Du, Yu, Lavarec, Erwann, and Lalouette, Colin (2025) Effectiveness of commercial text embedding models for multilingual, multi-class SaaS software classification: A practical study. Journal of Computer-Assisted Linguistic Research 9: 1-17. ISSN: 2530-9455.

Effectiveness of commercial text embedding models for multilingual, multi-class SaaS software classification: A practical study

Yu Du*[1], Erwann Lavarec[1], Colin Lalouette[1]
[1]Cloud-is-Mine R&D, Montpellier, France
*Corresponding author: du.yu.1411@gmail.com
Received: 25 June 2025 / Accepted: 30 July 2025 / Published: 20 November 2025

Abstract
In the rapidly evolving field of Software as a Service (SaaS), the accurate categorization of multilingual SaaS applications represents a significant challenge due to the inherent linguistic diversity and continuous growth in available software categories. This study investigates the application of commercial text embedding models, which transform textual data into numerical representations, for multilingual, large-scale, multi-class software classification tasks. We systematically compare various text embedding models integrated with classification algorithms, examining their predictive performance and transfer learning capabilities across multiple languages. Our experiments demonstrate that these embedding models exhibit substantial robustness and efficacy in both monolingual and cross-lingual classification contexts. Notably, a multi-layer perceptron classifier trained on bilingual datasets (French and English) using OpenAI's *text-embedding-3-large* embedding model achieved high accuracy (0.90) and F1-score (0.78), even when evaluated on languages not represented in the training corpus. This research not only offers valuable insights for professionals and practitioners in the SaaS sector but also lays the groundwork for further research in advanced applications, crucial for handling the extensive textual data in the contemporary digital marketplace.

Keywords: Natural Language Processing, Text Classification, Text Embedding, Large Language Model, CamemBERT, Software-as-a-Service

1. Introduction

In the dynamic Software as a Service (SaaS) landscape, the task of accurately categorizing software products is a crucial yet complex challenge. This categorization is not just a matter of organizational convenience, but a fundamental process that influences customer discovery, targeted marketing, and the overall user experience in digital marketplaces (Raghavan et al.

DOI: https://doi.org/10.4995/jclr.2025.24200
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

2020). The diversity and rapid evolution of SaaS offerings exacerbate this challenge, as new categories emerge and existing ones evolve. In this context, traditional software classification methods, which often rely on manual categorization, are increasingly inadequate in terms of both scalability and accuracy.

The global SaaS market has undergone rapid expansion in recent years. For instance, in the Customer Information System market alone, projections estimate that the market will reach USD 3.26 billion by 2030, driven by a significant compound annual growth rate of 12.8% (MarketsandMarkets 2025). This growth has been driven by increased adoption across diverse sectors, including healthcare, finance, and education. This further accentuates the multilingual and multicultural challenges inherent in software categorization. Furthermore, the increasing prevalence of SaaS solutions designed to cater to global customer bases has led to a growing need for effective multilingual processing capabilities. These capabilities are essential for ensuring the accurate categorization of products, the personalization of user experiences, and the efficient discovery of products by users.

The rapid evolution of Natural Language Processing (NLP) techniques, particularly text embedding models, has catalyzed a wave of innovation across multiple domains. The evolution of the field from basic methods such as Term Frequency-Inverse Document Frequency (TF-IDF) to sophisticated deep learning-based models such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer (GPT) reflects a significant shift in capabilities and performance (Spärck Jones 1972; Radford et al. 2018; Devlin et al. 2019). The aim of this paper is to explore the potential of advanced NLP techniques to rethink the way SaaS software is categorized.

The difficulties in categorizing SaaS software lie in three main aspects: the large-scale classification challenge (involving hundreds of unique categories), the class imbalance problem (leading to biased classification results that favor majority classes) (Aravamuthan et al. 2022), and the multilingualism challenge. Previous research has shown that the large-scale classification and class imbalance issues can be alleviated by using pre-trained BERT models with text augmentation techniques (Du et al. 2023). The present study goes one step further and explores how the multilingualism aspect can be addressed by using commercial text embedding models such as OpenAI's *text-embedding-ada-002* model (Greene et al. 2022). Given the global nature of the SaaS industry and the myriad of languages used by its stakeholders, the effectiveness of these commercial text embedding models in understanding and categorizing software in different languages and their ability to transfer learning across languages are key aspects that this study aims to investigate.

Furthermore, a key concern with the use of these commercial embedding models is their lack of transparency, particularly with regard to the nature of their training data (Wang et al. 2024). This opacity raises critical questions about the genericity of these models when applied to a diverse, multilingual scenario. In this context, the study aims to provide a nuanced examination of commercial text embedding models applied to the multilingual, large-scale multi-class classification of SaaS software. By evaluating the performance of these models and exploring their limitations and capabilities, we aim to provide insights that could shape the future of

software categorization in the SaaS industry, a sector defined by its global reach and linguistic diversity. Through this study, we contribute to the broader discourse on the applicability of NLP technologies to real-world, industry-specific challenges, highlighting both their potential and the critical need for transparency and inclusivity in their development and deployment.

The structure of the paper is as follows. Section 2 provides a comprehensive overview of the development of text embedding methods, as well as existing research with similar aims. Section 3 describes the dataset, the text embedding models and the evaluation protocols used in our study. Section 4 presents and discusses the results regarding the performance of these models in the context of multilingual, multi-class SaaS software classification. Section 5 focuses on the main limitation of the study. Finally, Section 6 concludes the study by summarizing the key findings, outlining practical implications, and suggesting avenues for future research.

2. Related Works

Text embedding models have become an integral part of NLP, allowing text to be converted into numerical vector representations while capturing the rich semantic information and relationships inherent in the text (Incitti et al. 2023; Patil et al. 2023). The evolution of text representation approaches began with purely statistical text vectorization methods such as Bag of Words (BoW) and TF-IDF (Spärck Jones 1972; Zhang et al. 2010). These techniques were essential during the initial development of NLP, as they primarily aimed to convert text into numerical vectors based on the frequency and importance of words. However, their inability to capture the deeper semantic meanings or contextual nuances of words was a significant limitation. This led to the development of more sophisticated text embedding models based on neural networks.

The introduction of Peters et al.'s (2018) Embeddings from Language Models marked a major step forward by providing context-dependent embeddings. This innovation was followed by models such as BERT, RoBERTa (Robustly Optimized BERT approach), GPT and E5 (EmbEddings from bidirEctional Encoder rEpresentations), which were based on the Transformer architecture (Radford et al. 2018; Devlin et al. 2019; Y. Liu et al. 2019; Wang et al. 2024; Vaswani et al. 2017). These models greatly improved the understanding of context and semantics in text by looking at the whole sequence of words rather than just individual words in isolation. Further advances came with the development of multilingual embeddings such as m-BERT (Pires et al. 2019) and XLM-R (Conneau et al. 2020). These models, trained on text from different languages, extended the capabilities of embedding models to multiple linguistic contexts, which is particularly useful for applications involving multiple languages.

The vector representation of textual data (i.e., embedding) remains latent and unintelligible to humans until we couple it with downstream tasks such as text classification/clustering, sentiment analysis, question answering, etc. To this end, the output of text embedding models is typically used to train machine learning algorithms, which then serve as inference engines for various types of prediction tasks. In contrast to the traditional methods of vectorizing textual data in the early phases of NLP, which typically required extensive data preprocessing (such as dictionary construction, removal of stop-words, lemmatization, and feature engineering), the

advantage of using advanced text embedding models lies in their ability to offer standardized vector representations that are readily usable for machine learning algorithms.

Ravi and Kulkarni (2023) conducted a comparative study of four open-source text embedding techniques for clustering social media data using the K-means algorithm. The authors compared BERT with three previous models – Word2Vec (Mikolov et al. 2013), Doc2Vec (Le and Mikolov 2014), GloVe (Pennington et al. 2014) – and showed that the BERT model gave the best results. Licht (2023) conducted a comparative analysis of two types of methods for binary text classification tasks, in the context of classifying political texts from the Comparative Manifestos Project corpus[1]. The first method combined BoW with Machine Translation (i.e., BoW+MT), while the second used open-source Multilingual Sentence Embeddings (MSE). Based on their analysis, MSE generally produce more reliable results than BoW+MT, especially when dealing with limited labelled data. These results underline the effectiveness of MSE in cross-lingual text classification tasks, offering a valuable alternative to traditional methods, especially in scenarios with limited data labelling resources.

The main limitation of the previously mentioned open-source text embedding models, such as BERT, RoBERTa, m-BERT, E5, etc., lies in their restricted input length, which is typically capped at 512 tokens. This constraint is due to the significant memory and computational resources required during the training phase of these models. Consequently, for longer texts, it is necessary to use truncation strategies, which inevitably leads to a loss of information.

In contrast, proprietary commercial text embedding models have emerged that can handle texts of up to 8,191 tokens. However, relatively few studies have focused on analyzing the performance of these advanced commercial models in applied domains. Balikas (2023) presented a comparative study of leading commercial embedding models for the question-answering task, including those from OpenAI[2] and Cohere[3], using their internal Salesforce custom datasets. Focused on standardized metrics across three domains, their analysis aims to help practitioners and researchers make informed decisions when selecting the most appropriate solution for specific needs. Pattun and Kumar (2023) delved into the capabilities of OpenAI's commercial embedding models (i.e., the *text-embedding-ada-002* model) for detecting emotions and analyzing sentiment in user-generated text. The research illuminates the relatively underexplored yet substantial benefits of GPT embeddings in emotion analysis, with an emphasis on how these embeddings enhance machine learning algorithms. Their findings reveal that GPT embeddings significantly boost the accuracy and efficiency of these algorithms. This underscores the effectiveness of OpenAI's embedding models in capturing complex linguistic data, marking a notable advance in emotion classification techniques. In Aytekin and Ayhan Erdem's (2023) research, OpenAI's *text-embedding-ada-002* model was also used for multi-class irony classification. Their study showed that these obtained vector representations, when integrated with the Naive Bayes classifier, produced the highest F1 score, demonstrating the effectiveness of the embeddings.

[1] https://manifesto-project.wzb.eu/datasets/MPDS2020a
[2] https://platform.openai.com/docs/guides/embeddings
[3] https://docs.cohere.com/reference/embed

3. Methodology and Experimental setup

This section provides in detail the methodology for analyzing the effectiveness of commercial text embedding models in the context of multilingual, multi-class SaaS software classification. It encompasses the specifics of the dataset used, the models compared, and the evaluation protocol of the conducted study.

3.1. Dataset

The dataset used in the study is an enriched version of the one used in our previously conducted research (Du et al. 2023), which focused on using text augmentation techniques to handle imbalanced classification. Specifically, to construct the dataset, we gathered software description data from three different data sources (cf. Table 1 below): Appvizer[4], GetApp[5], and Capterra[6].

Data source	# of samples (percentage)	Languages
Appvizer	79,440 (24%)	French, English
GetApp	92,680 (28%)	French, English
Capterra	158,880 (48%)	French, English

TABLE 1. THE DATASET OF SAAS SOFTWARE DESCRIPTIONS USED IN THE STUDY

To mitigate the class imbalance issue where certain popular SaaS categories such as Customer Relationship Management have much more training samples compared to niche categories such as Hospital Management, we adopted text data augmentation techniques, as discussed in the previously conducted work (Du et al. 2023). Differently yet, we employed a more robust augmentation approach as described in Dai et al.'s (2025) ChatAug model. The authors propose to prompt large language models (LLM), such as GPT-3.5, to generate new text samples. Specifically, in our study, OpenAI's *gpt-3.5-turbo* model was used to rephrase existing samples, thereby augmenting the volume of data for underrepresented software categories.

As a result, the enhanced dataset comprises approximately 331,000 SaaS software descriptions, with data in two languages: French and English. Each language comprises the 50% of the dataset, evenly distributed across 331 unique labels. These labels correspond to SaaS software categories, which are derived from the categorization schemes used by the considered SaaS comparison platforms (Appvizer, GetApp, and Capterra). While not based on a formal ontology, these categories reflect practical, real-world software classifications used in the SaaS domain. It is worth noticing that the French (resp. English) samples are not a simple translation of the English (resp. French) ones in the dataset. This is because these software descriptions were

[4] https://www.appvizer.com/
[5] https://www.getapp.com/
[6] https://www.capterra.com/

independently collected from localized websites, capturing distinct linguistic nuances and market-specific terminology.

3.2. Compared models

The primary objective of this study is to assess the effectiveness of commercial text embedding models for multilingual, large-scale multi-class classification tasks (encompassing over 300 unique classes) in the SaaS industry. To achieve this, we evaluated the most prominent text embedding models hosted by OpenAI: *text-embedding-ada-002*, released in December 2022, and its more recent, enhanced version *text-embedding-3-large*, released in January 2024 (OpenAI 2024).

As discussed in the preceding section, the numerical representations generated by text embedding models are typically utilized as input for training machine learning models. However, since the focus here is not on comparing different machine learning models but rather on examining the performance of the embeddings themselves, we opted for the Multi-Layer Perceptron (MLP) model as the classifier in the experiments. MLP is a robust neural network architecture well-suited for handling nonlinear data such as text.

A robust embedding model ought to be able to place semantically similar texts in proximity within the latent vector space, regardless of the language of the input texts. In other words, consider the following two texts: "I love reading romantic books" and "J'aime lire les livres romantiques". To a human, these two sentences convey the same meaning; thus, a competent text embedding model should be capable of positioning them closely in the latent vector space. To evaluate the effectiveness of commercial embedding models in providing robust representations that facilitate cross-lingual transfer learning, we construct and test the following classifiers[7] in the study:

- *MLP-CamemBERT (fr)*: The MLP classifier built with embeddings generated by the open-source CamemBERT model on the French samples within the dataset. We also opted to include Martin et al.'s (2020) CamemBERT model, which is considered the French version of BERT. This was done to draw comparisons with OpenAI's commercial text embedding models.
- *MLP-OpenAI-Embed-V2 (fr)*: The MLP classifier built with embeddings generated by OpenAI's *text-embedding-ada-002* model on the French samples within the dataset.
- *MLP-OpenAI-Embed-V2 (en)*: The MLP classifier built with OpenAI's *text-embedding-ada-002* model on the English samples within the dataset.
- *MLP-OpenAI-Embed-V2 (fr-en)*: The MLP classifier built with OpenAI's *text-embedding-ada-002* model on the entire dataset, including English and French text (cf. Table 1).
- *MLP-OpenAI-Embed-V3 (fr-en)*: The MLP classifier built with OpenAI's *text-embedding-3-*

[7] The names of these classifiers are the name of the machine learning model (e.g., MLP), combined with the name of the text embedding model

large model on the entire dataset, including English and French text (cf. Table 1).

The rationale for building different classifiers for different languages is to evaluate the effectiveness of text embeddings in cross-lingual transfer learning scenarios. Specifically, the goal is to evaluate how well a classifier trained on embeddings of French texts performs when applied to texts in other languages, thereby testing the generalization and transfer capabilities of the embedding model across linguistic boundaries.

3.3. Evaluation protocol

Each classifier mentioned in the previous section was trained using a 5-fold cross-validation approach on the corresponding training set. This approach involves dividing the training data into five parts, using four for training and one for validation, and rotating this process five times to enhance the model's robustness and accuracy. For example, the classifier named *MLP-OpenAI-Embed-V2 (en)* consists of training an MLP classifier with OpenAI's *text-embedding-ada-002* embedding model on the English samples within the dataset.

3.3.1. Evaluation metrics

To assess the performance of the classifiers built upon the embedding models, the following metrics are considered:

- *Accuracy@K*. This metric measures the proportion of testing samples for which the correct label is among the top-k predictions made by the classifier. In the experiments, we considered Accuracy@1 and Accuracy@3.
- *Precision*. It represents the average of the precision metric for each class[8], which is the ratio of correctly predicted instances of a given class to the total instances predicted as that class.
- *Recall*. This metric represents the average of the recall metric for each class, defined as the ratio of correctly predicted instances of a given class to the actual instances of that class.
- *F1-score*. It represents the harmonic mean of Precision and Recall, providing a balanced measure of the classifier's precision and recall across all classes.

3.3.2. Transfer learning capability for multilingual classification

In order to evaluate the effectiveness of the embeddings obtained by the OpenAI models in dealing with multilingualism, we decided to test each of the constructed classifiers on test samples that are not in the same language as their training data. For example, consider *MLP-OpenAI-Embed-V2 (fr)*, which was trained using French text. In addition to testing its performance on French test data with the metrics (cf. Section 3.3.1), we also tested its predictive performance on test samples in languages other than French. To this end, we also collected a test dataset of 16,108 SaaS software descriptions in four other languages, such as Spanish,

[8] We used macro-averaging for precision, recall, and F1-score computations, meaning that metrics are computed independently for each class and then averaged.

Italian, German and Portuguese—none of which were present in the training data. This setup allows us to assess whether the embeddings themselves capture sufficient cross-lingual semantic alignment to enable effective classification across languages.

4. Results and discussion

This section provides an in-depth analysis of the metrics and results obtained from the experiments detailed in the preceding section (cf. Section 3). It focuses on a comparative evaluation of the predictive performance of classifiers that are based on the tested commercial text embedding models, within the scope of a large-scale multi-class classification task. This task encompasses hundreds of unique labels pertinent to the SaaS software industry. Moreover, the section also explores how these classifiers perform on multilingual text not previously encountered in the training data.

First, let us delve into the comparative performance results of the models tested for predicting software categories using French descriptions of SaaS software, as depicted in Table 2. The metrics with the highest scores are emphasized in bold. Notably, the *MLP-OpenAI-Embed-V2 (fr)* model, trained solely on the French dataset with embeddings from OpenAI's *text-embedding-ada-002*, emerges as the top performer. Surprisingly, for this large-scale multi-class text classification task in French, the second most effective model was *MLP-CamemBERT (fr)*, leveraging embeddings from the open-source CamemBERT model. This suggests that language-specific, open-source embedding models can be viable and yield results comparable to those of commercial embedding models in certain application scenarios. Classifiers trained on the complete dataset (including both French and English texts) demonstrated similar performance levels to the best-performing models, with the latest version of OpenAI's embedding model (*text-embedding-3-large*) slightly outperforming its predecessor (*text-embedding-ada-002*). However, when assessing the transfer learning capability of these embeddings, such as the performance of *MLP-OpenAI-Embed-V2 (en)*, the classifier trained on English texts, a marked decline in effectiveness was noted. For instance, the F1-score dropped by 61% from 0.996 for *MLP-OpenAI-Embed-V2 (fr)* to 0.383 for *MLP-OpenAI-Embed-V2 (en)*.

Model	**Metric**				
	Accuracy@1	Accuracy@3	Precision	Recall	F1-score
MLP-CamemBERT (fr)	**0.998**	**1**	0.995	0.995	0.995
MLP-OpenAI-Embed-V2 (fr)	0.996	**1**	**0.996**	**0.996**	**0.996**
MLP-OpenAI-Embed-V2 (en)	0.369	0.637	0.577	0.368	0.383
MLP-OpenAI-Embed-V2 (fr-en)	0.993	0.999	0.994	0.993	0.993
MLP-OpenAI-Embed-V3 (fr-en)	0.994	0.999	0.995	0.994	0.994

TABLE 2. COMPARATIVE RESULTS OF DIFFERENT CLASSIFIERS ON THE FRENCH DATASET

The comparative performance results of the tested models on the English dataset are shown in Table 3 below. Contrary to the previous observations on the French dataset (cf. Table 2), the *MLP-OpenAI-Embed-V2 (en)* model, trained exclusively on the English dataset, ranked only third. It was outperformed by both the *MLP-OpenAI-Embed-V2 (fr-en)* and *MLP-OpenAI-Embed-V3 (fr-en)* models. These models were trained on a combination of English and French texts, potentially highlighting the superiority of a bilingual training approach in improving model performance. Furthermore, the classifier using OpenAI's latest *text-embedding-3-large* model achieved the best results, both in terms of prediction accuracy and cross-language transfer learning capability.

Model	**Metric**				
	Accuracy@1	Accuracy@3	Precision	Recall	F1-score
MLP-CamemBERT (fr)	0.247	0.372	0.356	0.246	0.252
MLP-OpenAI-Embed-V2 (fr)	0.224	0.369	0.358	0.224	0.232
MLP-OpenAI-Embed-V2 (en)	0.712	0.915	0.724	0.712	0.707
MLP-OpenAI-Embed-V2 (fr-en)	0.736	0.929	0.743	0.736	0.733
MLP-OpenAI-Embed-V3 (fr-en)	**0.743**	**0.933**	**0.749**	**0.743**	**0.741**

TABLE 3. COMPARATIVE RESULTS OF DIFFERENT CLASSIFIERS ON THE ENGLISH DATASET

The last evaluation was carried out on a test dataset containing four languages other than English and French, i.e., the two languages in which the training dataset was available. The aim here is to evaluate the transfer learning capability of the commercial text embedding models (cf. Section 3.3.2). The performance of the tested classifiers yielded remarkable results, as shown in Table 4. The ranking of the models according to their performance was as follows: the *MLP-OpenAI-Embed-V3 (fr-en)* model was the best, followed by *MLP-OpenAI-Embed-V2 (fr-en)*, *MLP-OpenAI-Embed-V2 (fr)*, *MLP-OpenAI-Embed-V2 (en)*, and finally *MLP-CamemBERT (fr)*.

For this multilingual, large-scale multi-class classification task involving 331 unique labels, the best-performing model, *MLP-OpenAI-Embed-V3 (fr-en)*, achieved an Accuracy@3 of 0.904 and an F1-score of 0.781. This level of accuracy is particularly remarkable given that the dataset consisted of multilingual data that these models had not previously encountered. The success of the *MLP-OpenAI-Embed-V3 (fr-en)* model in this setting demonstrates its superior ability to adapt to and accurately classify text in languages on which it has not been explicitly trained. Moreover, the fact that the models trained on bilingual data (French and English) outperformed those trained on monolingual data, even when applied to languages on which they were not trained, indicates the potential benefit of bilingual training in enhancing the flexibility and generalization capabilities of these models.

Figure 1 above provides a comprehensive view of the performance of different embedding models across six different languages. In the figure, the *MLP-OpenAI-Embed* models (both V2 and V3) refer to those trained on both English and French data. In addition, the open-source,

non-commercial *MLP-CamemBERT* model is also shown for easy comparison. Figure 1 suggests that commercial text embedding models, such as OpenAI's *text-embedding-ada-002* and *text-embedding-3-large*, perform reasonably well in multilingual text classification tasks. This is particularly noteworthy given that the constructed classifiers were not explicitly trained on data from the target languages.

Model	**Metric**				
	Accuracy@1	Accuracy@3	Precision	Recall	F1-score
MLP-CamemBERT (fr)	0.234	0.309	0.567	0.231	0.269
MLP-OpenAI-Embed-V2 (fr)	0.664	0.842	0.744	0.686	0.678
MLP-OpenAI-Embed-V2 (en)	0.445	0.702	0.591	0.492	0.473
MLP-OpenAI-Embed-V2 (fr-en)	0.692	0.868	0.747	0.741	0.712
MLP-OpenAI-Embed-V3 (fr-en)	**0.762**	**0.904**	**0.791**	**0.806**	**0.781**

TABLE 4. COMPARATIVE RESULTS OF DIFFERENT CLASSIFIERS ON COLLECTED MULTILINGUAL DATASET (CF. SECTION 3.3.2)

In summary, the results across all three evaluation settings—French, English, and multilingual datasets—consistently demonstrate the robustness of commercial embedding models, particularly when trained on bilingual data. For example, the *MLP-OpenAI-Embed-V3 (fr-en)* classifier demonstrated the highest overall performance, with an Accuracy@1 of 0.762 and an F1-score of 0.781 on multilingual test dataset, despite never having been trained on those languages. These findings substantiate the practical utility of commercial embedding models for multilingual SaaS classification and establish the foundation for further exploration of scalable and flexible approaches.

In the subsequent section, we address the key limitation of the current embedding-based classification approach, particularly regarding its adaptability in dynamic SaaS classification environments.

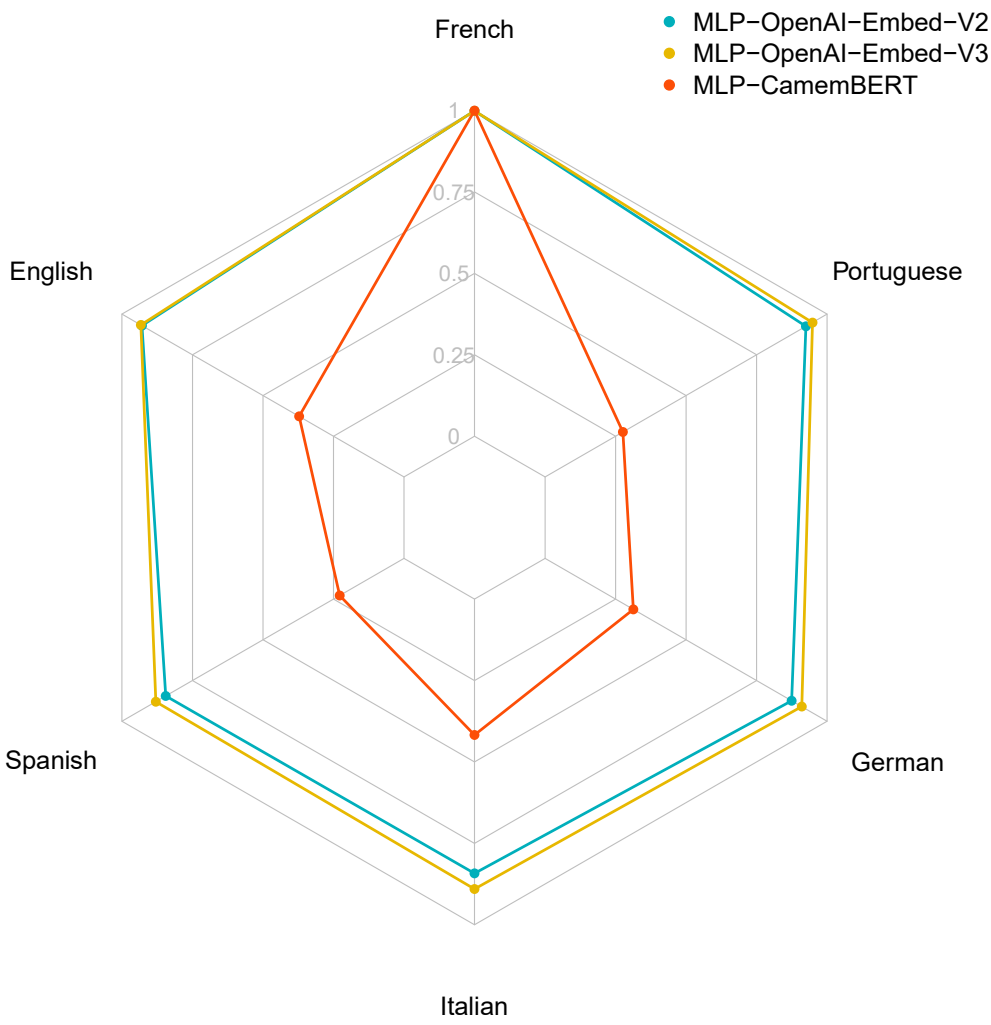

FIGURE 1. COMPARISON OF TEXT EMBEDDING MODELS IN CROSS-LANGUAGE TRANSFER LEARNING CAPABILITY. THE ACCURACY@3 METRICS ARE USED.

5. LIMITATION

While the presented approach demonstrated robust effectiveness in multilingual, large-scale, multi-class SaaS software classification, one key limitation must be acknowledged: the methodology is inherently inflexible because the classifiers must be retrained whenever the category catalog evolves. This is an inevitable scenario in the rapidly developing SaaS market. Consequently, the maintenance of classifier accuracy necessitates continuous, computationally intensive retraining, which results in substantial resource expenditure. The training of machine learning models, such as the MLP utilized in this study, necessitates significant computational power, which may not be cost-effective for smaller organizations.

To address this limitation, several alternative methods have been proposed to circumvent the necessity of extensive retraining. Among these methods, zero-shot classification has gained prominence as it primarily relies on vector similarity searches, thereby resulting in a significant reduction in both computational and temporal costs (Xian et al. 2016; Xie and Virtanen 2019). These approaches obviate the necessity for classifier retraining, thereby facilitating expeditious adaptability to alterations in the category catalog. Specifically, zero-shot classification models leverage similarity computation among the semantic vector representations (i.e., the embeddings) of both the input data and the labels (classes). Thereby, it enables the categorization of new classes without the necessity of explicit training data for those categories. For instance, Xian et al. (2016) demonstrated that latent embedding spaces could generalize well to unseen classes by relying solely on vector similarity between text embeddings and class-label embeddings. Similarly, Xie and Virtanen (2019) investigated zero-shot audio classification employing semantic embeddings, underscoring the adaptability and flexibility of these methodologies. These approaches obviate the necessity for classifier retraining, thereby facilitating expeditious adaptability to alterations in the category catalog, an essential feature for rapidly evolving SaaS markets.

Furthermore, the use of LLMs through prompting has emerged as a compelling and flexible solution for text classification tasks (Liu and Shi 2024; Brown et al. 2020). In this paradigm, categorization is achieved through the utilization of natural language prompts, which are provided to an LLM without the necessity for explicit model retraining. Unlike embedding-based classifiers or zero-shot approaches that rely on semantic vector representations, prompt-based classification operates end-to-end within the LLM itself. The classification decision is made based directly on the model's internal reasoning over the prompt, without computing or comparing external embeddings. LLM-based methods have demonstrated noteworthy efficacy in a range of NLP tasks, attributable to their capacity for generalization, which is derived from training on extensive and diverse textual datasets. Recent studies have highlighted that carefully crafted prompts can significantly enhance classification performance, rendering LLMs attractive for practical scenarios where retraining is infeasible (Brown et al. 2020; Gao et al. 2021). Liu and Shi (2024) specifically addressed the effectiveness of prompt engineering, demonstrating how domain-specific prompts can substantially improve the accuracy of classification tasks in specialized industries. However, the successful deployment of LLM-based classification necessitates meticulous prompt design that encapsulates the comprehensive context of the category catalog, potentially augmenting manual effort and necessitating domain-specific expertise.

Despite the encouraging nature of these approaches, it is important to note that each alternative is accompanied by a unique set of challenges. Zero-shot classification techniques, while efficient, heavily depend on the quality of semantic embeddings and might exhibit limited performance on highly specialized or nuanced categories due to insufficient semantic alignment. In a similar vein, classification methods that rely exclusively on prompting LLMs may incur augmented complexity and expense due to prompt optimization processes, the necessity for extensive contextualization, and potential inconsistencies emanating from LLM output variability.

6. Conclusions and Future works

Advances in NLP, and text embedding models in particular, have revolutionized the way applications are built across multiple domains (Devlin et al. 2019; Aytekin and Ayhan Erdem 2023; Greene et al. 2022). This article is devoted to providing significant insights into the use of commercial text embedding models for multilingual, large-scale, multi-class classification of SaaS solutions. Based on rigorous experiments on various sources of software description data in the SaaS marketplace community, the results of the article demonstrate the effectiveness of commercial embeddings in a real and challenging context.

The findings of the study are multifaceted. Notably, it was observed that while non-commercial text embedding models, such as Martin et al.'s (2020) pre-trained CamemBERT, show promise in monolingual settings, their performance declines significantly in cross-language scenarios. In contrast, commercial text embedding models, such as OpenAI's *text-embedding-ada-002* model (Greene et al. 2022), show robustness in both language-specific and cross-language classification tasks. In particular, the study showed that a bilingual approach of training classifiers using commercial text embeddings can effectively handle multilingual classification challenges. For example, classifiers trained with OpenAI's *text-embedding-3-large* model (OpenAI 2024) on French and English data showed high accuracy (0.90) and F1-score (0.78) even when applied to texts of languages not included in the training set. These results highlight the potential of commercial text embedding models to meet the complex classification needs of the SaaS industry.

The implications of this study are extensive, particularly for professionals in the SaaS sector and related fields, as well as for researchers and data scientists focusing on text embedding models for practical applications. The ability to accurately classify SaaS software in multilingual contexts offers concrete benefits for market analysis, targeted marketing, and customer support. These insights can support strategic decision-making for practitioners operating in the SaaS sector. These insights particularly pertain to the critical trade-offs that practitioners must navigate when deliberating the implementation of text embedding and classification methodologies. More specifically, these trade-offs center on the nexus between the computational cost of these methodologies and the adaptability of the classifiers they produce. The demonstrated efficacy of bilingual training with commercial embedding models underscores a pragmatic approach to addressing multilingualism challenges inherent in global SaaS platforms.

Nevertheless, ethical considerations related to transparency and fairness in model deployment should be an integral part of future discussions. Commercial embedding models, while powerful, often lack transparency regarding their training datasets, potentially embedding biases that may inadvertently propagate through automated classification decisions. Future studies should therefore address these ethical dimensions, proposing guidelines and transparency measures that help mitigate potential biases and ensure fair and equitable application across diverse linguistic and cultural contexts (Bender et al. 2021).

Moreover, this study highlights the importance of developing adaptable classification systems capable of efficiently handling dynamic category expansions. Future research should explore hybrid approaches, combining the computational efficiency and flexibility of zero-shot and/or LLM-based classification with traditional embedding-based classifiers. Such integrated methodologies could provide robust solutions capable of handling evolving SaaS categorization demands, ultimately contributing to more resilient and cost-effective NLP systems in real-world scenarios (Gao et al. 2021).

Looking ahead, the research opens up promising avenues, particularly in the area of zero-shot classification using a commercial text embedding model (Xian et al. 2016; Xie and Virtanen 2019). Indeed, the diversity and rapid evolution of the SaaS industry pose the challenge of flexibility requirements as software categories evolve. Zero-shot classification with embeddings has advantages in this respect, as there is no need to train classifiers; it is challenging but has great potential. Future studies could explore the extent to which commercial text embeddings can generalize to unseen categories and languages, thereby increasing the flexibility and utility of NLP models in diverse and evolving market scenarios.

References

Aravamuthan, Sarang, Prasad Jogalekar, and Jonghae Lee. 2022. "Extracting Features from Textual Data in Class Imbalance Problems." *Journal of Computer-Assisted Linguistic Research* 6 (November): 42–58. https://doi.org/10.4995/jclr.2022.18200.

Aytekin, Mustafa Ulvi, and O. Ayhan Erdem. 2023. "Generative Pre-Trained Transformer (GPT) Models for Irony Detection and Classification." *2023 4th International Informatics and Software Engineering Conference (IISEC)*, December, 1–8. https://doi.org/10.1109/IISEC59749.2023.10391005.

Balikas, Georgios. 2023. "Comparative Analysis of Open Source and Commercial Embedding Models for Question Answering." *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (New York, NY, USA), CIKM '23, October 21, 5232–33. https://doi.org/10.1145/3583780.3615994.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA), FAccT '21, March 1, 610–23. https://doi.org/10.1145/3442188.3445922.

Brown, Tom B., Benjamin Mann, Nick Ryder, et al. 2020. "Language Models Are Few-Shot Learners." *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA), NIPS '20, December 6, 1877–901.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, et al. 2020. "Unsupervised Cross-Lingual Representation Learning at Scale." *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, July, 8440–51. https://doi.org/10.18653/v1/2020.acl-main.747.

Dai, Haixing, Zhengliang Liu, Wenxiong Liao, et al. 2025. "AugGPT: Leveraging ChatGPT for Text Data Augmentation." *IEEE Transactions on Big Data* 11 (3): 907–18. https://doi.org/10.1109/TBDATA.2025.3536934.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, June, 4171–86. https://doi.org/10.18653/v1/N19-1423.

Du, Yu, Erwann Lavarec, and Colin Lalouette. 2023. "Text Data Augmentation to Manage Imbalanced Classification: Apply to BERT-Based Large Multiclass Classification for Product Sheets." *International Journal of Computational Linguistics (IJCL)* 14: 1–18.

Gao, Tianyu, Adam Fisch, and Danqi Chen. 2021. "Making Pre-Trained Language Models Better Few-Shot Learners." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.295.

Greene, Ryan, Ted Sanders, Lilian Weng, and Arvind Neelakantan. 2022. "New and Improved Embedding Model." December 15. https://openai.com/index/new-and-improved-embedding-model/.

Incitti, Francesca, Federico Urli, and Lauro Snidaro. 2023. "Beyond Word Embeddings: A Survey." *Information Fusion* 89: 418–36. https://doi.org/10.1016/j.inffus.2022.08.024.

Le, Quoc, and Tomas Mikolov. 2014. "Distributed Representations of Sentences and Documents." *Proceedings of the 31st International Conference on Machine Learning*, June 18, 1188–96. https://doi.org/10.48550/arXiv.1405.4053.

Licht, Hauke. 2023. "Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings." *Political Analysis* 31 (3): 366–79. https://doi.org/10.1017/pan.2022.29.

Liu, Menglin, and Ge Shi. 2024. "PoliPrompt: A High-Performance Cost-Effective LLM-Based Text Classification Framework for Political Science." SSRN Scholarly Paper 4940136. Social Science Research Network, August 19. https://papers.ssrn.com/abstract=4940136.

Liu, Yinhan, Myle Ott, Naman Goyal, et al. 2019. "Roberta: A Robustly Optimized Bert Pretraining Approach." *arXiv Preprint arXiv:1907.11692*, ahead of print. https://doi.org/10.48550/arXiv.1907.11692.

MarketsandMarkets. 2025. "Customer Information System Market Size, Share Report | Industry Analysis and Growth 2030." MarketsandMarkets. https://www.marketsandmarkets.com/Market-Reports/customer-information-system-market-183489801.html.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, et al. 2020. "CamemBERT: A Tasty French Language Model." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July, 7203–19. https://doi.org/10.18653/v1/2020.acl-main.645.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." arXiv:1301.3781. Preprint, arXiv, September 7. https://doi.org/10.48550/arXiv.1301.3781.

OpenAI. 2024. "New Embedding Models and API Updates." March 13. https://openai.com/index/new-embedding-models-and-api-updates/.

Patil, Rajvardhan, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. 2023. "A Survey of Text Representation and Embedding Techniques in NLP." *IEEE Access* 11: 36120–46. https://doi.org/10.1109/ACCESS.2023.3266377.

Pattun, Geeta, and Pradeep Kumar. 2023. "Emotion Classification Using Generative Pre-Trained Embedding and Machine Learning." *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, December, 1–6. https://doi.org/10.1109/ICMLANT59547.2023.10372980.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, et al. 2018. "Deep Contextualized Word Representations." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, edited by Marilyn Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202.

Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. "How Multilingual Is Multilingual BERT?" arXiv:1906.01502. Preprint, arXiv, June 4. https://doi.org/10.48550/arXiv.1906.01502.

R, Srinivasa Raghavan, Jayasimha K.R, and Rajendra V. Nargundkar. 2020. "Impact of Software as a Service (SaaS) on Software Acquisition Process." *Journal of Business & Industrial Marketing* 35 (4): 757–70. world. https://doi.org/10.1108/JBIM-12-2018-0382.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training." Preprint. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

Ravi, Jayasree, and Sushil Kulkarni. 2023. "Text Embedding Techniques for Efficient Clustering of Twitter Data." *Evolutionary Intelligence* 16 (5): 1667–77. https://doi.org/10.1007/s12065-023-00825-3.

Spärck Jones, Karen. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation* 28 (1): 11–21. https://doi.org/10.1108/eb026526.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2017. "Attention Is All You Need." *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA), NIPS'17, December 4, 6000–6010. https://doi.org/10.48550/arXiv.1706.03762.

Wang, Liang, Nan Yang, Xiaolong Huang, et al. 2024. "Text Embeddings by Weakly-Supervised Contrastive Pre-Training." arXiv:2212.03533. Preprint, arXiv, February 22. https://doi.org/10.48550/arXiv.2212.03533.

Xian, Yongqin, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. 2016. "Latent Embeddings for Zero-Shot Classification." arXiv:1603.08895. Preprint, arXiv, April 10. https://doi.org/10.48550/arXiv.1603.08895.

Xie, Huang, and Tuomas Virtanen. 2019. "Zero-Shot Audio Classification Based on Class Label Embeddings." arXiv:1905.01926. Preprint, arXiv, August 7. https://doi.org/10.48550/arXiv.1905.01926.

Zhang, Yin, Rong Jin, and Zhi-Hua Zhou. 2010. "Understanding Bag-of-Words Model: A Statistical Framework." *International Journal of Machine Learning and Cybernetics* 1 (1): 43–52. https://doi.org/10.1007/s13042-010-0001-0.