Leveraging Open Banking Data for SME Finance: Clustering and Forecasting SME Cashflows

Brandi Jess[a, b,*,1], **Pietro Alessandro Aluffi**[a,**,1], **Marya Bazzi**[a, b, d], **Matthew Arderne**[a], **Daniel Rodrigues**[a], **Kate Kennedy**[a] **and Martin Lotz**[b]

[a]SME Capital
[b]University of Warwick
[d]The Alan Turing Institute

Abstract. Small and Medium-sized Enterprises (SMEs) make up over 90% of businesses, yet access to sufficient financing options remains limited and challenging, necessitating innovative lending solutions that are better-tailored to SME data and needs. This paper proposes a novel approach to cashflow analysis, crucial for supporting SME financing mechanisms, such as cashflow lending. Leveraging Open Banking data, the proposed methodology integrates embedding and clustering techniques to set the basis for dynamic, real-time, and forward-looking assessment tools of SME cash positions. More broadly, our approach serves as an early warning system for potential financial distress and facilitates proactive interventions and informed data-driven decision-making for lenders. By employing sentence transformers for bank transaction description embedding and clustering for monthly transaction segmentation, early results uncover clusters of distinct spending behaviors and our forecasting approach outperforms initial baselines on empirical SME bank transaction data. Improving the categorisation and forecasting pipeline by adapting embedding models to the idiosyncrasies of SME bank transaction data and incorporating metadata for transaction data enrichment is ongoing work.

1 Introduction

In the complex ecosystem of the economy, Small and Medium-sized Enterprises (SMEs) play a pivotal role due to their significant contributions to employment, innovation, and GDP [13]. However, one of the challenges SMEs face is securing adequate financing, which is critical for their survival and growth. SMEs frequently encounter significant barriers to accessing traditional bank financing due to having less easily accessible and usable data relative to larger corporations, as well as more nuanced risk profiles (e.g., more volatile cashflows and business operations). This leads to stringent collateral requirements and risk-aversion from financial institutions, creating the need for alternative financing models better tailored to SMEs' data and needs [4].

Recent studies have advocated for innovative lending practices that leverage predictive analytics and data-driven decision-making to improve access to finance and enhance lending models [9]. Cashflow lending, for instance, allows SMEs to access finance based on cashflow rather than existing assets. This makes it more effective and affordable than collateral-based lending, particularly for SMEs with limited tangible assets [1].

The success of cashflow lending hinges on accurate and financial analysis of an SME's cashflows. The advent of Open Banking has revolutionised the access to financial data, providing an unprecedented opportunity to innovate how we analyse SME financial transactions. A report by Mastercard [20] shows that SME owners are willing to allow access to their company's bank transaction data on the basis that it decreases reliance on credit scores for loans. This suggests the potential for significant increase in Open Banking uptake and an eventual change in lending dynamics for SMEs.

Machine learning plays a pivotal role in this new landscape as it can provide the basis for powerful analytics tools, such as cashflow forecasting and bank transaction categorisation. In natural language processing, embeddings refer to learned representations of words or phrases as high-dimensional vectors, capturing semantic and syntactic relationships and meanings. Bank transaction data is commonly presented in a non-standardised, semantically ambiguous, noisy, and unstructured format, for example, a $25 subscription fee for OpenAI's ChatGPT may appear in various forms depending on the bank, such as `6789 01JAN24 CHATGPT SUBSCRIPTION SAN FRANCISCO US USD 25.00VRATE 1.2345N-S TRN FEE 0.50` or `OPENAI *CHATGPT SU US 25.00 VISAXR 1.2345 CD 6789`. Therefore, data pre-processing and embedding models can be crucial to infer the business context of a transaction and for effective analysis of bank transaction data, extracting valuable information and forming the basis for feature engineering in any subsequent machine learning pipeline. Furthermore, bank transaction data can exhibit high variance and a wide range of temporal patterns across different bank accounts of a given company, different companies, and different portfolios. Therefore, generalisability is a key consideration when developing techniques for analysing such datasets.

Accurate embeddings can significantly enhance financial analysis potential by identifying patterns and anomalies in transaction data, providing early warnings of financial distress and enabling proactive interventions [8]. Techniques such as character-level embeddings or subword embeddings can help capture the nuances of misspelled or abbreviated words prevalent in open banking data [24]. Leveraging detailed transaction embeddings allows financial institutions to bet-

[*] Corresponding Author. Email: brandi.jess@smecapital.com.
[**] Corresponding Author. Email: pietro.aluffi@smecapital.com.
[1] Equal contribution.

ter assess the creditworthiness of SMEs, reduces reliance on traditional credit scores and facilitates access to financing [28]. Embedding models can also contribute to the automation of categorisation and analysis of transaction data, minimising the need for manual intervention and reducing the risk of human error [16].

While the opportunities provided by embedding models for bank transactions are substantial, the development of such models involves addressing several challenges. Firstly, insufficient data and/or data labels from SMEs with limited transaction history and business idiosyncrasies can hinder the development and generalisability of embedding representations. Secondly, despite the increased access to bank transaction data, this data is often noisy and non-standardised. Bank transactions often include typographical errors and unconventional abbreviations or shorthand descriptions that vary in format, language, and detail [12], and often require labelling at a borrower level rather than any models able to generalise across lending portfolios. This makes standardisation and accurate interpretation difficult. Thirdly, the length of bank transaction descriptions is typically short, which limits contextual understanding. Unlike longer texts where the meaning and context can be inferred from surrounding text, bank transaction descriptions are often short and and need to be enriched with metadata to capture meaning. Not only is the word-level meaning important, the sentence, or description, level also provides valuable contextual information to bank transactions.

This work reviews the performance of different sentence-level embedding techniques applied to business bank transactions data within a broader machine learning pipeline. The sentence-level embeddings of bank transaction data can be used in various downstream tasks, including clustering, prediction, and change-point detection.

The focus of this work is to introduce a machine learning pipeline for cashflow forecasting that can aid the analysis of SME's financial behavior using open banking data, employing embedding models to extract meaningful information from the noisy, shorthand descriptions of bank transactions. We build on the work by Kotios et al. [15] and Toran et al. [30], and our key contributions are as follows. Firstly, our machine learning pipeline for clustering bank transactions is unsupervised and can thus be adapted without the need for quality labels to a wide variety of businesses and use-cases. Importantly, this approach may mitigate overfitting and generalise more effectively across different companies and portfolios. Secondly, we show that using cluster-level features from prior time-points can increase cashflow forecasting accuracy. Thirdly, we consider the idiosyncrasies of bank transactions data and highlight the importance of developing embedding models specifically tailored to this data.

We note that our contribution has applications beyond SME finance, in broader business-to-business (B2B) interactions, such as commercial insurance underwriting and supplier risk assessments. This work is also relevant in the business-to-consumer (B2C) lending space, where cashflow forecasting is often the only viable way to assess borrower risk.

2 Related work

Bank transaction data has grown considerably with the expansion of electronic banking [14]. The banking sector is well aware of the value of customer information covering demographics, leisure, wealth, insurance, financial transactions, and so on. This section reviews literature relevant to our study, focusing on embedding techniques for bank transactions. We explore methods for embedding bank transactions, which transform transactional data into a structured format amenable to machine learning algorithms. This review establishes the groundwork for our proposed methodology by delineating the current state of research and identifying gaps that our in-progress work aims to contribute to.

2.1 Financial data for SMEs

The classical approach to SME assessment uses financial ratio–based variables extracted from classical financial statements [17, 2]. The risk assessment of SMEs using bank transaction data has gained significant attention in recent years, highlighting the role of machine learning in enhancing financial oversight and risk management. Kou et al. [16] propose a bankruptcy prediction model for SMEs that uses transactional data to demonstrate its predictive capability and economic benefits. Similarly, Teng [29] discusses a financial risk monitoring system that utilises both structured and unstructured transaction data, enhancing the efficiency of customer information processing and fraud risk prediction. In the context of risk modelling, Startseva et al. [26] present algorithms for analysing textual labels in transaction data, enhancing the identification of high-risk transactions and improving the accuracy of financial monitoring systems.

2.2 Embedding bank transactions

The effectiveness of text analysis methods like bag-of-words (BOW) and word2vec is well-documented, and despite their simplicity, these techniques remain prevalent in the field [21]. However, the introduction of sentence embeddings marks a significant advancement, offering a more nuanced representation by capturing contextual relationships within sentences [23]. This development is particularly beneficial in the financial sector, improving tasks such as credit risk assessment, financial health analysis, and categorisation of spend by providing a deeper understanding of semantic meanings.

In the lending sector, the integration of textual features has been shown to improve the predictive power of credit risk models [32]. Transaction analysis provides key insights into user spending behavior and financial health monitoring. Nevertheless, challenges such as data sparsity, rare and missing words, misspellings, and unconventional abbreviations complicate the analysis of short-text bank transactions [11]. To mitigate some of these issues, researchers have explored various strategies, including the use of ontology and Wikipedia data to enrich datasets and reveal hidden topics, thus addressing data sparsity issues [22, 31, 10, 3]. Additionally, the robustness of FastText in handling misspelling errors has been particularly noted, with its application in analyzing bank transaction descriptions proving effective [30, 24]. To our knowledge, developing embedding approaches that explicitly mitigate the main idiosyncrasies prevalent in bank transaction data (e.g., unconventional abbreviations, shorthand descriptions that vary in format, language, and detail) is still an open research area.

2.3 Machine learning applications of embedded bank transactions

Recent research has also highlighted the effectiveness of using sentence embeddings within machine learning models to categorise personal bank transactions, as evidenced by recent studies [27, 11, 6, 8]. Although these methodologies can also be applicable to business transactions, unique challenges need to be addressed for business data due to the nature of the transactions. That is, SMEs will have widely different customers and suppliers depending on the businesses sector, as well as spending categories that are specific to business (e.g. advertising and consultants).

In response to these challenges, recent studies have started focusing specifically on business and SME transaction data. Begicheva and Zaytsev [5] use bank transaction embeddings to calculate macroeconomic indices, while Kou et al. [16] conduct bankruptcy prediction for SMEs. Both of these approaches demonstrate the value of using transaction data to analyse a business' financial position instead of relying on the traditional method of using financial statements and accounting data. Furthermore, Toran et al. [30] developed a model that uses bank transaction text embeddings to train a discriminative deep neural network classifier for bank transaction categorisation, achieving an average accuracy of 91%. In the context of their experiments, this represents a significant improvement over the labels provided by Plaid. Similarly, Kotios et al. [15] leveraged transaction embeddings for categorisation and extended this to cash flow forecasting within each category, finding that an XGBoost model delivered the highest accuracy. Both of these approaches, however, still rely on a fixed set of predefined categories and manual labeling, both of which often unavailable in practise.

We propose an unsupervised approach for categorising bank transactions. Such an approach avoids the need for manual labels (often unavailable or requiring significant manual resources from domain experts). It also mitigates overfitting, particularly important in this context as business transactions vary widely across different sectors and even across companies within a given sector.

3 Methods

In this section, we describe our approach for clustering and forecasting bank transaction data, and review a number of embedding techniques for short-hand bank transaction descriptions. Recall that bank transaction data includes different and inconsistent formats and descriptions, making interpretability and analysis challenging. One of our primary objectives is thus to develop a robust embedding approach that not only simplifies the representation of transaction descriptions but also preserves their semantic and syntactic integrity, in order to capture financial insight and inform decision-making.

3.1 Pre-processing

As mentioned above, when analysing bank transaction descriptions we face several challenges because of the nature of our dataset. For example, descriptions often include typographical errors and unconventional abbreviations or shorthand descriptions that vary in format, language, and detail, making standardisation and accurate interpretation difficult. As in Toran et al. [30], the initial step in our pipeline involves pre-processing the raw transaction data to ensure its quality and suitability for analysis. We start this process by obtaining transaction datasets that include essential attributes such as the date, time, description, and amount of each transaction. Given our focus on understanding expenditure patterns in the context of SME analysis, we specifically filter these datasets to retain only outgoing transactions. An important step of our pre-processing involves the application of NLP techniques to clean and standardise transaction descriptions. This standardisation is not only a linguistic correction but is aimed at identifying various similar descriptions and turning them into a standard format that can be aggregated. For example, slight variations in wording or abbreviations used across descriptions are standardised to ensure that transactions with similar purposes are recognised as such.

Once cleaned, the transaction descriptions are aggregated based on their standardised form and the corresponding month. This step allows us to compute description level features, such as the total, mean, maximum and minimum expenditure, associated with each type of transaction per month, offering a clear view of spending patterns over time. We currently use the data without any filtering criteria, but future work could include the removal of outliers (e.g., a large loan) or restrictions on time (e.g., starting the day after a loan is dispersed) to ensure the consistency and relevance of the data being analysed. The raw data included 20,171 transactions, and after post processing and aggregation we reduce it to 9,608 transactions. An example of this process is shown in Figure 1.

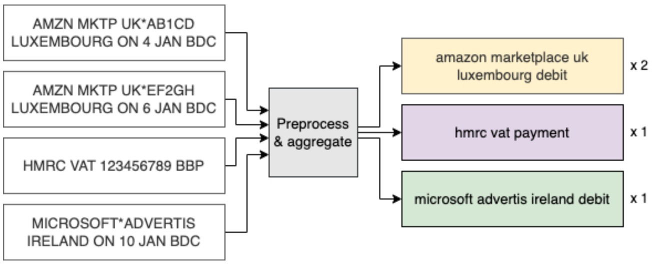
AMZN MKTP UK*AB1CD LUXEMBOURG ON 4 JAN BDC
AMZN MKTP UK*EF2GH LUXEMBOURG ON 6 JAN BDC
HMRC VAT 123456789 BBP
MICROSOFT*ADVERTIS IRELAND ON 10 JAN BDC
Preprocess & aggregate
amazon marketplace uk luxembourg debit
x 2
hmrc vat payment
x 1
microsoft advertis ireland debit
x 1

Figure 1. An example of the pre-processing and aggregation steps on raw bank transaction description text.

3.2 Embeddings

Our objective is to identify an embedding method that is able to represent bank transactions in an embedding space based on their descriptions. Business transactions often have diverse descriptions, making it challenging to track them accurately. By converting transaction descriptions into embeddings, similar transactions can be grouped together based on their semantic and syntactic similarity.

To generate high-quality embeddings for short bank transaction descriptions related to SME loans, we employ a comprehensive three-phase methodology. This methodology encompasses the evaluation of existing pre-trained embedding techniques, fine-tuning of selected models on our dataset, and the development of custom embedding models specifically tailored for this domain.

As a first step, we propose a comparison of several pre-trained sentence embedding models. Our benchmarking efforts include a range of widely used models, such as:

- **FastText** We consider a range of FastText models, encompassing those pre-trained on extensive corpora such as Common Crawl and Wikipedia.
- **BERT** Various BERT-based models have been adapted for sentence embeddings, including Sentence-BERT (SBERT) and SRoBERTa. Introduced by Reimers and Gurevych [25], the SBERT model generate sentence embeddings by fine-tuning BERT models specifically for semantic similarity tasks.
- **Mixedbread-AI** We include a recently introduced embedding model provided by Mixedbread-AI, termed mxbai-embed-large-v1, which has demonstrated effectiveness in capturing contextual information in textual data [19].
- **OpenAI Embeddings**: Leveraging the language models developed by OpenAI, this embedding technique utilises pre-trained models such as GPT-3 to generate rich, context-aware embeddings. These models are characterised by their understanding of language nuances, making them effective in capturing the subtle meanings embedded within complex transaction descriptions.

- **Sent2Vec**: Sent2Vec extends word2vec to sentences by learning sentence representations directly, rather than averaging word embeddings [23].

In our initial work, we compare the performance of the two proposed embedding models, SBERT and the pre-trained embeddings from FastText trained on Common Crawl. After these pre-trained embeddings are tried and compared, we plan to investigate enhancing the quality of the embeddings by fine-tuning or retraining the embedding models specifically on bank transaction description data.

To assess and compare the efficacy of these pre-trained embedding methods, we first analyse the singular value decay (SVD) of the generated embeddings. We then utilise dimensionality reduction techniques, such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA), to reduce the dimensions of the embeddings based on the SVD to reduce the noise and increase the clustering interpretability and to improve scalability, while preserving essential information.

3.3 Clustering and forecasting

The next step in our proposed pipeline is to apply a clustering algorithm to the reduced embeddings. A variety of clustering techniques can be tried here, including the traditional k-means and more recent clustering approaches, such as CLASSIX [7], and one can use variants of the Hungarian algorithm to track clusters over time [18]. For a given set of clusters, one can then forecast cashflow spend within each cluster over time.

Given that the clustering will group transactions with similar semantic meaning or structure together, such as expenses or revenue categories, our forecasting model uses historical spending within each cluster to predict future spend. This approach allows for a more dynamic and accurate forecast of cashflows, with the ability to identify recurrent patterns and temporal behaviors specific to each cluster. For example, recurring payments typically follow regular intervals (e.g., monthly or quarterly), whereas operational expenses can vary depending business type and sector. By embedding these transactions into clusters, the model uses patterns to improve the accuracy of future predictions. This clustering approach enables the model to distinguish between periodic and more irregular financial flows, thus informing the forecasting process with greater granularity.

4 Preliminary results

In the initial phase of our research, we explored two distinct embedding models, SBERT [25] and FastText pre-trained on the Common Crawl corpus, for representing and clustering bank transaction descriptions associated with SME loans. Our objective was to assess these models' ability to facilitate the understanding of SME transaction data effectively. Our dataset included SME transaction records, which we pre-processed following the methodology described above. This step was essential to ensure the consistency and relevance of our data for analysis. We then extracted embeddings for each cleaned transaction description using the SBERT and FastText models. SBERT, known for its deep contextual understanding due to the underlying BERT architecture, provides nuanced language embeddings. In contrast, FastText captures a broader range of linguistic features from its training on the extensive Common Crawl corpus. In this work, we use the default embedding dimension for FastText and SBERT, which are 300 and 384 respectively.

Singular value decomposition (SVD) is a matrix decomposition on which PCA is based. To analyse and reduce the complexity of our embeddings, we used PCA, a statistical procedure that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This step reduces the dimensionality by transforming the original variables into a new set of variables, which are easier to analyse and visualise. In both examples, the optimal number of components returned by a SVD of the data matrix was 2.

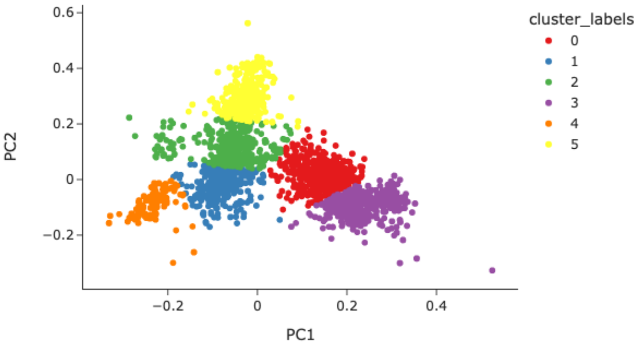

Figure 2. An example of the clusters from the reduced sentence embeddings of the FastText model pre-trained on Common Crawl.

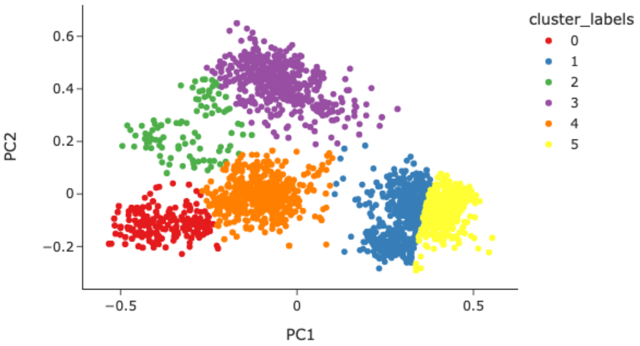

Figure 3. An example of the clusters from the reduced sentence embeddings of the SBERT model.

After reducing the dimensionality of our embeddings, we apply a k-means clustering algorithm with a predetermined choice of six clusters to partition the reduced embeddings into distinct groups based on similarity in feature space. Figure 2 and Figure 3 show scatter plots of the first two principal components with clusters indicated by color for the FastText and SBERT embedding models respectively. Without the presence of labelled data at this stage, one way to assess the embedding techniques is to visually examine the cluster plots to assess whether distinct transaction groups are identified. The cluster segmentation shown in these plots demonstrates how transactions with similar characteristics cluster together. Our preliminary results highlight the potential of embedding techniques like SBERT and FastText in transforming raw financial data into structured insights that can enhance decision-making processes for SMEs. However, without a labelled dataset it is more complicated to evaluate which underlying mechanism is picked up by the embedding techniques. By further refining these techniques (e.g., optimising the number of clusters and clustering approach) and expanding our dataset, we aim to develop more detailed and accurate methods for categorising and analysing SME transactions.

4.1 Forecasting

This section presents the preliminary results of our forecasting models, focusing on the efficacy of embedding techniques and clustering in improving cashflow forecasting accuracy for SMEs. We compare the performance of our baseline and machine learning forecasting models, highlighting the comparative results obtained from applying SBERT and FastText embeddings.

We apply the SBERT and FastText embedding models to an SME's transaction data, using the first two principal components to cluster our data, as described in the section above. Using the resulting data with cluster labels, we then developed two forecasting models: a baseline model and an XGBoost model. The baseline model simply uses the total spend per cluster in the previous month as our predictions for the next month, that is, the baseline assumes the total spend in any given cluster in any given month will remain the same for that cluster in the next month. In contrast, the XGBoost model employs a more sophisticated approach, including the cluster size (i.e., the number of unique embeddings/transaction descriptions) from the previous month and the total spend per cluster from the previous month, to forecast the total spend for the forthcoming month.

Clustering was performed on the reduced data using k-means, experimenting with a range of cluster numbers from 5 to 9. For both SBERT and FastText embedded data, cluster labels were generated using the same process as defined in the previous section, and the performance of the XGBoost model was compared against the baseline model across these different cluster configurations. A simple grid search was used to tune the parameters (maximum depth, learning rate, number of estimators, and subsample) of the XGBoost model, and the model with the tuned parameters was run over 10 different random seeds to ensure robustness and stability of the results. The primary metric for comparison was the Root Mean Squared Error (RMSE) of the predicted total spend versus the actual total spend for the next month per cluster. The average reduction in RMSE was calculated to assess the efficacy of the XGBoost model relative to the baseline.

The results, shown in Table 1, demonstrate that the XGBoost model outperforms our baseline model. For SBERT embeddings, the XGBoost model achieved an average RMSE reduction of 9.87% across the various cluster sizes (5 to 9). Similarly, for FastText embeddings, the XGBoost model demonstrated an average RMSE reduction of 11.86% compared to the baseline model. These percentage reductions in RMSE indicate a notable improvement over the simplistic baseline model. They suggest that the integration of more sophisticated feature engineering and machine learning techniques, such as those leveraging transaction embeddings and clustering, holds promise for enhancing the accuracy of cashflow forecasting for SMEs. The improvements observed, particularly with FastText embeddings, underscore the potential benefits of further refinement and tuning of these models.

Table 1. Baseline vs. XGBoost forecasting results for the clustered data from both SBERT and FastText embeddings. Results are given as the average RMSE and average standard deviation across iterations, rounded to the nearest 1000 GBP.

SBERT		
Clusters	Baseline	XGBoost
5	109000 (332000)	100000 (121000)
6	92000 (307000)	83000 (109000)
7	79000 (286000)	71000 (113000)
8	70000 (269000)	63000 (110000)
9	63000 (254000)	56000 (99000)

FastText		
Clusters	Baseline	XGBoost
5	114000 (336000)	100000 (121000)
6	94000 (309000)	82000 (106000)
7	81000 (288000)	72000 (113000)
8	71000 (271000)	65000 (100000)
9	65000 (255000)	56000 (93000)

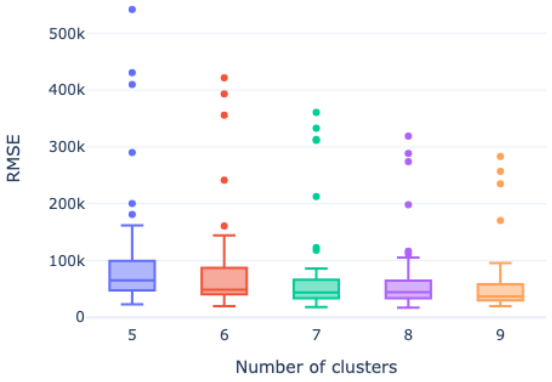

Figure 4. Boxplot showing the RMSE results by iteration for the FastText model pre-trained on Common Crawl.

Figures 4 and 5 show boxplots of the average RMSE values obtained from the XGBoost model for each random seed across various cluster configurations (5 to 9 clusters) using FastText and SBERT embeddings respectively. As the number of clusters increases, both the median and mean RMSE values generally decrease. This trend aligns with expectations as distributing the total spend across a greater number of clusters will inevitably lead to a smaller average spend per cluster. For both experiments, XGBoost consistently outperforms the baseline model across all cluster configurations. The standard deviations of the RMSE values are lower for the XGBoost model compared to the baseline in both embedding techniques. This indicates that the XGBoost model not only improves accuracy but also enhances the stability and reliability of the predictions.

These preliminary results demonstrate that leveraging embedding techniques and more advanced machine learning models, such as XGBoost, can improve cashflow forecasting accuracy for SMEs. The observed reductions in RMSE highlight the potential for further investigation and model refinement, in particular the embedding approach. Additionally, multi-output regression models could be applied in the forecasting pipeline to simultaneously predict total spend across all clusters, which accounts for the correlation across cluster predictions, rather than implementing separate independent models for each cluster.

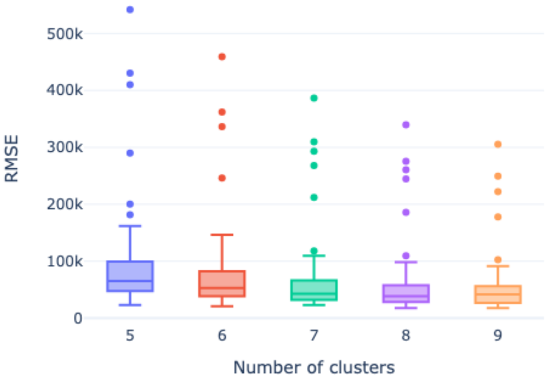

Figure 5. Boxplot showing the RMSE results by iteration for the SBERT model.

5 Conclusion

In this work, we present an approach for clustering and forecasting SME cashflow time series using bank transaction data. The ability to effectively leverage SME bank transaction data has a wide range of financial applications, however, this data presents a number of idiosyncrasies that one must cater to. These include short sentence lengths, lack of standardisation, unconventional abbreviations, and shorthand descriptions that vary in format, language, and detail.

We describe a pre-processing pipeline that one can apply to increase the standardisation and initial aggregation of bank transaction data. We then consider several approaches to embed and cluster the resulting dataset. Preliminary forecasting results obtained with XGBoost and pre-trained embedding approaches using FastText and SBERT outperform our baseline model. Our ongoing work consists of two key steps. The first is to build on our discussion in Section 3.2 and to identify an embedding approach that best caters to the idiosyncrasies of bank transaction data. The second is to consider a wider range of feature (e.g., metadata enrichment) and cross-correlations as part of the machine pipeline to gain deeper insight into the repeated patters and anomalies that underpin SME cashflows.

Our proposed machine learning pipeline for cashflow forecasting can serve as a valuable monitoring and underwriting tool for lenders to understand the SMEs' financial health and operational strategies. Such a tool can identify potential financial risks or opportunities early, enabling better investment choices and/or more proactive management of existing investments.

6 Future work

While our current findings demonstrate the potential of clustering and embedding techniques for improving cashflow forecasting accuracy, the scope of this study is limited by the subset of data used. Specifically, we rely on a subset of our dataset representing a limited portfolio view of SMEs, which may not fully capture the diverse business types and sectors across different industries. Exploring the generalizability of our results is an important next steps, as cashflow patterns can vary significantly depending on the operational sector, scale, and region of businesses. Recall that our use of unsupervised techniques was adopted with this inherent variance in mind, as it may mitigate overfitting and generalise more effectively across different companies and portfolios.

Seasonality is an important factor that we initially captured through a time-dependent clustering approach. Many SMEs experience cyclical financial patterns driven by season-specific business cycles or periodic events. By grouping transactions into clusters, our goal is to identify and account for these recurring patterns, which are important for cashflow monitoring. Such an approach allows the model to detect seasonal trends and adjust forecasts accordingly, enhancing its robustness when applied to businesses with strong seasonal dynamics. In future work, we will include longer timeframes and a broader variety of SMEs across different industry sectors, as well as explore further clustering and machine learning methods to account for seasonality and more irregular or unexpected temporal variation.

Additionally, we will work with domain experts to generate ground truth labels for a diverse set of bank transactions. This will offer a valuable additional evaluation mechanism as we enhance our methodology to both cluster and forecast cashflow time series, as well as help with user interpretability across different notable spend categories within a business.

References

[1] B. Amoako-Adu and J. Eshun. SME Financing in Africa: Collateral Lending vs Cash Flow Lending. International Journal of Economics and Finance, 10(6):p151, May 2018. ISSN 1916-971X. doi: 10.5539/ijef.v10n6p151. URL https://ccsenet.org/journal/index.php/ijef/article/view/74436. Number: 6.

[2] S. Balcaen and H. Ooghe. 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. The British Accounting Review, 38(1):63–93, 2006. Publisher: Elsevier.

[3] S. Banerjee, K. Ramanathan, and A. Gupta. Clustering short texts using wikipedia. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 787–788, 2007.

[4] T. Beck and A. Demirguc-Kunt. Small and medium-size enterprises: Access to finance as a growth constraint. Journal of Banking & Finance, 30(11):2931–2943, Nov. 2006. ISSN 0378-4266. doi: 10.1016/j.jbankfin.2006.05.009. URL https://www.sciencedirect.com/science/article/pii/S0378426606000926.

[5] M. Begicheva and A. Zaytsev. Bank transactions embeddings help to uncover current macroeconomics, Dec. 2021. URL http://arxiv.org/abs/2110.12000. arXiv:2110.12000 [cs, q-fin].

[6] C. B. Bruss, A. Khazane, J. Rider, R. Serpe, A. Gogoglou, and K. E. Hines. DeepTrax: Embedding Graphs of Financial Transactions, July 2019. URL http://arxiv.org/abs/1907.07225. arXiv:1907.07225 [cs, stat].

[7] X. Chen and S. Güttel. Fast and explainable clustering based on sorting, Feb. 2024. URL https://arxiv.org/abs/2202.01456.

[8] A. B. Dayioglugil and Y. S. Akgul. Continuous Embedding Spaces for Bank Transaction Data. In M. Kryszkiewicz, A. Appice, D. Ślęzak, H. Rybinski, A. Skowron, and Z. W. Raś, editors, Foundations of Intelligent Systems, pages 129–135, Cham, 2017. Springer International Publishing. ISBN 978-3-319-60438-1.

[9] C. Dhruv, D. Paul, M. H. Kumar, A. K. M, and M. S. Reddy. Framework for Bank Loan Re-Payment Prediction and Income Prediction. In 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC), pages 833–840, Jalandhar, India, May 2023. IEEE. ISBN 9798350300710. doi: 10.1109/ICSCCC58608.2023.10176363. URL https://ieeexplore.ieee.org/document/10176363/.

[10] S. Fodeh, B. Punch, and P.-N. Tan. On ontology-driven document clustering using core semantic features. Knowledge and information systems, 28:395–421, 2011. Publisher: Springer.

[11] S. García-Méndez, M. Fernández-Gavilanes, J. Juncal-Martínez, F. J. González-Castaño, and O. B. Seara. Identifying Banking Transaction Descriptions via Support Vector Machine Short-Text Classification Based on a Specialized Labelled Corpus. IEEE Access, 8:61642–

61655, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2983584. URL https://ieeexplore.ieee.org/document/9047935/.

[12] S. García-Méndez, F. d. Arriba-Pérez, O. Barba-Seara, M. Fernández-Gavilanes, and F. J. González-Castaño. Demographic Market Segmentation on Short Banking Movement Descriptions Applying Natural Language Processing. In 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), pages 141–146, 2021. doi: 10.1109/ISCSIC54682.2021.00035.

[13] S. C. Gherghina, M. A. Botezatu, A. Hosszu, and L. N. Simionescu. Small and Medium-Sized Enterprises (SMEs): The Engine of Economic Growth through Investments and Innovation. Sustainability, 12(1):347, Jan. 2020. ISSN 2071-1050. doi: 10.3390/su12010347. URL https://www.mdpi.com/2071-1050/12/1/347. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[14] A. M. Hormozi and S. Giles. Data Mining: A Competitive Weapon for Banking and Retail Industries. Information Systems Management, 21(2):62–71, Mar. 2004. ISSN 1058-0530. doi: 10.1201/1078/44118.21.2.20040301/80423.9. URL https://doi.org/10.1201/1078/44118.21.2.20040301/80423.9.

[15] D. Kotios, G. Makridis, G. Fatouros, and D. Kyriazis. Deep learning enhancing banking services: a hybrid transaction classification and cash flow prediction approach. Journal of Big Data, 9(1):100, Oct. 2022. ISSN 2196-1115. doi: 10.1186/s40537-022-00651-x. URL https://doi.org/10.1186/s40537-022-00651-x.

[16] G. Kou, Y. Xu, Y. Peng, F. Shen, Y. Chen, K.-S. Chang, and S. Kou. Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decis. Support Syst., 140:113429, 2021. doi: 10.1016/j.dss.2020.113429.

[17] M. Kregar. Cash flow based bankruptcy risk and stock returns in the US computer and electronics industry. The University of Manchester (United Kingdom), 2011.

[18] A. Kume and S. G. Walker. The utility of clusters and a Hungarian clustering algorithm. PLOS ONE, 16(8):e0255174, Aug. 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0255174. URL https://dx.plos.org/10.1371/journal.pone.0255174.

[19] X. Li and J. Li. AnglE-optimized Text Embeddings, Nov. 2023. URL http://arxiv.org/abs/2309.12871. arXiv:2309.12871 [cs].

[20] Mastercard. Empowering Small and Mid-sized Business Growth by Unleashing Its Data. Technical report, Mastercard, 2022. URL https://www.mastercard.us/content/dam/public/mastercardcom/na/us/en/large-enterprises/other/empowering-smb-study.pdf. Published: Mastercard.

[21] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space, 2013. _eprint: 1301.3781.

[22] M. A. Osman, S. A. M. Noah, and S. Saad. Ontology-based knowledge management tools for knowledge sharing in organization—a review. IEEE access, 10:43267–43283, 2022. Publisher: IEEE.

[23] M. Pagliardini, P. Gupta, and M. Jaggi. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In M. Walker, H. Ji, and A. Stent, editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 528–540, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1049. URL https://aclanthology.org/N18-1049.

[24] A. Piktus, N. B. Edizel, P. Bojanowski, E. Grave, R. Ferreira, and F. Silvestri. Misspelling Oblivious Word Embeddings. In J. Burstein, C. Doran, and T. Solorio, editors, 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), pages 3226–3234, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1326. URL https://aclanthology.org/N19-1326.

[25] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Aug. 2019. URL http://arxiv.org/abs/1908.10084. arXiv:1908.10084 [cs].

[26] A. Startseva, A. Vulfin, V. Vasilyev, A. Nikonov, and A. Kirillova. Analysis of Financial Payments Text Labels in the Dynamic Client Profile Construction. 2020 International Conference on Information Technology and Nanotechnology (ITNT), pages 1–10, 2020. doi: 10.1109/ITNT49337.2020.9253280.

[27] R. Sukumaran. Improved Customer Transaction Classification using Semi-Supervised Knowledge Distillation, Feb. 2021. URL http://arxiv.org/abs/2102.07635. arXiv:2102.07635 [cs].

[28] Z. Tao, W.-g. Zhang, W. Xu, and H. Hao. Multiple instance learning for credit risk assessment with transaction data. Knowl. Based Syst., 161: 65–77, 2018. doi: 10.1016/j.knosys.2018.07.030.

[29] T. Teng. Research and Application of Computer Internet Technology in Intelligent Financial Risk Monitoring System. 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), pages 699–702, 2021. doi: 10.1109/TOCS53301.2021.9688750.

[30] L. Toran, C. Van Der Walt, A. Sammarone, and A. Keller. Scalable and Weakly Supervised Bank Transaction Classification, June 2023. URL http://arxiv.org/abs/2305.18430. arXiv:2305.18430 [cs].

[31] T. N. T. Zakaria, M. J. A. Aziz, M. R. Mokhtar, and S. Darus. Text Clustering for Reducing Semantic Information in Malay Semantic Representation. Asia-Pac. J. Inf. Technol. Multimed, 9:11–24, 2020.

[32] W. Zhang, C. Wang, Y. Zhang, and J. Wang. Credit risk evaluation model with textual features from loan descriptions for P2P lending. Electronic Commerce Research and Applications, 42:100989, July 2020. ISSN 15674223. doi: 10.1016/j.elerap.2020.100989. URL https://linkinghub.elsevier.com/retrieve/pii/S1567422320300661.