An Approach to Business Workflow Software Architectures: A Case Study for Bank Account Transaction Type Prediction

Fatma Gizem Çallı[1(✉)], Çağdaş Ayyıldız[2], Berke Kaan Açıkgöz[2], and Mehmet S. Aktas[2]

[1] Software Development Department, Eçözüm Bilgi Teknolojileri A.Ş., Istanbul, Turkey
gizem.calli@ecozum.com

[2] Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey
{cagdas.ayyildiz,berke.acikgoz}@std.yildiz.edu.tr, aktas@yildiz.edu.tr

Abstract. Today, practically every bank's computer system can automatically categorize transactions. If someone uses their debit/credit card to buy groceries or clothing, they can see the sort of expense on their user account in seconds. Even though banks provide this level of categorization for individual users, there is no categorization solution for accounting systems. In this article, the main objective is to design and develop a business workflow that can predict bank account transaction types. Various machine learning and deep learning algorithms are used to accomplish this purpose. In the prototype implementation, Support Vector Machines, Random Forest, Long Short-Term Memory Networks, and Frequent Pattern Growth algorithms are used, and the prediction successes of these techniques are analyzed.

Keywords: Electronic bank transaction categorization · Electronic bank transaction prediction · Association mining rules · Deep learning · Machine learning · LSTM · FP-Growth

1 Introduction

People generally use cash to buy and sell products and services, as well as to keep their savings. The use of cash as a medium of exchange has fallen dramatically in recent years. People are increasingly turning to electronic transaction methods, such as debit cards and digital payments. As an example, only 23% of all payments in 2019 were made using cash [1]. This increase in banks' online money transfers has dramatically increased the data volume of banks.

Banks categorize the transfer types for their customers. Although these categories are very diverse, the definition of the type of transaction is usually chosen by the customer performing the transfer. During this selection, most customers

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022 Workshops, LNCS 13381, pp. 709–723, 2022.
https://doi.org/10.1007/978-3-031-10548-7_51

do not care much about the type of transaction to complete the transaction as soon as possible and fill this section with incorrect or unnecessary information. The fact that most customers carry out their transactions in this way reduces the value of the data held by the banks and makes it incomprehensive.

Motivation: The banks offer expense categorization to individual users. However, there are no solutions for the accounting categorization for enterprises. Bank account transaction types for enterprises consist of many variable categories. These categories facilitate the periodic summarization for an accountant in the enterprise. There are categories such as Electronic Funds Transfer (EFT), Wire Transfer, bank giro credit, pos machine transactions, etc. Our motivation in this research project is to categorize bank account transactions into specified categories.

Research Questions: A business process is introduced within the scope of this research paper to predict different types of bank account transactions. The prototype application of this business process was developed and the success of this application is calculated. In this context, the following research questions were examined:

- How should be the business process implemented to predict the type or category of bank money transfer?
- What should be the correct embedding approach to model bank account transaction types?
- Which approaches (classical machine learning algorithms, deep learning algorithms, association rule mining algorithms) will give more successful prediction results in bank account transaction type prediction?

Contribution: The goal of this article is to propose a business workflow to process and estimate the bank account transaction types using the created models. A prototype application is created to demonstrate the effectiveness of the suggested business workflow. The results acquired from the developed prototype application is tested to determine whether it accurately categorizes actual transaction. Within the scope of this research, the estimation of bank account transaction types was studied with three different approaches:

- Machine Learning Algorithms
- Deep Learning Algorithms
- Association Rule Mining Algorithms

In line with this research, an experimental study was conducted to see how successful different artificial intelligence approaches predict the category of bank account transactions. The results show that deep learning algorithms outperform other methods.

Paper Structure: The structure of this article is organized as follows. In chapter II, the basic concepts of machine learning, deep learning, association rule mining, and then the literature on big data architectures are reviewed. Section 3 provides the methodology and describes the proposed model. In Sect. 4, prototype, evaluation and experimental results are given. In Sect. 5, the study results are evaluated and future studies are mentioned.

2 Fundamental Concepts and Literature Review

2.1 Fundamental Concepts

Machine Learning: Machine learning uses mathematical models to assist a computer in learning without being given explicit instructions. This is a subcategory of artificial intelligence. In machine learning, algorithms are used to find patterns in data. These patterns are also used to create a data model that predicts the outcome. Machine learning results improve with more data and experience, just as people do with more practice.

Because of its adaptability, machine learning is an excellent option in scenarios where data, requests or tasks are constantly changing or when it is impossible to code a solution effectively [4].

Deep Learning: Deep learning is a type of machine learning that employs artificial neural networks to assist digital systems in learning from unstructured, unlabeled data and making decisions.

Machine learning, in general, teaches AI systems to learn by analyzing data, finding patterns, generating recommendations, and adapting. Deep learning allows digital systems to learn from examples rather than just following predefined rules, and then it uses that knowledge to act, behave, and perform like people.

Deep learning software is used to evaluate vast and complicated data sets, conduct complex and non-linear tasks, and reply to a text, voice, or pictures more quickly and accurately than humans by data scientists and developers.

Companies in healthcare, transportation, technology, and other industries invest in deep learning to speed innovation, open new possibilities, and stay ahead of their competition as data volumes expand. As consequence of cheaper and more powerful computers, deep learning becomes incredibly valuable [3].

Association Rules: Association Rules are used to find relations between items in an item set [9–11]. Analyzing connections of customers' purchasing habits is a real-life example of association rule mining. "If X and Y exist, then Z must exist," is how the rule mining is usually described. For the purchasing habits example mentioned before, one can describe the association rules as "if a customer bought X and Y then this customer is inclined to buying Z". Association Rule Mining methods generally use two fundamental formulas: Support and Confidence. The ratio of a subset of all possible transactions to all transactions completed is shown by support and the ratio of items that can be seen if the given condition is correct is shown by confidence.

Most association rule mining algorithms use minimum support and minimum confidence values to remove some unnecessary rules. These values should be determined in the algorithm's parameter. The algorithm used in this paper is described further down.

2.2 Literature Review

While there are no studies in the field of accounting that categorize bank transactions, there are some classification and categorization studies in electronic banking. Adam Wamai Egesa's work in financial transaction analysis was achieved 90% and 84% success in Support Vector Machine (SVM) and Naive Bayes (NB) algorithms, respectively. In the same study, it is seen that one of the success factors is the n-gram data representation method [2]. In another study conducted with the same concept, short text classification was applied on Banking Transaction Descriptions by using Specialized Labeled Corpus [5]. In this study, the performances of SVM as a Machine Learning algorithm and different word n-gram features were compared according to various train test split ratios. As a result of this comparison, SVM and ALL-IN-1 model [6] produced the most successful result with 95.05% precision.

Within the scope of machine learning, there are also similar studies like credit card transactions fraud detection in electronic banking. Some of the methods used in these studies are Random Forest (RF) and Long Short-Term Memory (LSTM) networks. Since 97.67% success was achieved in a similar study with RF [7], the RF algorithm was also included in this study. In equivalent research on fraud detection in credit card transactions, Bénard Wiese's study showed that LSTM produces more successful results than SVM and Feed Forwarding Neural Network (FFNN) [8].

Association rules are used mainly to list other actions that people can take due to reviewing their previous efforts. However, this listing and analysis process takes too long for systems with big data. Programming frameworks such as MapReduce and Spark are often used to manipulate data in such systems [14–16].

FP-Growth algorithm is one of the most widely used algorithms for deleting repeating patterns and generating rules today [16]. But such algorithms are slow in terms of working speed. The FP-Growth algorithm can run in parallel on distributed machines [17]. In the Parallel FP-Growth (PFP) algorithm, the MapReduce approach divides large processing loads into parts, and each machine connected to the algorithm runs on a different processing group. In this way, while contributing to progress, the workload per machine is reduced on the one hand. This strategy, however, does not account for load balancing. The BPFP method separates the total task into equal subtasks by removing the mining unit's duty [18].

Data pre-processing steps in big data processing platforms have been the subject of research [21,22]. Studies on big data processing libraries for data mining were observed [23]. In the business and science domains, we also observe workflows for various purposes [24,25]. In the e-commerce field, there have been studies on supervised learning business workflows for predicting purchase intent [24]. This study also uses big data processing libraries for data pre-processing and data mining algorithms and differs from previous works as it introduces a business workflow for bank transaction type prediction.

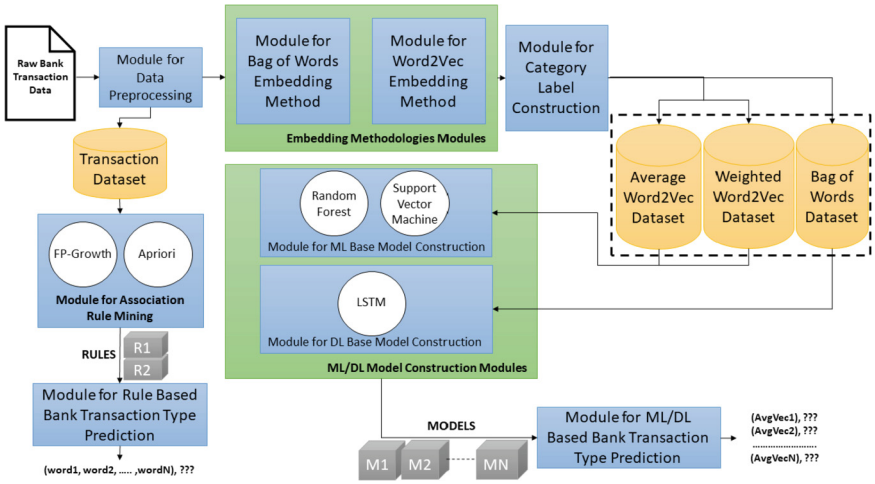

Fig. 1. A system architecture that can predict bank transaction type

3 Methodology

As part of this study, first and foremost, some bank transaction information of Eçözüm Bilgi Teknolojileri A.S. customers were collected in the form of an excel file.

The architecture of the methodology proposed in this study is as shown in Fig. 1. The architecture's main goal is to analyze customer account transaction data and extract useful information from it using Machine Learning, Deep Learning and, Association Rule Mining methods.

In this proposed model, the bank account transactions that customers performed is going to be predicted. Within the context of this study, a prototype is created to illustrate the usability of the suggested architecture. In Sect. 4, the prototype application is briefly explained. In addition, the performance and scalability issues of the application are addressed in Sect. 5. The following sections go over the data analysis and classification processes in detail.

3.1 Module for Data Preprocessing

The initial version of the data consists of 32 columns and 100,000 rows of transaction information. Due to hardware-based computational issues, the sample size is reduced accordingly at this level. Description1, Description2, and Category Name are three columns that were considered useful in the data set. The Description1 and Description2 columns contain the customer and bank's description while performing the transaction. The column Category Name is the transfer type specified by the customer. The values of these columns are the categories specified by the accountants of enterprises.

To analyze the words in Description1 and Description2 columns, the Description1 and Description2 columns were combined into a specific column named as Description. The letters in this column were converted to lowercase with the lowercase function to avoid case-sensitivity problems. The frequencies of the words in the Description column are used to indicate their importance for the algorithm. Because of this, the frequencies of the words in the Description columns in all rows in the entire data was calculated, and those that fell below a certain threshold value were removed from the data. This eliminates fewer words that aren't important in the decision-making phase before they enter the algorithm.

3.2 Module for Bag of Words Embedding Method

One of the main factors that are of great importance and increase the success is how the data is represented before training a model. The bag-of-words (n-gram) method, one of the data representation methods, is frequently used in natural language processing problems. It is seen that the n-gram method gives successful results in similar problems where the sequential or combined probability of the words in the data is essential. Within the project's scope, each Description line is divided into separate tokens from n=1 to n=n from the beginning of the line, according to the maximum length of transaction. E.g., Suppose there is a bank transaction "eft from account A to account B". In the suggested method, the data representation of the original data is produced by n-gram sequences that were obtained separately as "eft", "eft account", "eft account A", "eft account A account ", "eft account A account B".

3.3 Module for Word2Vec Embedding Method

To make the data more meaningful, the data types were converted from word to vector after calculating the frequencies of the words contained in the data. In this way, a sequence is created where each word is represented by a set of 128 vectors. The word vectors which were in the same sentences are combined and the averages of created vectors are calculated. The mean vector was created in two ways to make different comparisons while averaging: Weighted Vector Average and Equally Weighted Vector Average. The function of this module is illustrated in Fig. 2.

The weights of the words were calculated using the frequency values of the words in the Weighted Vector Average. In the Equally Weighted Vector Average, the weights of all words were accepted as 1.

3.4 Module for Category Label Construction

During the construction of the category labels, it is observed that the same category names were repeated for reasons such as extra punctuation marks or English/Turkish characters which can be used interchangeably. For these repetitions not to be processed as different categories by the algorithm, and not to

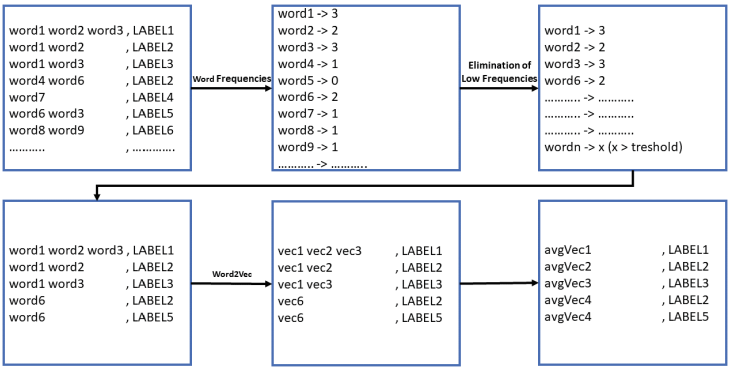

Fig. 2. Module for Word2Vec Embedding Method

reduce the success of the training results, some required operations are applied to the category name list. All letters are converted to lowercase, punctuation is cleaned, similar characters in Turkish/English are mapped and singularized, and finally, words in categories containing more than one word are concatenated with an underscore. In this way, the category names are cleaned and can be expressed with a single vector. As a result of these processes, the number of categories is reduced from the count of 173 to 162.

3.5 Modules for ML/DL Model Construction

Two different learning methods are used for training to predict categories using the data set: machine learning and deep learning.

Module for ML-Based Model Construction: In this project, Support Vector Machines (SVM) and Random Forest (RF) ML algorithms are used to predict transaction types. SVM is appropriate for the data set because it outperforms other classification algorithms based on training results when the data set size is large. SVM also has the advantage of being memory efficient.

Random Forest is a decision tree algorithm that was created to address the problem of overfitting in commonly used decision tree algorithms. The Random forest algorithm works similarly to a typical decision tree algorithm. Random Forest differs in that it employs multiple decision trees and accepts the most common output as the label. The label is determined by taking the average of the obtained results.

Module for DL-Based Model Construction: As a Deep Learning Based Model, Long Short-Term Memory (LSTM) networks are used to make category estimation. LSTM which is a type of Recurrent Neural Network (RNN) has been applied in this project due to its success in the problems requiring sequential estimation [13].

LSTM networks consist of structures called cells. Each consecutively feed forwarding cells encapsulate 3 separate gates called Forget Gate, Input Gate, and Output Gate. Due to the feed forwarding structure, two states are registered in the cell: the current state and the hidden state from the last cell. During the cell stream, in the Forget Gate, it decides whether the hidden state from the previous loop will be kept or not. According to the sigmoid function, if the result is "0" the information is forgotten and a new one from the previous cell is saved. If it is "1", the loop continues as it is. The Input Gate calculates the importance of the new data. If the result is negative (−1), the new information is removed from the cell. If it is positive (1), the current loop state of the cell is saved with new data. Finally, the Output Gate calculates the hidden state value for the next cell [19].

Within the project's scope, a model is designed with 100 Embedding Layer, 100 LSTM Layer, and 162 Dense Layer and trained with the bank account transaction data set. The results are shown in the Experimental Results section according to Accuracy and F1-Score values.

3.6 Module for Association Rule Mining

Methods like Association Rule Mining are used to extract meaningful relationships from data. As one of the Association Rule Mining algorithms, FP-Growth Algorithm is implemented in this module. As a result of the algorithm, Association Rules are generated.

FP-Growth Algorithm: The FP-Growth Algorithm is frequently used in Association Rule Mining as an enhanced version of the Apriori Algorithm. The FP-Growth Algorithm aims to reduce the high disk I/O and computational costs encountered in Apriori [11,12]. Instead of accessing the data at each step, the frequency and association information of the data is kept in the tree structure called FP-Growth Tree by providing access to data set two times. First of all, the frequency of each item is calculated. Based on the specified minimum support value, the Fp-Growth Tree is created by the items that are placed in descending order from the tree's root according to their frequency.

The list of routes for each item in the tree is saved as the Conditional Pattern Base, starting with the items with the lowest frequency, and the most used patterns for each item in this list are placed in the Conditional Pattern Frequency tree with their frequency values. Finally, the list of the rules is extracted from the Conditional Pattern Frequency tree. These are association rules developed for the provided data set using the FP-Growth Algorithm [20].

3.7 Module for Rule Based Bank Transaction Type Prediction

The 'Diger' label appears frequently in the data set used for this research project. This is a generic tag that has no bearing on inference. Using the association rules algorithm, it is attempted to predict whether the transactions labelled as

'Diger' have meaningful values. Association rules are produced with the help of Fp-Growth algorithm. The purpose of the obtained association rules is to categorize the data whose Category Name is 'Diger' in the data set if it complies with the rules. This estimation model is used to estimate rates on the test data set.

3.8 Module for ML/DL Based Bank Transaction Type Prediction

The models that can make the ideal predictions are obtained after the training. The success rates of these models' predictions are provided. When the data from the data set with the Category Name column of 'Diger' is given into the same models, the previously unknown transaction type becomes predictable.

4 Prototype and Evaluation

4.1 Prototype

This section explains how the technologies in this study are used. Python 3 is preferred as the software language because it works fast and is easy to use. The code is written using the Anaconda toolkit's Jupyter Notebook. Jupyter Notebook is a development environment that allows you to keep your code in different notebooks and run and analyze it in parts.

The transaction data received from Eçözüm Bilgi Teknolojileri A.Ş. is converted from excel to CSV format. For data in CSV format, there are ready-to-read commands such as 'read_csv(data)' in the Pandas library and 'read.csv(data)' in the Apache Spark library. Furthermore, reading and writing operations on CSV files are faster than those on XLSX files. A DataFrame containing the edited versions of the preprocessed data is created and saved using the Pandas library. Apache Spark is used to run and evaluate big data analytics operations across distributed systems.

The Spark MLlib which is a collection of machine learning algorithms is used in Spark. Association Rule Mining algorithms were implemented in Spark using this library.

4.2 Data Set

A data set containing customer bank account transaction information is used. This data set was obtained from Eçözüm Bilgi Teknolojileri A.S., a company that can report e-collection, bank integrations, virtual and physical POS transactions, as well as develops open banking technologies. The transfer information of 100,000 transactions was sufficient, considering the server's operating speed and the complexity analysis of the algorithms. The data set used includes 32 columns and 100,000 rows of transaction data. There are 68,281 rows in the data set, with the 'Category Name' column being 'Diger'. There are 173 different category types in the remaining 31,719 lines.

4.3 Test Design

In this study, Machine Learning, Deep Learning, and Association Rule Mining approaches are applied to distributed networks using Apache Spark. Random Forest (RF) and Support Vector Machines (SVM), as well as Long Short Term Memory (LSTM) and Frequent Pattern (FP) Growth algorithms, are used in Bank Account Transaction Type Prediction.

It was decided to focus on the data size to be used for training in order to make the training of the models more efficient. The data set is divided into two parts: a training set of 80% and a test set of 20%, with 80.000 in the training set and 20.000 in the test set.

RF and SVM algorithms in Python's Sklearn library are used within the scope of machine learning. Average Word2Vec embedding and Weighted Word2Vec embedding are the two types of vectors available for RF and SVM algorithms. To speed up reading and writing operations from the file, the data frame obtained during the conversion of words to vectors is saved as parquet file. The Sklearn library's "model selection" subheading is used to achieve the desired model training functions. The Accuracy and F1 Score values of these models are calculated. These values represent the success rates of the models.

Within the context of deep learning, the LSTM model is trained with the batch size 64 parameter. In the model creation, prediction, and evaluation stages, Python's Tensorflow Keras, SckitLearn, and nltk libraries are used.

The efficient FP-Growth Python library is used to apply the FP-Growth algorithm to the data set in order to determine the association rules, as mentioned in the Methodology section. This library provides the opportunity to save a Dataframe with the details of the rules it generates. The results acquired are used to develop the estimating module.

The algorithms were run on an Intel XEON x5650 processor with 8 GB of RAM in a server. This server is provided by Eçözüm Bilgi Teknolojileri A.S. to ensure that company information is not leaked.

The next section explains how to compare the outcomes of the algorithms and calculate the efficiency.

4.4 Experimental Results

This section presents deep learning and machine learning training along with testing results. The accuracy and F1-Score metrics are used to comprehensively summarize these results.

Accuracy Tests for Machine Learning Algorithms: Figure 3 shows the accuracy of the tested machine learning algorithms, while Fig. 4 shows the F1-Score comparison.

The RF and SVM values shown in the figure are the versions of the Weighted Word2Vec data set, as mentioned in the Sect. 3. Random Forest and SVM algorithms have additional "AVG" versions, as shown in the figure. The models are trained using the Average Word2Vec data set in "AVG" versions.

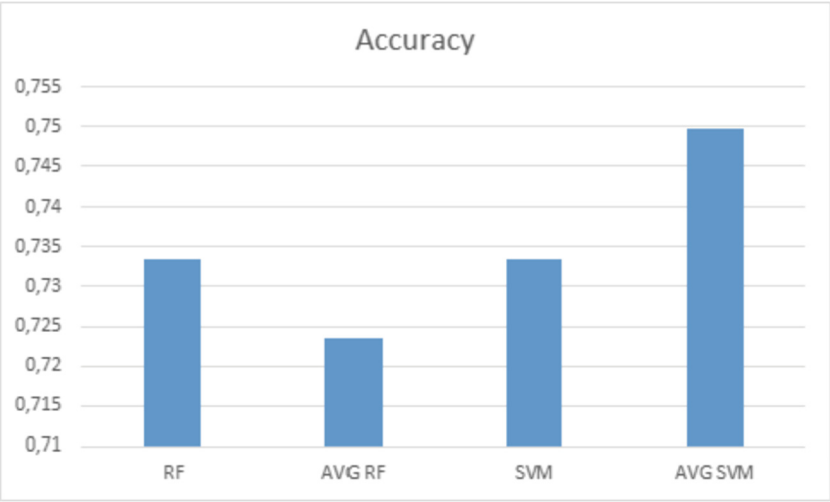

Fig. 3. Accuracy rates of ML algorithms

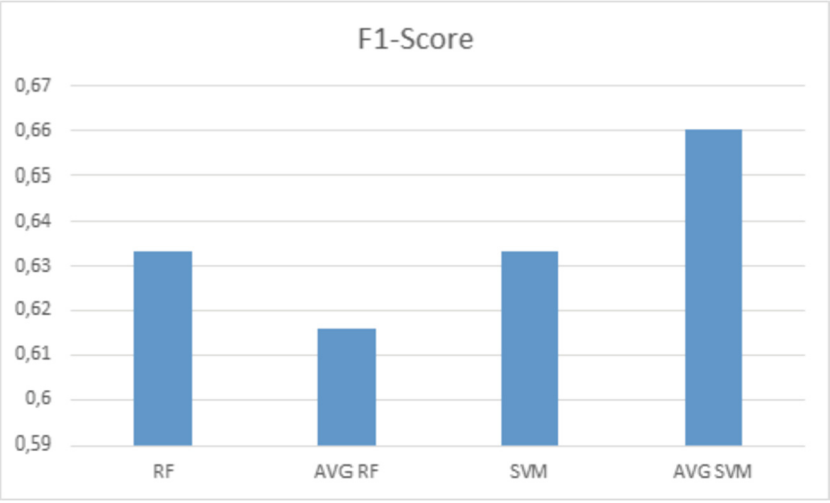

Fig. 4. F1-score rates of ML algorithms

The Random Forest and SVM algorithms both have accuracy rates of 73.3%. On this basis, it can be concluded that both algorithms are capable of training the desired model. When it comes to the "AVG" versions, AVG-RF has a 72.3% accuracy rate, while AVG-SVM has a 75% accuracy rate. While the rate of Random Forest decreases as the feature vector changes, the rate of SVM increases.

F1-Score rates of the Random Forest and SVM algorithms are 63.3%. The F1-Score rate is 61.6% for AVG-RF algorithm and 66% for AVG-SVM algorithm.

When these two figures are compared, we can conclude:

- The SVM algorithm performs the same or better than the Random Forest algorithm in all cases.
- Since the highest Accuracy rate obtained with machine learning is 75%, different methods can be tried to get better results.
- More successful models can be found as the highest F1-Score rate obtained by machine learning is 66%.

Accuracy Tests for Deep Learning Algorithms: Figure 5 shows the accuracy of the tested deep learning algorithms, while Fig. 6 shows the F1-Score comparison. One of the deep learning algorithms, LSTM, is used, as mentioned in Sect. 3 when explaining the methodology. The "AVG" version of LSTM refers to the data set's test result, in which Word2Vec vectors are equally weighted during training and the "Weighted" version of LSTM refers to the data set's test result, in which Word2Vec vectors are weighted during training. The "N-GRAM" version of LSTM refers to the data set's test result, in which the Bag of Words Data set was used in the training phase.

The accuracy rate of the Weighted LSTM, AVG-LSTM, and N-GRAM-LSTM methods are 73.2%, 76%, and 82.9%, respectively. There appears to be an increase in LSTM with the change in the feature vector.

F1-Score rates are 63.6% for the Weighted LSTM algorithm, 70% for the AVG-LSTM algorithm, and 82% for the N-GRAM-LSTM algorithm.

When these two figures are compared, we can conclude:

- N-Gram embedding method gives the best values for Accuracy and F1-Score.
- The Average data set gives higher values for both Accuracy and F1-Score compared to the Weighted data set.
- N-Gram gives better results than Word2Vec in all metrics.

Prediction Rates for Association Rules: Within the scope of the research, FP-Growth from association rule mining algorithm was used to produce Association Rule. There can be no accuracy rate established because the algorithm does not represent a scalable accuracy rate. The accuracy with which the algorithm's rules can label the data set was compared and deemed a success factor in determining the algorithm's success rate. The support and confidence values were set as 0.01 for support and 0.6 for confidence.

In this part, the experimental results for the FP-Growth algorithm are explained in detail. The data set is split into Train and Test data with 80% and 20% ratios, respectively. Different minimum confidence and minimum support values were tried, but 0.06 minimum confidence and 0.01 minimum support values produced the most reasonable result. In this part of the study, 43.6% of the test value could not be assigned to a category, 28.4% was correctly classified, and 27.9% was incorrectly classified.

Since the results show that the Association Rule Mining algorithm is successful in general, FP-Growth requires more data to create reliable rules.

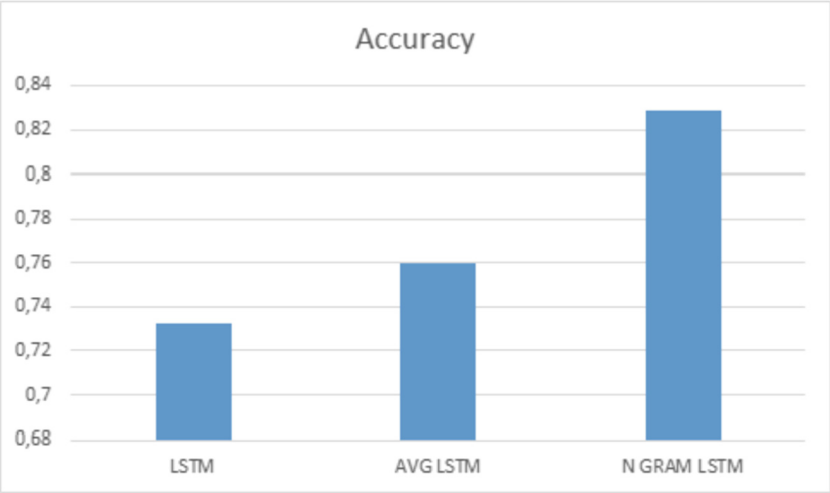

Fig. 5. Accuracy rates of DL algorithms

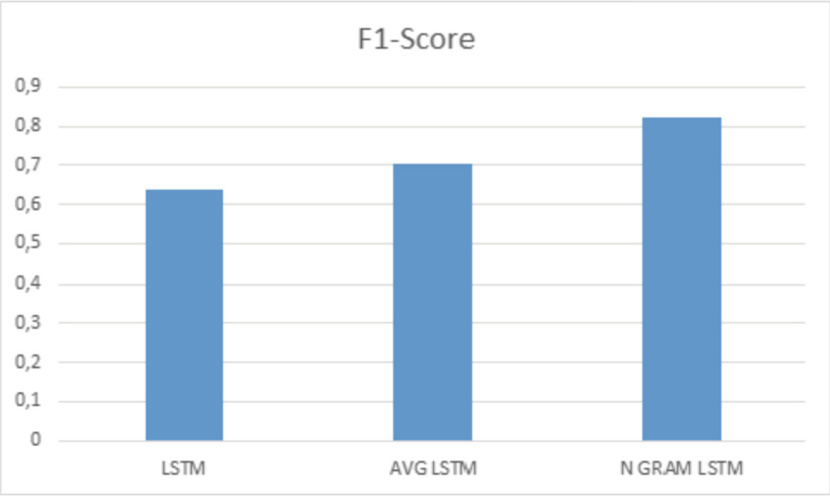

Fig. 6. F1-score rates of DL algorithms

5 Conclusion and Future Work

A software architecture proposal was made as part of this study to determine and make sense of the uncertain data encountered in the banking sector. A prototype application was designed to test the proposed architecture. The proposed architecture uses Random Forest and SVM machine learning algorithms, LSTM deep learning algorithm, FP-Growth association mining algorithm.

By measuring the system's accuracy and execution times, experimental research was conducted to determine the prototype's utility. The results show that the proposed architecture can handle large amounts of data while keeping processing costs low.

The state of art methods such as BERT embedding and association rule mining algorithms can be explored in future studies. The learning performance of the models can be improved by incorporating different bank transfer data and different categories into the data set.

Acknowledgments. We want to extend our sincere thanks to Eçözüm Bilgi Teknolojileri A.Ş. for providing the data set.

References

1. Caswell, E., Smith, H., Learmonth, D., Pearce, G.: Cash in the Time of Covid. Bank of England Quarterly Bulletin, Q4 (2020)
2. Wamai Egesa, A.: Analysis of Financial Transactions using Machine Learning. Department of Computer Science, Lund University (2016). http://lup.lub.lu.se/student-papers/record/8871594
3. Zhong, G., Ling, X., Wang, L.-N.: From shallow feature learning to deep learning: benefits from the width and depth of deep architectures. Wiley Interdisp. Rev. Data Mining Knowl. Disc. **9**, e1255 (2019). https://doi.org/10.1002/widm.1255
4. Razno, M.: Machine learning text classification model with NLP approach. In: COLINS 2019, vol. II, ISSN 2523–4013 (2019). http://colins.in.ua
5. García-Méndez, S., Fernández-Gavilanes, M., Juncal-Martínez, J., González-Castaño, F.J., Seara, Ó.B.: Identifying banking transaction descriptions via support vector machine short-text classification based on a specialized labelled corpus. IEEE Access **8**, 61642–61655 (2020). https://doi.org/10.1109/ACCESS.2020.2983584
6. Plank, B.: ALL-IN-1: short text classification with one model for all languages (2017). arXiv:abs/1710.09589
7. Melcher, K., Silipo, R.: Fraud detection using random forest, neural autoencoder, and isolation forest techniques. In: KNIME (2019)
8. Wiese, B., Omlin, C.: credit card transactions, fraud detection, and machine learning: modelling time with LSTM recurrent neural networks. In: Bianchini, M., Maggini, M., Scarselli, F., Jain, L.C. (eds) Innovations in Neural Information Paradigms and Applications. Studies in Computational Intelligence, vol. 247, pp. 231–268. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04003-0_10
9. Khurana, K., Sharma, S.: A comparative analysis of association rules mining algorithms. Int. J. Sci. Res. Publ. **3**(5), 1–4 (2018) (ISSN: 2250–3153). http://www.ijsrp.org/research-paper-0513.php?rp=P171249
10. Akbar T., Gandomi, A.H., Shahbahrami, A.: A survey of evolutionary computation for association rule mining. Inf. Sci. **524**, 318–352 (2020). ISSN 0020–0255. https://doi.org/10.1016/j.ins.2020.02.073
11. Mustakim, M., et al.: Market basket analysis using Apriori and FP-growth for analysis consumer expenditure patterns at Berkah mart in Pekanbaru Riau. J. Phys. Conf. Seri. **1114**, 012131 (2018). https://doi.org/10.1088/1742-6596/1114/1/012131

12. Mythili, M.S., Shanavas, A.M.: Performance evaluation of Apriori and FP-growth algorithms. Int. J. Comput. App. **79**, 34–37 (2013) https://doi.org/10.5120/13779-1650
13. Jurgovsky, J., et al.: Sequence classification for credit-card fraud detection. Expert Syst. App. **100**, 234–245 (2018). ISSN 0957–4174. https://doi.org/10.1016/j.eswa.2018.01.037
14. Salloum, S., et al.: Big data analytics on Apache spark. Int. J. Data Sci. Anal. **1**, 145–164 (2016). https://doi.org/10.1007/s41060-016-0027-9
15. Tannir, K.: Optimizing Hadoop for MapReduce (2014). https://www.packtpub.com/product/optimizing-hadoop-for-mapreduce/9781783285655
16. Raj, S., Ramesh, D., Sreenu, M., Sethi, K.K.: EAFIM: efficient apriori-based frequent itemset mining algorithm on Spark for big transactional data. Knowl. Inf. Syst. **62**(9), 3565–3583 (2020). https://doi.org/10.1007/s10115-020-01464-1
17. Zhou, L., Zhong, Z., Chang, J., Li, J., Huang, J.Z., Feng, S.: Balanced parallel FP-growth with MapReduce. In: 2010 IEEE Youth Conference on Information, Computing and Telecommunications, pp. 243–246 (2010). https://doi.org/10.1109/YCICT.2010.5713090
18. Zeng, Y., Yin, S., Liu, J., Zhang, M.: Research of improved FP-growth algorithm in association rules mining. Sci. Programm. **2015**, 6 (2015). Article ID 910281. https://doi.org/10.1155/2015/910281
19. Sundermeyer, M., Schlüter, R., Ney, H.: LSTM neural networks for language modeling. Proc. Interspeech. **2012**, 194–197 (2012). https://doi.org/10.21437/Interspeech.2012-65
20. Lin, K.-C., Liao, I.-E., Chen, Z.-S.: An improved frequent pattern growth method for mining association rules. Expert Syst. App. **38**(5), 5154–5161 (2011). https://doi.org/10.1016/j.eswa.2010.10.047
21. Catalkaya, M.B., et al.: Data feature selection methods on distributed big data processing platforms. In: International Conference on Computer Science and Engineering, UBMK 2018 (2018). https://doi.org/10.1109/UBMK.2018.8566451
22. Celik, O., et al.: Implementation of data preprocessing techniques on distributed big data platforms. In: International Conference on Computer Science and Engineering, UBMK 2019 (2019). https://doi.org/10.1109/UBMK.2019.8907230
23. Yildiz, E.C., et al.: Data mining library for big data processing platforms: a case study-sparkling water platform. In: International Conference on Computer Science and Engineering, UBMK 2018 (2018). https://doi.org/10.1109/UBMK.2018.8566278
24. Gurbuz, A., et al.: Prediction of purchase intention on the e-commerce clickstream data. In: SIU 2019 (2019). https://doi.org/10.1109/SIU.2019.8806311
25. Tufek, A., et al.: Provenance collection platform for the weather research and forecasting model. In: SKG 2018 (2018). https://doi.org/10.1109/SKG.2018.00009