LLMs for the categorisation of SME bank transactions

Brandi Jess*
SME Capital; The University of Warwick
brandi.jess@warwick.ac.uk

Pietro Alessandro Aluffi*
SME Capital; The University of Warwick
pietro.aluffi@smecapital.com

Marya Bazzi
SME Capital; The University of Warwick; The Alan Turing Institute
marya.bazzi@warwick.ac.uk

Matt Arderne
SME Capital
matthew.arderne@smecapital.com

Daniel Rodrigues
SME Capital
daniel.rodrigues@smecapital.com

Kate Kennedy
SME Capital
kate.kennedy@smecapital.com

Martin Lotz
The University of Warwick
martin.lotz@warwick.ac.uk

Abstract

This study investigates the application of Large Language Models (LLMs) for automating the categorization of bank transactions from Small and Medium-sized Enterprises (SMEs) in the manufacturing sector. Categorized bank transaction data provides real-time insights into an SME's financial health, yet automating this process is still a significant challenge due to the lack of standardization, sparse and ambiguous metadata, and the wide breath of potential transaction types.

We analysed a dataset of 14,799 transactions, labeled into 24 categories, and achieved an accuracy of 80% for LLM-generated labels when compared to human annotations by domain experts. Predictive models using transaction embeddings were developed, with performance evaluated through various train-test splits, including a leave-one-company-out approach. The results highlight the LLMs ability to categorise frequent transaction types but struggle with low-frequency and ambiguous categories. Our findings highlight the potential of LLMs to improve the efficiency and scalability of financial data processing, enhancing access to financial services for SMEs.

1 Introduction

Small and Medium-sized Enterprises (SMEs) are vital to the global economy, representing a majority of businesses worldwide and playing a key role in economic growth, employment, and innovation. Despite their importance, SMEs face challenges when securing financial resources, often being perceived as high-risk due to their inconsistent cash flows, limited financial histories, and non-standardised financial transactions. This creates a barrier to obtaining financing, which is essential for the growth and sustainability of these businesses, highlighting the need for an efficient approach to assessing SME's financial data.

The advent of open banking has the potential to reshape the financial landscape due to the increased access to bank transaction data. Although this data often presents challenges due to its ambiguous, inconsistent, or incomplete transaction descriptions, it also presents opportunities for innovation in assessing the financial health of SMEs. Access to such data has enabled the development of advanced machine learning (ML) and natural language processing (NLP) techniques, which have shown promise in addressing some of the traditional barriers faced by SMEs. These techniques facilitate the categorisation of bank transactions, cash flow prediction, and the broader assessment of financial viability.

However, despite the success, these approaches often depend on manually labeled data, which is time-consuming to generate and difficult to scale [4].

Weakly supervised techniques have emerged as a potential solution to reduce the burden of manual labeling; however, this approach still faces limitations, such as relying on the development of domain-specific heuristics (e.g., using specific rules to categorize recurring transactions like rent or taxes) and the difficulty in appropriately handling complex or ambiguous transaction descriptions (e.g., "office supplies" could be categorized under inventory, utilities, or general expenses depending on industry and business model of the company). Furthermore, these models often suffer from overfitting due to the wide variety of bank transactions and category variation depending on the context (e.g., the nature of the business—what's considered "suppliers" for a manufacturing firm may differ for a consulting company), necessitating further manual labeling over time.

To mitigate these challenges, we explore the use of Large Language Models (LLMs) to automatically generate labels for SME bank transactions. LLMs, which are pre-trained on vast datasets, have demonstrated abilities in understanding and generating text, making them well-suited to capture the nuanced meanings in transaction descriptions, even when those descriptions are noisy or incomplete [10]. This is particularly important in the financial domain, where transaction data can vary widely in format and detail. LLM-generated labels can provide categorisation at scale, allowing for the rapid and automated processing of large datasets without the need for manual intervention [13]. This scalability is crucial for financial institutions dealing with high volumes of transaction data on a daily basis.

Our research builds on previous work in transaction classification by incorporating LLMs into the label generation process, moving beyond the limitations of weak supervision models and manual labeling [24]. Our system is designed to handle a wide variety of transaction types, making it adaptable to different financial environments. Much of the existing literature on transaction classification focuses on retail or personal banking datasets [23] [16]

*Both authors contributed equally to this research.

[6], which tend to be more homogeneous, cleaner and more predictable, while our research leverages a diverse proprietary dataset from a portfolio of SMEs. This dataset contains a wide variety of transaction types, with challenges unique to SMEs, including variable transaction volumes, sector-specific behaviors, and short hand annotations. By improving the automated labeling processes, financial institutions can improve risk assessments, and develop tailored financial products for SMEs.

In the following sections, we review related work on text data labeling, embedding bank transactions, and LLMs in finance. We then outline our methodology, including the dataset, pre-processing, LLM-based labeling, and model evaluation. After presenting preliminary results on labeling performance and predictive modeling, we conclude with a discussion of future directions and key findings.

2 Related Work

Labeling bank transactions is crucial for analysing financial behaviors, assessing credit risk, and improving financial health monitoring [30]. Traditional approaches rely on manual labeling by experts, which, despite high accuracy, proved resource-intensive and unscalable, especially given the diverse and volatile financial activities of SMEs. Early ML models introduced rule-based systems and supervised learning to automate transaction labeling, yet these methods demanded extensive labeled datasets and struggled with the unstructured and inconsistent nature of SME transaction descriptions [22]. However, the reliance on labeled data remains a significant limitation for achieving scalable, generalisable, and real-time analysis [14]. Recent developments in NLP and LLMs have been adopted to enhance the automation of transaction labeling [8]. These methods provide more context-aware interpretations of transaction data, reducing the dependency on manual interventions and enhancing scalability [11]. The following sections explore the current state of labeling bank transactions, the role of embeddings in improving classification, and the introduction of LLMs in FinTech applications, particularly for handling the unique challenges posed by SME financial data.

2.1 Labeling Text Data

Transaction labeling has relied on manual annotations performed by domain experts. Despite being highly accurate this process demands significant time and expertise, especially when processing large volumes of transaction data, such as those of SMEs. Furthermore, manual labeling presents scalability challenges, particularly when encountering complex or non-standardised short hand transaction descriptions. These challenges often result in bottlenecks, limiting the speed and efficiency of financial data analysis [24]. In fact, developing large-scale labeled datasets manually can be prohibitively expensive and time-consuming, making it impractical for dynamic environments [26]. Weak supervision has emerged as an alternative to address the limitations of manual annotation, enabling the creation of labeled datasets without ground truth labels. Weak supervision leverages domain-specific heuristics, user-defined rules, or semi-supervised models to generate probabilistic labels from noisy or incomplete data sources [21]. This approach significantly reduces the time and effort required for labeling, improving scalability in handling large datasets [3]. Weak supervision techniques, like data programming, have proven effective in automatically labeling datasets programmatically, relying on labeling functions and heuristics rather than hand-labeling each data point [12]. However, the use of rules and heuristics to programmatically label dataset, is also one of the limitation of weak labels. In particular such dependence on predefined heuristics or labeling functions, while reducing manual effort, may be difficult to scale and generalise across diverse and complex datasets [25], especially in SME businesses where transaction descriptions vary widely, and the interpretation of the same transaction can differ significantly depending on the context (e.g., company's operations, nature of its suppliers and customers, etc). Furthermore, these heuristics may not be robust enough to handle ambiguous or noisy transaction data, necessitating continuous refinement and domain-specific tuning to maintain accuracy [9]. While weak supervision has proven valuable for scaling transaction labeling, it still faces hurdles in generalising across diverse datasets and dealing with noisy or ambiguous texts. These limitations highlight the ongoing need for improved techniques, such as leveraging LLMs and advanced ML algorithms, to better handle the complexities of modern financial transaction data.

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 [17]. However, the introduction of sentence embeddings marks a significant advancement, offering a more nuanced representation by capturing contextual relationships within sentences [19]. 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 [30]. 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 [6]. 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 [2, 5, 18, 29]. Additionally, the robustness of FastText in handling misspelling errors has been particularly noted, with its application in analysing bank transaction descriptions proving effective [20, 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 Large Language Models in Finance

The integration of LLMs into the financial sector represents an innovative shift in how financial transactions are labeled and analysed. As introduced above, transaction labeling relied on manual annotation or weakly supervised methods, both of which required domain expertise and significant time investment. These approaches lack

scalability and generalisability, particularly in handling the complexity and variety of financial transaction descriptions [15, 24]. Recent advancements in the development of LLMs offer an alternative or complimentary approach. LLMs can autonomously generate high-quality labels for transaction data by leveraging their context-aware understanding of incomplete or inconsistent descriptions [7]. This reduces the need for domain-specific heuristics and enables scalable, real-time transaction labeling systems. Notably, models like FinGPT and BloombergGPT have been developed specifically for the financial sector, demonstrating superior performance in financial natural language tasks such as sentiment analysis and transaction categorisation [27, 28]. The ability of LLMs to understand and process sector-specific language has further enhanced their utility in the Fintech space. For instance, in applications where transaction descriptions are often non-standardised and vary significantly, LLMs have demonstrated the capacity to handle this complexity through pre-training on diverse datasets [1]. Despite their potential, LLMs face several limitations when applied in the financial sector. One of the most significant challenges is the unstructured and noisy nature of financial data. LLMs often struggle to interpret incomplete or inconsistent data, which is common in short hand financial transactions or reports.

3 Methods

In this section, we outline our proposed approach to categorise SME bank transaction data. We begin by describing the dataset and the pre-processing steps undertaken to prepare the data for analysis. We then present an LLM-based labeling model and discuss the evaluation metrics applied to assess model performance.

3.1 Dataset

The dataset consists of transactional data from three companies, all operating within the manufacturing sector. Each dataset contains attributes such as date, time, transaction description, and amount. Transactions are labelled into 24 categories as shown in Table 1.

These categories capture a wide range of financial activities relevant to the companies' operations.

Our goal was to assign labels to the transactions using both manual (human) and automated (LLM) approaches, followed by the development of a ML model capable of predicting the labels based on transaction embeddings.

3.2 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. [24], 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. 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 for the three selected companies included 14,799 transactions, and after post processing and aggregation we reduce this to 1,720 transactions.

Table 1: The categories used for the bank transaction data labels.

ATM Withdrawals
Charges / Fees
Cheques
Credit Cards
Debt / Loan Repayments
Directors Loans
Inventory
Interest
Intra-company Transfers
Loan Inflows
Marketing / Advertising
Other Income
Other Outgoing
Payroll / Consultants
Refunds
Rent
Revenue
Sundries
Software / IT
Suppliers
Tax
Travel
Unpaid
Utilities

3.3 LLM-based labelling

For each transaction, the model was given the cleaned and standardised description as input and tasked with predicting one of the 24 categories in Table 1. We utilised OpenAI's GPT-4 Mini for generating the LLM labels. The performance of the LLM was evaluated by comparing its labels against the human-labelled dataset.

3.4 Label prediction model performance and evaluation

Once the transactions were labelled using an LLM, we explored the feasibility of building a predictive model using the transaction embeddings. Transaction embeddings, generated using a transformer-based architecture, were used as input features for the classifier. We experimented with various train-test splits to assess the generalisability of the model. These splits included a random 80/20 split across the selected three companies and a leave-one-company-out split. For both splits, we used accuracy as the primary evaluation metric, and explored the use of other metrics such as F1-score to handle the imbalance across transaction categories.

4 Preliminary results

This section presents preliminary findings on the performance of the labeling methodologies and predictive models. The accuracy of labels generated by the LLM is evaluated against domain expert annotations, highlighting strengths and weaknesses across categories. The effectiveness of the embedding-based predictive model is assessed using various training and testing strategies. Additionally, an error analysis identifies common challenges, particularly with ambiguous transaction descriptions and variations across companies.

4.1 LLM labelling performance

When comparing the LLM-generated labels to the human-annotated ground truth, we observed that the LLM achieved an accuracy of 80%. The model performed well in categories such as Tax, Payroll / Consultants, and Suppliers, where transaction descriptions were relatively consistent. However, its performance dropped in categories with more ambiguous or vague descriptions, such as Sundries and Other Outgoing, where additional context or domain-specific knowledge was required to assign the correct label.

Figure 1 shows that for the categories Other Income, ATM Withdrawals, Cheques, and Debt / Loan Repayments the accuracy compared to manual labels was 0% due to the low frequency of these transactions, whereas more common categories, such as Charges / Fees and Tax saw an accuracy near 100%.

4.2 Embedding-based predictive model

Following the label generation from the LLM, an investigation was conducted to determine whether a predictive model could effectively assign labels to a larger dataset without relying on LLMs. This approach aims to enhance scalability and efficiency in the labeling process by using transaction embeddings and reducing the reliance on LLMs for data security reasons.

We evaluated several models, including logistic regression for its interpretability, XGBoost for its robustness against overfitting, and deep neural networks (DNN) for their ability to capture complex relationships in the data. The performance of these predictive models is evaluated through different training and testing strategies.

4.2.1 Random 80/20 split across all companies. When training the labelling model using transaction embeddings on a randomly selected 80% of the data and testing on the remaining 20%, we achieved

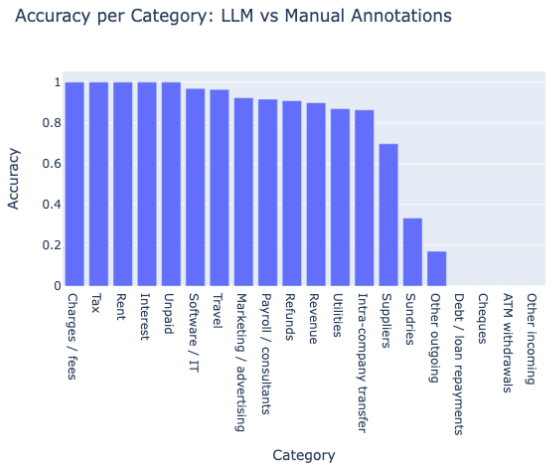

Figure 1: Accuracy of the LLM v manual annotations for each category.

an accuracy of 79%. The results of the three tested models is illustrated in Figure 2. This result demonstrates the model's ability to

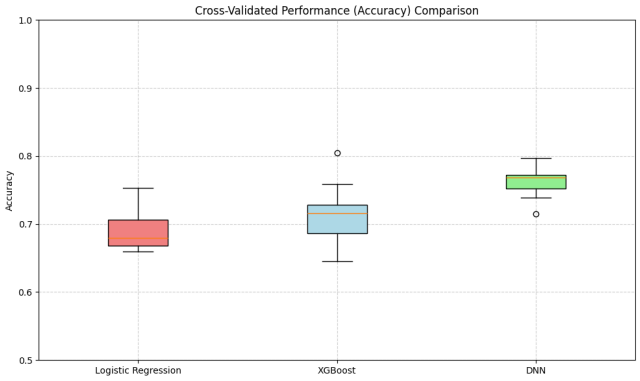

Figure 2: A box plot showing the accuracy of the embedding-based predictive models, including logistic regression, xgboost, and DNN, for labelling transactions using a random 80/20 test/train split.

reasonably predict the label of unseen transactions within the same company or across companies when the training data is representative.

4.2.2 Leave-one-company-out split. The model's performance dropped to an accuracy of 48% when training on two companies and testing on the third. This drop in accuracy suggests that transaction patterns and descriptions may vary across companies even within the same sector, making generalisation more challenging. One possibility is that the transaction format is dependent on the bank provider. For the three companies examined, data across three different bank providers was included. This highlights the need for either additional training data or more sophisticated techniques to improve performance when generalising to new companies.

4.3 Error analysis

Figure 3 reveals transactions where the LLM-generated labels deviate from the manually annotated ground truth. Categories such as Cheques, ATM Withdrawals, Other Income, and Debt / Loan Repayments demonstrate the most pronounced discrepancies. These errors can be attributed to two primary factors: the low frequency of transactions in these categories and the ambiguity in transaction descriptions. The label-imbalanced nature of the dataset poses a challenge for classification accuracy. Categories like Revenue, Suppliers, and Payroll / Consultants have higher transaction counts, resulting in more training data for these labels. As a result, the model demonstrates strong performance in these categories, achieving high accuracy rates. For instance, the model correctly classified 160 transactions under Revenue, with minimal misclassifications. However, for categories such as Cheques and Debt / Loan Repayments, the limited number of examples negatively impacted the model's ability to generalise. Without prior training or fine-tuning on specific transaction categories, the zero-shot model relies entirely on its pre-trained language understanding to assign labels. In the case of low-frequency categories, the model does not receive sufficient context to accurately predict these labels, potentially leading to misclassification. The ambiguity of transaction descriptions presents another key source of error. Categories such as Sundries and Other Outgoing—which inherently lack clear a definitions. These categories require greater context or domain-specific knowledge to be accurately labeled. The confusion matrix shows that Other Outgoing and Sundries were frequently confused with other categories due to their broad nature, which often encompasses transactions with unclear or generalised descriptions. Bank transaction description are usually made up of non-standard abbreviations and short hand text, which complicates the model's ability to understand context and give the correct label, often confusing similar categories as we can observe in the classification patterns. For example, Payroll / Consultants and Suppliers were frequently confused by the model. This suggests that the it could not differentiate between payments for suppliers and those for consultants as both categories often involve similar descriptions, such as invoices or payment references. Similarly, the confusion between Revenue and Other Income illustrates challenges in distinguishing between different types of income transactions. While the model achieved high accuracy for Revenue, several Other Income transactions were misclassified as Revenue.

For the predictive model, errors in the leave-one-company-out setup were more pronounced in company-specific categories, highlighting the need for either more training data or improved feature representation to capture the nuances of different companies' financial activities.

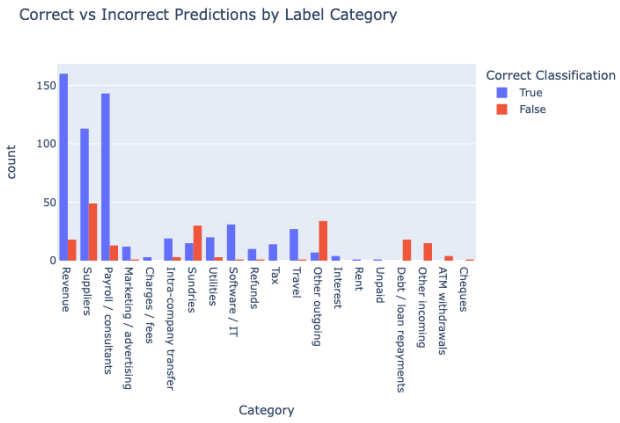

Figure 3: Counts of correct and incorrect labels for each category.

5 Future work

While our results demonstrate promising accuracy in LLM-generated labels for SME bank transactions, particularly when compared to human annotations, several challenges remain. One area for future research is the handling of sensitive financial data. Given the privacy concerns associated with sensitive bank transaction data, making sure that these models are both secure and compliant with regulations is essential.

In terms of model performance, future work should focus on improving accuracy for low-frequency or ambiguous categories, such as Sundries and Other Outgoing, where additional domain-specific context may be required. Leveraging few-shot learning and fine-tuning LLMs on selected transaction types could help the models generalise better in these cases, enhancing performance on less common categories. Additionally, an investigation into whether predictive models can effectively assign labels without relying on LLMs was conducted. This approach aims to enhance scalability and efficiency by utilizing transaction embeddings while addressing data security concerns. Further work is needed to improve the generalizability of these models to new companies across various sectors.

Additionally, a hybrid approach combining different techniques, such as FastText embeddings or other word vector models, with weak labeling frameworks could also be explored. This method would allow the combination of weak supervision with various models to further enhance label accuracy and tackle the limitations of any single technique.

Finally, improving interpretability remains a significant challenge. LLMs function as black-box models, which raises concerns especially for non-expert users if the main users interaction with the model are non-experts. Developing methods to explain model predictions in a clear, explainable, and interpretable way is important, particularly in real-time decision-making scenarios. Ensuring that these models are transparent and interpretable will improve trust and facilitate their broader adoption by financial institutions.

6 Conclusions

In this paper, we explored the use of LLMs to automate the categorisation of SME bank transactions, demonstrating promising accuracy compared to human-labeled data. Our findings suggest that LLMs are effective at generating labels for the majority of transaction categories, particularly those with recurring descriptions such as Tax and Payroll. However, challenges remain in handling ambiguous or infrequent categories like where additional domain-specific or even company-specific knowledge may improve performance.

The scalability of LLMs offers significant potential for financial institutions looking to reduce the time and cost associated with manual transaction labeling. By automating this process, institutions can improve their risk assessments and develop more tailored financial products for SMEs. This automation can pave the way for easier and more transparent access to funding for SMEs that often struggle to secure traditional financing due to inconsistent financial histories or limited access to formal credit assessments. With improved portfolio monitoring, these businesses can increasing their chances of obtaining funding for growth and sustainability, despite, in some instances, not having a clear and sounds financial history. Nevertheless, the black-box nature of these models and the sensitive nature of financial data highlight the need for further work on model interpretability and privacy-preserving techniques.

Overall, this study underscores the potential of LLMs to modernise transaction classification for SMEs, facilitating more accessible and equitable financial services. As future research addresses privacy concerns, improves accuracy and model interpretability, LLMs could become important tools in supporting SMEs' access to funding and financial growth.

References

[1] [n. d.]. SlopeGPT: The first payments risk model powered by GPT | by Jason Huang | Slope Stories | Medium. https://medium.com/slope-stories/slopegpt-the-first-payments-risk-model-powered-by-gpt-4-cd444ab5242d

[2] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. 2007. Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 787–788.

[3] Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. 2021. Interactive Weak Supervision: Learning Useful Heuristics for Data Labeling. In *International Conference on Learning Representations*.

[4] Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically Labeled Data Generation for Large Scale Event Extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 409–419. https://doi.org/10.18653/v1/P17-1038

[5] Samah Fodeh, Bill Punch, and Pang-Ning Tan. 2011. On ontology-driven document clustering using core semantic features. *Knowledge and information systems* 28 (2011), 395–421. Publisher: Springer.

[6] Silvia Garcia-Mendez, Milagros Fernandez-Gavilanes, Jonathan Juncal-Martinez, Francisco Javier Gonzalez-Castano, and Oscar Barba Seara. 2020. Identifying Banking Transaction Descriptions via Support Vector Machine Short-Text Classification Based on a Specialized Labelled Corpus. *IEEE Access* 8 (2020), 61642–61655. https://doi.org/10.1109/ACCESS.2020.2983584

[7] Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch, and Alan Akbik. 2023. Fabricator: An Open Source Toolkit for Generating Labeled Training Data with Teacher LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Yansong Feng and Els Lefever (Eds.). Association for Computational Linguistics, Singapore, 1–11. https://doi.org/10.18653/v1/2023.emnlp-demo.1

[8] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*. Association for Computational Linguistics, Mexico City, Mexico, 165–190. https://doi.org/10.18653/v1/2024.naacl-industry.15

[9] Cheng-Yu Hsieh, Jieyu Zhang, and Alexander Ratner. 2022. Nemo: Guiding and Contextualizing Weak Supervision for Interactive Data Programming. *Proceedings of the VLDB Endowment* 15, 13 (Sept. 2022), 4093–4105. https://doi.org/10.14778/3565838.3565859

[10] Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. http://arxiv.org/abs/2305.05862 arXiv:2305.05862.

[11] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large Language Models in Finance: A Survey. In *4th ACM International Conference on AI in Finance*. ACM, Brooklyn NY USA, 374–382. https://doi.org/10.1145/3604237.3626869

[12] Pierre Lison, Jeremy Barnes, and Aliaksandr Hubin. 2021. skweak: Weak Supervision Made Easy for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, Heng Ji, Jong C. Park, and Rui Xia (Eds.). Association for Computational Linguistics, Online, 337–346. https://doi.org/10.18653/v1/2021.acl-demo.40

[13] Lefteris Loukas, Ilias Stogiannidis, Odysseas Diamantopoulos, Prodromos Malakasiotis, and Stavros Vassos. 2023. Making LLMs Worth Every Penny: Resource-Limited Text Classification in Banking. In *Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF '23)*. Association for Computing Machinery, New York, NY, USA, 392–400. https://doi.org/10.1145/3604237.3626891

[14] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham. 2008. A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data. In *2008 Eighth IEEE International Conference on Data Mining*. 929–934. https://doi.org/10.1109/ICDM.2008.152 ISSN: 2374-8486.

[15] Mohammad M. Masud, Clay Woolam, Jing Gao, Latifur Khan, Jiawei Han, Kevin W. Hamlen, and Nikunj C. Oza. 2012. Facing the reality of data stream classification: coping with scarcity of labeled data. *Knowledge and Information Systems* 33, 1 (Oct. 2012), 213–244. https://doi.org/10.1007/s10115-011-0447-8

[16] Artem Mateush, Rajesh Sharma, Marlon Dumas, Veronika Plotnikova, Ivan Slobozhan, and Jaan Übi. 2018. Building Payment Classification Models from Rules and Crowdsourced Labels: A Case Study. In *Advanced Information Systems Engineering Workshops*, Raimundas Matulevičius and Remco Dijkman (Eds.). Vol. 316. Springer International Publishing, Cham, 85–97. https://doi.org/10.1007/978-3-319-92898-2_7 Series Title: Lecture Notes in Business Information Processing.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. _eprint: 1301.3781.

[18] Mohamad Amin Osman, Shahrul Azman Mohd Noah, and Saidah Saad. 2022. Ontology-based knowledge management tools for knowledge sharing in organization—a review. *IEEE access* 10 (2022), 43267–43283. Publisher: IEEE.

[19] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings Using Compositional n-Gram Features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 528–540. https://doi.org/10.18653/v1/N18-1049

[20] Aleksandra Piktus, Necati Bora Edizel, Piotr Bojanowski, Edouard Grave, Rui Ferreira, and Fabrizio Silvestri. 2019. Misspelling Oblivious Word Embeddings. In *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)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 3226–3234. https://doi.org/10.18653/v1/N19-1326

[21] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: creating large training sets, quickly. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3574–3582.

[22] Mingtian Shao and Naijie Gu. 2021. Anomaly Detection Algorithm Based on Semi-Supervised Collaborative Strategy. *Journal of Physics: Conference Series* 1944, 1 (June 2021), 012017. https://doi.org/10.1088/1742-6596/1944/1/012017 Publisher: IOP Publishing.

[23] Duc Tuyen Ta, Wajdi Ben Saad, and Ji Young Oh. 2023. Specialized text classification: an approach to classifying Open Banking transactions. In *2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT)*. IEEE, Lviv, Ukraine, 1–4. https://doi.org/10.1109/CSIT61576.2023.10324203

[24] Liam Toran, Cory Van Der Walt, Alan Sammarone, and Alex Keller. 2023. Scalable and Weakly Supervised Bank Transaction Classification. http://arxiv.org/abs/2305.18430 arXiv:2305.18430 [cs].

[25] Albert Tseng, Jennifer J. Sun, and Yisong Yue. 2022. Automatic Synthesis of Diverse Weak Supervision Sources for Behavior Analysis. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2201–2210. https://doi.org/10.1109/CVPR52688.2022.00225 ISSN: 2575-7075.

[26] Paroma Varma and Christopher Ré. 2018. Snuba: automating weak supervision to label training data. *Proceedings of the VLDB Endowment* 12, 3 (Nov. 2018), 223–236. https://doi.org/10.14778/3291264.3291268

[27] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. http://arxiv.org/abs/2303.17564 arXiv:2303.17564.

[28] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. http://arxiv.org/abs/2306.06031 arXiv:2306.06031.

[29] Tuan Norhafizah Tuan Zakaria, Mohd Juzaiddin Ab Aziz, Mohd Rosmadi Mokhtar, and Saadiyah Darus. 2020. Text Clustering for Reducing Semantic Information in Malay Semantic Representation. *Asia-Pac. J. Inf. Technol. Multimed* 9 (2020),

11–24.
[30] Weiguo Zhang, Chao Wang, Yue Zhang, and Junbo Wang. 2020. Credit risk evaluation model with textual features from loan descriptions for P2P lending. *Electronic Commerce Research and Applications* 42 (July 2020), 100989. https://doi.org/10.1016/j.elerap.2020.100989