp. 333–340.

Simperl E., Corcho O., Grobelnik M., Roman D., Soylu A., Ruíz M. J. F., Gatti S., Taggart C., Klima U. S., Uliana A. F., Makgill I. and Lech T. C. (2018). Towards a knowledge graph based platform for public procurement. In *Research Conference on Metadata and Semantics Research*. Cham: Springer, pp. 317–323.

Taş B. K. O. (2022). Effect of public procurement regulation on competition and cost-effectiveness. *Journal of Regulatory Economics* **58**(1), 59–77.

Taş B. K. O., Dawar K., Holmes P. and Togan S. (2019). Does the wto government procurement agreement deliver what it promises? *World Trade Review* **18**(4), 609–634.

Van der Heijden N., Yannakoudakis H., Mishra P. and Shutova E. (2021). Multilingual and cross-lingual document classification: A meta-learning approach. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1966–1976.

Wachs J., Fazekas M. and Kertész J. (2021). Corruption risk in contracting markets: A network science perspective. *International Journal of Data Science and Analytics* **12**(1), 45–60.

Williams T. P. and Gong J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction* **43**, 23–29.

Wu S. and Dredze M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, pp. 833–844.

Zhang Z., Jasaitis T., Freeman R., Alfrjani R. and Funk A. (2023). Mining healthcare procurement data using text mining and natural language processing–reflection from an industrial project. *arXiv preprint* arXiv:2301.03458.

Cite this article: Acikalin UU, Gorgun MK, Kutlu M and Tas BKO (2024). How you describe procurement calls matters: Predicting outcome of public procurement using call descriptions. *Natural Language Engineering* **30**, 1255–1276. https://doi.org/10.1017/S135132492300030X