Business text classification with imbalanced data and moderately large label spaces for digital transformation

Muhammad Arslan, Christophe Cruz

Applied Network Science

RESEARCH **Open Access**

Business text classification with imbalanced data and moderately large label spaces for digital transformation

Muhammad Arslan[1*] and Christophe Cruz[1]

*Correspondence: muhammad.arslan@u-bourgogne.fr

[1] Laboratoire Interdisciplinaire Carnot de Bourgogne (ICB), Université de Bourgogne, 9 Av. Alain Savary, Dijon, France

Abstract

Digital transformation refers to an organization's use of digital technology to improve its products, services, and operations, aligning them with evolving business requirements. To demonstrate this transformative process, we present a real-life case study where a company seeks to automate the classification of their textual data rather than relying on manual methods. Transitioning to automated classification involves deploying machine learning models, which rely on pre-labeled datasets for training and making predictions on new data. However, upon receiving the dataset from the company, we faced challenges due to the imbalanced distribution of labels and moderately large label spaces. To tackle text classification with such a business dataset, we evaluated four distinct methods for multi-label text classification: fine-tuned Bidirectional Encoder Representations from Transformers (BERT), Binary Relevance, Classifier Chains, and Label Powerset. The results revealed that fine-tuned BERT significantly outperformed the other methods across key metrics like Accuracy, F1-score, Precision, and Recall. Binary Relevance also displayed competence in handling the dataset effectively, while Classifier Chains and Label Powerset exhibited comparatively less impressive performance. These findings highlight the remarkable effectiveness of fine-tuned BERT model and the Binary Relevance classifier in multi-label text classification tasks, particularly when dealing with imbalanced training datasets and moderately large label spaces. This positions them as valuable assets for businesses aiming to automate data classification in the digital transformation era.

Keywords: Business, Digital transformation, News documents, Performance comparison

Introduction

In today's digital era, businesses are experiencing an unprecedented surge in textual data sourced from various channels such as online news articles, press releases, and internal websites (Arslan and Cruz 2022). This wealth of data holds invaluable insights into new product launches, service enhancements, and other business endeavors. Prior to delving into any business analysis with sourced textual data, effective text classification stands as a crucial prerequisite. This process entails organizing text into distinct categories based on specific features, facilitating streamlined analysis and decision-making (Kim et al. 2006; Ur-Rahman and Harding 2012). Traditionally, text classification has been reliant on the company's taxonomy, which organizes business concepts of interest in a hierarchical manner (Arslan and Cruz 2022, 2023a). In this taxonomic-based approach, the aim is to identify relevant business concepts from the taxonomy within the business articles. The relevance of an article is then determined based on the frequency of occurrences of taxonomy concepts within it. However, manual text classification methods pose challenges, particularly when dealing with large volumes of business information, often leading to errors and inefficiencies.

In the contemporary landscape of data abundance, digital transformation (Arslan and Cruz 2023b) has become imperative for businesses striving not just to survive but to thrive in an intensely competitive environment. Digital transformation, which involves strategically integrating technology to streamline operations, improve decision-making, and unlock value on a large scale (Read et al. 2011), has emerged as the driving force behind reshaping how organizations operate, communicate, and innovate. A crucial aspect of digital transformation lies in automating tasks that were previously performed manually. This shift holds particular significance for enterprises grappling with vast amounts of textual data (Trincado-Munoz et al. 2023; He and Sun 2023; Kiener et al. 2023). Transitioning from manual text classification to automated machine-based methods marks a substantial stride towards harnessing data for actionable insights. This shift not only enhances the efficiency of business text classification but also minimizes the inherent risks of errors often associated with manual text classification.

To effectively drive the digital transformation process, transitioning from manual to machine-based text classification requires robust methods. Existing text classification methods involve selecting relevant features from the data to assign target categories or labels, typically classified into two forms: single-label and multi-label classification. This paper focuses solely on multi-label classification, a task where one text document can be assigned one or more labels (Read et al. 2011). After conducting a literature review, we identified BERT (González-Carvajal et al. 2020) and Problem Transformation approaches (Liu et al. 2017) as widely used models for multi-label text classification. These machine learning models are adept at handling extensive amounts of text, uncovering significant features embedded within the text and classifying it into various categories. Given these observations, we believed that these methods could be beneficial for our case study. However, their performance in scenarios where the dataset is imbalanced and features moderately large label spaces remains unexplored.

To bridge this gap, our study evaluates the effectiveness of BERT (González-Carvajal et al. 2020) and Problem Transformation approaches, including Binary Relevance, Classifier Chains, and Label Powerset, in classifying business texts using an imbalanced dataset containing a moderately large label spaces (a total of 80 distinct labels in our case). Our evaluation methodology involves several critical stages. Initially, we prepare the data, striving to reduce the issue of class imbalance inherent in the dataset. Subsequently, we proceed to model training, wherein each method undergoes training on the preprocessed dataset, with BERT fine-tuned for optimal performance. Finally, we conduct model evaluation, assessing the performance of each model using metrics such as Accuracy, Precision, Recall, and F1-score.

The paper's structure is outlined as follows: "Background" Section offers a review of the multi-label text classification approaches used in this article. "Analyzing text classification models for business-related text" Section introduces the proposed work. In "Results" Section, we delve into the results. "Discussion" Section presents the discussion, and lastly, "Conclusion" Section concludes the paper.

Background

Problem Transformation approaches constitute a versatile family of techniques employed in multi-label classification tasks. Their primary objective is to systematically convert the inherent complexity of the original multi-label problem into one or more simpler, more tractable classification tasks (Spolaôr et al. 2013). These approaches prove invaluable when dealing with multi-label problems that encompass an extensive array of potential labels. In such scenarios, the sheer number of possible labels can render the multi-label classification problem computationally expensive and particularly challenging.

Among the Problem Transformation techniques, several have gained prominence due to their effectiveness and adaptability. Three of the most widely employed Problem Transformation approaches include Binary Relevance, Classifier Chains, and Label Powerset (Luaces et al. 2012). Each of these techniques offers distinct advantages in handling multi-label classification challenges, and their selection often depends on the specific characteristics of the dataset and the nature of the problem at hand.

In Binary Relevance method (Read et al. 2021), a separate binary classifier is trained for each label, and each classifier predicts whether the input belongs to that particular label. The main advantage of the Binary Relevance method is its simplicity and flexibility. It can work with any binary classifier, and the classifiers can be trained independently, making it easy to add or remove labels without affecting the performance of other classifiers. However, the method does not consider any correlations between the labels, which may affect the overall accuracy of the multi-label classification task.

The Classifier Chain method (Read et al. 2021) uses a chain of binary classifiers to predict the labels. In this method, the labels are treated as a sequence, and the classifiers are trained in the order of the label sequence. The main advantage of the Classifier Chain method is its ability to model the correlations between labels, which can lead to improved accuracy in the multi-label classification task. However, the method can be computationally expensive, especially if there are many labels in the dataset.

The Label Powerset method (Read et al. 2014) involves transforming the multi-label problem into a multiclass problem. In this method, each unique combination of labels is treated as a separate class, and a multiclass classifier is trained to predict the class for each input. The main advantage of the Label Powerset method is its ability to handle any number of labels, and it can capture complex dependencies between labels. However, the method suffers from the curse of dimensionality, as the number of classes grows exponentially with the number of labels in the dataset.

In addition to Problem Transformation approaches, fine-tuning a pre-existing BERT model has gained significant traction as a popular and effective strategy in the existing literature (Lee et al. 2020). The process of fine-tuning a pre-trained BERT model for multi-label text classification involves training the model on a specific dataset, providing both labels and corresponding text inputs. During this training phase, the weights of the BERT model are iteratively adjusted to optimize its performance on the designated multi-label text classification task. Numerous studies have delved into sophisticated techniques for multi-label classification (Bogatinovski et al. 2022; Haghighian Roudsari et al. 2022; Huang et al. 2023; Zeng et al. 2024; Lefebvre et al. 2024). Nevertheless, their adaptability to imbalanced business-related datasets with moderately large label spaces remains unexplored in the literature. To bridge this gap, this article endeavors to present a comprehensive comparative analysis of four distinct techniques using a business-related dataset. This study seeks to furnish valuable insights into the efficacy of these methods for multi-label classification tasks within a business context, with the aim of informing and inspiring future research in this vital domain.

Analyzing text classification models for business-related text

In this section, we will present an in-depth overview of the dataset selected for our multi-label classification tasks. We will explore the dataset's structure, composition, and the preprocessing steps we conducted to ensure its suitability for our analysis. Subsequently, this dataset will serve as the foundation for training, fine-tuning, and the rigorous evaluation of multiple multi-label classification models. Through this exploration, we aim to provide a comprehensive understanding of the dataset's role in our study and offer transparency regarding our methodology for business text classification.

Dataset

As a critical initial step in our analysis of various multi-label classification models, we used the business dataset sourced from the French company, FirstECO. This dataset underscores our dedication to utilizing real-world data to ensure the integrity of our study. Constructed by extracting business news from various online sources spanning the period from 2017 to 2022, it comprises 28,941 texts, each potentially corresponding to one or more of 80 distinct labels. These labels encompass diverse aspects of the business domain, including Intangible Development, Activities, Products, Material Investment, Increased Standby, Financial Development, Company Life, Geographical Development, and Public Finances, among others. To maintain confidentiality, we cannot disclose the entire list of labels, and therefore, we present only seven labels from the dataset. Table 1 showcases text examples from the business dataset. Each text pertaining to business is tagged with two or more labels.

Each label in the business dataset is represented by varying numbers of text examples, ranging from 25 to over 4000 (as illustrated in Fig. 1). This variance highlights the dataset's imbalance, where some labels contain significantly more text examples than others. To address this issue, we have taken measures to ensure balance by setting a minimum threshold of 50–100 text examples per label. This confirms that each label has a sufficient representation in the dataset, reducing the effects of imbalance distribution of labels. Text preprocessing is an essential stage in the data pipeline, serving the critical purpose of cleansing and formatting text data to make it suitable for text classification. This multifaceted process takes raw text data as input and applies a series of essential preprocessing steps to each text entry. To begin, it carefully eliminates punctuation and numeric characters from the text, leveraging the power of regular expressions. This

Table 1 Sample extracts from the business dataset

No	Business text	Labels
1	Bordelais Transports GLS becomes Goëvia and recruits in 2020. The transport company GLS (Gonzalez Logistique Services—approx. 200 employees), based in Vayres (33), changes its name as it celebrates its 20th anniversary. Transports GLS thus becomes Goëvia. Furthermore, the company announces recruitments in transportation and logistics during the year 2020	"Recruitment", "Company identity"
2	MCA Ingénierie opens its capital to Capzanine and will accelerate its development in France and internationally. MCA Ingénierie (Levallois-Perret, 92–1000 employees—2016 revenue of €74.4 million), a service company operating in the field of engineering and high-technology consulting, has opened its capital to Capzanine (Paris, 75), which takes a minority stake. This operation will enable MCA Ingénierie to support its external growth and continue to accelerate its international development, initiated since 2013	"National development", "External growth", "International development", "Equity investment"
3	Eneco and Ophiliam Management are leading a €28 million project for the construction of photovoltaic hangars in Nouvelle-Aquitaine. The Dutch green energy production group Eneco and the manager Ophiliam Management (Paris, 75—via FPCI Volta Entreprises IV) are joining forces to develop a €28 million construction project in Nouvelle-Aquitaine. This project involves the construction of 200 buildings equipped with photovoltaic roofs, with a combined capacity of 20 megawatts. The park will notably be in Landes, Gironde, and Pyrénées-Atlantiques. Furthermore, it will serve to accommodate the agricultural activities of part of the members of the Maïsadour cooperative agri-food group. Engie will be responsible for the construction, with delivery and connection expected during 2018	"Construction of premises", "Partnership"
4	The hospital in Pau is considering privatizing the management of its parking lot, with the tender expected to be launched by the end of 2017. The Pau hospital (64) is proposing a project to privatize its 1,450-space parking lot by September 2019. The parking would become paid, allowing for better management of available spaces. The tender is expected to be launched before the end of 2017 and would cover a 10-year management lease	"Public equipment", "Contract"
5	The Ardèche-based company Ekibio acquires the Rhône-Alpes brand Pléniday and intends to expand internationally. Ekibio—subsidiary of the holding company Compagnie Biodiversité: Périgny, 17) specializes in the production of organic and fair-trade food products sold in specialized stores. In late February 2020, it acquired the dietary food brand Pléniday, which includes 31 dietary food references distributed across three ranges (low-sodium, hypoallergenic, and low-carbohydrate) and sold in specialized stores. This operation will enable Ekibio to expand internationally and establish subsidiaries in the medium term	"Sale", "Acquisition", "Establishment of units abroad"

Table 1 (continued)

No	Business text	Labels
6	Neoen and the Community of Communes of Portes de Romilly will build two new photovoltaic power plants in the Aéromia zone in Romilly-sur-Seine starting in 2022. Aéromia, the area of the former airfield in Romilly-sur-Seine (10), which has had a solar power plant covering 22 hectares since 2011, will welcome two new photovoltaic power plants. These will cover nearly 37 hectares and produce nearly 50 GWh per year, equivalent to the annual consumption of around 180,000 inhabitants. The project is led by the Community of Communes of Portes de Romilly (CCPRS) and the company Neoen (Paris, 75), a major player in photovoltaic energy production. Construction work will begin in 2022	"Material investment", "Implantation"

initial cleaning step ensures that extraneous symbols and digits do not introduce noise into the subsequent analysis.

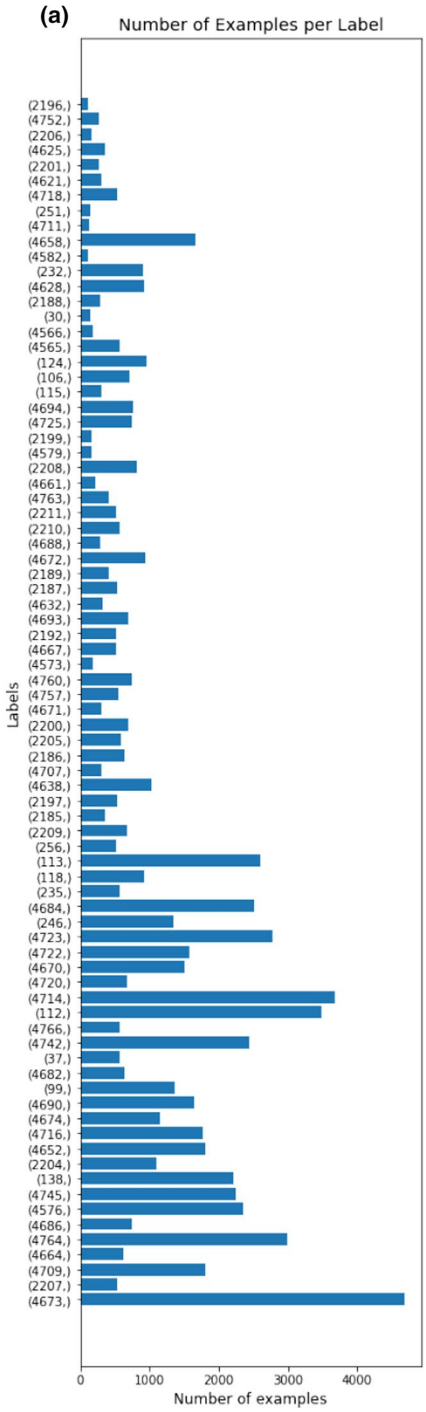
Following the initial cleanup, the text is subjected to a harmonizing transformation: it is converted to lowercase, ensuring uniformity in the text's case, which is vital for text analysis. Concurrently, the text is divided into distinct tokens using a tokenizer, segmenting it into manageable units for further processing. Once tokenized, the text undergoes another refinement process by having stop words excised from its content. These stop words, drawn from a predetermined set, are words like "the" "and" and "in" which are commonly occurring and often carry little meaningful information. Their removal streamlines the text and enhances the classification accuracy in subsequent analysis. Finally, the preprocessed words, now refined and devoid of unnecessary clutter, are thoughtfully reintegrated into a coherent string. This newly processed text is then systematically appended to a fresh list, which serves as a repository of the clean and formatted text data.

Implementation

The implementation centers around the execution of Problem Transformation techniques and the fine-tuning of the BERT model. This endeavor is not merely an exercise in technical prowess, but a strategic demonstration of which model can serve as the most practical choice for revolutionizing the landscape of business text classification. It caters specifically to companies poised to embark on digital transformation strategies, offering them a glimpse into the cutting-edge tools that can redefine how they interpret and utilize textual data in their evolving business ecosystems.

Problem transformation approaches

The process starts by importing necessary modules for data preparation, model training, and evaluation. The imported modules include GaussianNB and MultinomialNB for Naive Bayes classification, and Accuracy_score for evaluation metrics, train_test_split for splitting the data into training and testing sets, and TfidfVectorizer for transforming the text data into feature vectors. The scikit-multilearn library (Szymanski and Kajdanowicz 2019) is also imported to support multi-label classification problems. The process then creates an instance of TfidfVectorizer to convert the text data into feature vectors. The TfidfVectorizer is set to use inverse document frequency and normalization.



(b)

1	Label	Description
2	4757	Intangible development, activity, products
3	99	Material investment
4	41218	Increased standby
5	4682	Financial development
6	4625	Company life
7	134	Geographical development
8	27363	Public finances

Fig. 1 **a** Number of examples per label (left), **b** Description of each label (right)

Next, the process creates an instance of MultiLabelBinarizer (Pedregosa et al. 2011) and applies it to the labels. The MultiLabelBinarizer transforms the list of labels into a binary matrix where each row corresponds to an instance and each column corresponds to a unique label. Then, the data is split into training and testing sets using train_test_split, with a test size of 20%. Finally, the process returns X_train, X_test, Y_train, and Y_test, which are the feature matrices and label matrices for the training and testing sets, respectively. These matrices are used to train and evaluate multi-label classification models based on Problem Transformation approaches, which are; Binary Relevance, Classifier Chain, and Label Powerset in our case. Finally, the performance of these approaches is evaluated on the testing dataset using various metrics such as Accuracy, Precision, Recall, and F1-score (see Table 2).

Fine-tuning BERT

To fine-tune the existing BERT-based model for text classification, the model "bert-base-multilingual-cased" (Devlin et al. 2018) is chosen as it supports multiple languages. The process of fine-tuning starts with importing the necessary libraries such as NumPy (Oliphant 2006), Pandas (Reback et al. 2020; McKinney 2012), Scikit-learn (Kramer and Kramer 2016), PyTorch (Imambi et al. 2021), and Transformers (Wolf et al. 2020). Then, a number of hyperparameters are set, including Max_Len, which is set to 80 and represents the maximum length of input sequences. Train_Batch_Size is set to 16, and Valid_Batch_Size is set to 8. The process also specifies the number of Epochs to train the model, which is set to 5, and sets the learning rate to 1e−05. Additionally, a pre-trained BERT tokenizer using the BertTokenizer class is used from the transformer's library.

Furthermore, the data is split into training and testing datasets using a Train_Size of 0.8. The training dataset is created by randomly sampling 80% of the data from the original dataset, while the testing dataset is created by dropping the samples in the training dataset from the original dataset. The final training dataset has 23,153 samples, while the testing dataset has 5788 samples. The BERT model is trained on the training dataset by feeding batches of input sequences to the model, computing the loss, and optimizing the weights using backpropagation. Finally, the model's performance is evaluated on the testing dataset using Accuracy, Precision, Recall, and F1 score (see Table 2).

Accuracy, F1-score, Precision, and Recall are commonly used performance metrics that can be used to evaluate the effectiveness of a classifier. Accuracy measures the fraction of instances that are correctly classified by the classifier. Precision measures the fraction of correctly identified positive instances among all instances predicted as positive, while Recall measures the fraction of correctly identified positive instances among all positive instances in the data. Using these definitions, we can compute Accuracy, Precision and Recall as follows:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Table 2 Comparative analysis of different multi-label classification approaches

Parameter	Binary relevance	Classifier chains	Label powerset	Fine-tuned BERT
Accuracy	0.730	0.103	0.143	0.895
F1-score	0.936	0.539	0.278	0.978
Precision	0.952	0.590	0.350	0.948
Recall	0.922	0.495	0.230	0.988

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

where True positives (TP) are instances that are positive and are correctly classified as positive by the classifier. False positives (FP) are instances that are negative but are incorrectly classified as positive by the classifier. True negatives (TN) are instances that are negative and are correctly classified as negative by the classifier and False negatives (FN) are instances that are positive but are incorrectly classified as negative by the classifier. However, Accuracy may not be a suitable metric to use when the classes are imbalanced. This is because a classifier that simply predicts the majority class for all instances would achieve high accuracy even if it performs poorly on the minority class. To address this problem, we can use the F1-score, which is a harmonic mean of Precision and Recall. It combines both Precision and Recall into a single metric that balances the trade-off between them. The F1-score is defined as:

$$\text{F1} - \text{Score} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})$$

The F1-Score ranges between 0 and 1, with a value of 1 indicating perfect Precision and Recall. Note that Precision measures the accuracy of the positive predictions made by the classifier, while Recall measures the completeness of the positive predictions made by the classifier.

Results

Table 2 displays the performance of four different methods for multi-label text classification on a dataset with 80 possible labels. The first method, Binary Relevance, achieves an accuracy of 0.730, F1-score of 0.936, Precision of 0.952, and Recall of 0.922. This method creates a separate binary classifier for each label and assigns a label to each text independently. The second method, Classifier Chains, achieves an accuracy of 0.103, F1-score of 0.539, Precision of 0.590, and Recall of 0.495. This method builds a chain of classifiers where each classifier considers the predictions of the previous classifiers in the chain. The third method, Label Powerset, achieves the Accuracy of 0.143, F1-score of 0.278, Precision of 0.350, and Recall of 0.230. This method transforms the multi-label classification problem into a multi-class classification problem by assigning each unique combination of labels to a single class. The fourth method, fine-tuned BERT, achieves the highest accuracy of 0.895, F1-score of 0.978, Precision of 0.948, and Recall of 0.988.

The lowest F1-score (i.e. 0.278) of the Label Powerset method in the multi-label classification problem can be attributed to several factors (Read et al. 2014). Firstly, Label Powerset assumes label independence, disregarding potential correlations among labels present in real-world scenarios. This oversight makes it challenging to accurately predict label combinations, particularly with many labels. Moreover, the curse of dimensionality worsens the problem by exponentially expanding the feature space, resulting in overfitting and lower generalization capability. Finally, the computational complexity of training one model per label in the Label Powerset method can lead to inadequate training or overfitting, further impacting the F1-score.

Discussion

The study delves into the multi-label classification task, characterized by an imbalanced dataset and moderately large label spaces. While existing literature offers numerous studies on multi-label classification, this paper's contribution lies in its focus on text classification within a business-related dataset. The distinction between an ordinary dataset and a business dataset representing business opportunities lies in their specific focus, content, and purpose. Whereas ordinary datasets cover a wide range of information across various domains, business datasets are meticulously curated to capture data pertinent to potential business endeavors. However, classifying text within business datasets presents unique challenges. Unlike ordinary datasets where features or keywords are typically explicit and easily discernible, business texts may lack clearly defined keywords associated with specific concepts. For example, a text on hiring practices might not explicitly mention keywords like "recruitment". This inherent ambiguity poses a challenge for classification models in accurately tagging texts with associated taxonomy concepts. Consequently, text classification in the business domain necessitates more advanced methodologies to address implicit features and contextual intricacies, ensuring precise categorization and thorough analysis of business-related texts.

The business dataset supplied by the company served as the foundation for our experiment, where we compared the performance of four methods: Binary Relevance, Classifier Chains, Label Powerset, and fine-tuned BERT. We evaluated their effectiveness using metrics such as Accuracy, F1-Score, Precision, and Recall. Our analysis revealed that the fine-tuned BERT method outshone the other three, boasting high scores across all metrics. While Binary Relevance also demonstrated strong performance, Classifier Chains and Label Powerset lagged, particularly on the dataset. These results underscore the advantage of fine-tuning the pre-trained BERT model, as it allows for adaptation to specific applications. Despite BERT's pre-training on extensive text corpora, which grants it a deep understanding of language nuances, fine-tuning tailors the model to the task at hand by training it on a smaller, task-specific dataset. To replicate the results of a fine-tuned BERT model, one must adhere to the same pre-processing steps, architecture, and hyperparameters as the original experiment. Using identical evaluation metrics for comparison is also crucial. However, the choice of dataset for fine-tuning the BERT model heavily influences its performance. Different tasks necessitate distinct datasets, and the quality and size of the dataset greatly impact the model's generalization ability. Hence, selecting an appropriate dataset tailored to the specific task is vital for achieving optimal results.

Conclusion

Digital transformation necessitates reorganizing to maximize technology's potential and seamlessly integrating it across business operations. Through our case study, we showcased how leveraging machine learning models can automate text classification in business contexts. Despite encountering challenges related to imbalanced data and moderately large label spaces, our evaluation unveiled the superior performance of fine-tuned BERT compared to other methods. Additionally, the Binary Relevance classifier demonstrated good performance. This paper serves as a beacon, illuminating the path towards enhanced text classification and understanding within the realm of business-oriented applications. In the age of digital transformation, where the efficient processing and comprehension of vast volumes of textual data are paramount, this paper provides a strategic solution. By embracing the principles of fine-tuning BERT models or employing traditional Binary Relevance classifiers, companies can harness the power of existing models to accurately classify their textual datasets. This precision empowers them to extract valuable insights from classified business data, automate decision-making processes, and retain a competitive edge in a swiftly evolving business environment.

Acknowledgements
The authors thank the French company FirstECO (https://www.firsteco.fr/) for providing the dataset, the French government for the plan France Relance funding, and Cyril Nguyen Van for his assistance.

Author contributions
Muhammad Arslan, the author, took charge of model execution and manuscript preparation. Meanwhile, Christophe Cruz played a pivotal role in securing funding for the project.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Competing interests
The authors declare no competing interests.

Received: 13 September 2023 Accepted: 22 April 2024
Published online: 30 April 2024

References

Arslan M, Cruz C (2022) Semantic taxonomy enrichment to improve business text classification for dynamic environments. In: 2022 International conference on innovations in intelligent systems and applications (INISTA), IEEE. pp. 1–6, https://doi.org/10.1109/INISTA55318.2022.9894173

Arslan M, Cruz C (2023a) Imbalanced multi-label classification for business-related text with moderately large label spaces. arXiv preprint http://arxiv.org/abs/2306.07046

Arslan M, Cruz C (2023b) Enabling Digital transformation through business text classification with small datasets. In 2023 15th international conference on innovations in information technology (IIT), IEEE, pp. 38–42. https://doi.org/10.1109/IIT59782.2023.10366487

Bogatinovski J, Todorovski L, Džeroski S, Kocev D (2022) Comprehensive comparative study of multi-label classification methods. Expert Syst Appl 203:117215

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint http://arxiv.org/abs/1810.04805

González-Carvajal S, Garrido-Merchán EC (2020) Comparing BERT against traditional machine learning text classification. arXiv preprint http://arxiv.org/abs/2005.13012

Haghighian Roudsari A, Afshar J, Lee W, Lee S (2022) PatentNet: multi-label classification of patent documents using deep learning-based language understanding. Scientometrics 127(1):207–231

He J, Sun B (2023) Digital transformation, dynamic capability and green technology innovation: empirical evidence based on text analysis methods. In: Proceedings of the 2nd International conference on big data economy and digital management, BDEDM 2023, January 6–8, 2023, Changsha, China

Huang A, Xu R, Chen Y, Guo M (2023) Research on multi-label user classification of social media based on ML-KNN algorithm. Technol Forecast Soc Chang 188:122271

Imambi S, Prakash KB, Kanagachidambaresan GR (2021) PyTorch. Programming with TensorFlow: solution for edge computing applications, pp 87–104.

Kiener F, Eggenberger C, Backes-Gellner U (2023) The role of occupational skill sets in the digital transformation: how IT progress shapes returns to specialization and social skills. J Bus Econ 94(1):75–111

Kim SB, Han KS, Rim HC, Myaeng SH (2006) Some effective techniques for naive bayes text classification. IEEE Trans Knowl Data Eng 18(11):1457–1466

Kramer O, Kramer O (2016) Scikit-learn. Machine learning for evolution strategies, pp 45–53

Lee JS, Hsiang J (2020) Patent classification by fine-tuning BERT language model. World Patent Inf 61:101965

Lefebvre G, Elghazel H, Guillet T, Aussem A, Sonnati M (2024) A new sentence embedding framework for the education and professional training domain with application to hierarchical multi-label text classification. Data Knowl Eng 150:102281

Liu J, Chang WC, Wu Y, Yang Y (2017) Deep learning for extreme multi-label text classification. In: Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval, pp 115–124

Luaces O, Díez J, Barranquero J, del Coz JJ, Bahamonde A (2012) Binary relevance efficacy for multilabel classification. Progress Artif Intell 1:303–313
McKinney W (2012) Python for data analysis: data wrangling with Pandas, NumPy, and IPython. "O'Reilly Media, Inc."
Oliphant TE (2006) Guide to numpy, vol 1. Trelgol Publishing, USA, p 85
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Duchesnay É (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
Read J, Pfahringer B, Holmes G, Frank E (2011) Classifier chains for multi-label classification. Mach Learn 85:333–359
Read J, Pfahringer B, Holmes G, Frank E (2021) Classifier chains: a review and perspectives. J Artif Intell Res 70:683–718
Read J, Puurula A, Bifet A (2014) Multi-label classification with meta-labels. In: 2014 IEEE international conference on data mining, IEEE, pp 941–946
Reback J, McKinney W, Van Den Bossche J, Augspurger T, Cloud P, Klein A, Seabold S (2020) Pandas-dev/pandas: Pandas 1.0. 5. Zenodo
Spolaôr N, Cherman EA, Monard MC, Lee HD (2013) A comparison of multi-label feature selection methods using the problem transformation approach. Electron Notes Theor Comput Sci 292:135–151
Szymanski P, Kajdanowicz T (2019) Scikit-multilearn: a scikit-based Python environment for performing multi-label classification. J Mach Learn Res 20(1):209–230
Trincado-Munoz F, van Meeteren M, Rubin TH, Vorley T (2023) Digital transformation in the world city networks' advanced producer services complex: A technology space analysis. Geoforum. https://doi.org/10.1016/j.geoforum.2023.103721
Ur-Rahman N, Harding JA (2012) Textual data mining for industrial knowledge management and text classification: A business oriented approach. Expert Syst Appl 39(5):4729–4739
Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Rush AM (2020) Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 38–45
Zeng D, Zha E, Kuang J, Shen Y (2024) Multi-label text classification based on semantic-sensitive graph convolutional network. Knowl-Based Syst 284:111303

Publisher's Note