Business Process Discovery from Emails: Text Classification and Process Mining - A Case Study of Procurement Process

Innovatus: Digital Transformation in Business Information Systems, Vol. No. 5, Issue 1

Yaghoub Rashnavadi[1], Sina Behzadifard[1], Reza Farzadnia[2], Sina Zamani[1]
rashnavadi55@gmail.com, sina.behzadifard@gmail.com, Farzadniareza@gmail.com, sinazamani920@gmail.com
[1]*Kharazmi University, Tehran, Iran*
[2]*Pars Investment Casting co. Tehran, Iran,*

ISSN (Print:) 2651-6993
Pre-print DOI No. 10.20944/preprints202005.0007.v2

FOR PUBLISHERS ONLY:
Manuscript received: October 13, 2021; revised: November 30, 2021; accepted: December 14, 2021

ABSTRACT

Messages and emails are traveling with the speed of light and making communication more accessible than ever, which has transformed organizations to the degree that they generate billions of emails daily to facilitate their operations and processes. This vast corpus of human-generated content is a rich dataset that can benefit businesses. To address the potential application of such data, we propose a framework to mine and extract the implicit information behind the email loops. This article examines the opportunity that email logs can bring to organizations and proposes a framework to discover process models based on a supervised machine learning technique to classify emails to the activities and Fuzzy Miner to extract the process model from the labeled emails. We also examined the framework with a real-life dataset from the procurement department of the case study company in Iran. The findings demonstrated discrepancies between the discovered process model and the designed business process, highlighting the needed improvements.

KEYWORDS Process Mining, Business Processes, Natural Language Processing, Machine Learning, Email Analysis

1 INTRODUCTION

Email plays a crucial role in our daily lives, with over 124.5 billion[1] business emails sent and received daily; its importance cannot be ignored, especially in the business context. Although email has personal applications, many organizations use this convenient tool to facilitate internal and external communications and even manage complex projects via a large team across different locations.

While organizations are executing their business processes, employees generate a significant amount of unstructured natural language documents through their routine emailing (Shing *et al.*, 2019) stored and archived into databases. So, it can be assumed that this corpus of textual is concerned about organizations' daily operations. They can contain different relevant data to the tasks, activities, and processes. The problem with such data is its unstructured nature, not containing any explicit information about the related business processes. The lack of visibility over such processes can cause intentional or unintentional deviations from the business processes' central goal and affect the organizations' competencies in the long term.

It can be assumed that there is enough information in the email's body text that can provide sufficient information for the recipients to understand the message and take the required action. Therefore, a methodology that can extract the hidden information would enable organizations to have more clarity over their processes, the more structured processes in ERP software, and the ones followed in email loops.

Therefore, our research's objective was to address the potential information implicit in emails which can be used for business process management and highlight how business processes can be discovered from emails. The results can be used for process improvement purposes.

This article proposed a methodology with such characteristics to discover the implicit business process by leveraging a supervised machine learning technique and process mining algorithm. Supervised machine learning techniques are algorithms trained to find a specific mapping from patterns in a labeled training dataset. Such techniques can be used to discover activities based on combining words in an email and map them to a defined set of activities (labels). We used a classifier model from the fastText library in this research, created by Faceboovk's AI Research lab for text classification and word embedding applications[2]. fastText helps the reproducibility of the research over the same dataset with the same model configuration and makes the training and testing process standardized. fastText library supports 157 languages[3].

After the training process, the classified emails were mapped to their relevant activity. The consequence event log is mined with fuzzy Miner to extract the process model.

[1] https://www.campaignmonitor.com/blog/email-marketing/2019/05/shocking-truth-about-how-many-emails-sent/

[2] https://fasttext.cc/

[3] https://fasttext.cc/docs/en/crawl-vectors.html

Innovatus: Special Issue on Digital Transformation
UNIVERSITY OF ASIA AND THE PACIFIC
UA&P

To better examine the methodologies' feasibility, we experimented with a real dataset from a case study company[4] in Iran. We focused on the procurement process to simplify the process discovery. The case study company's main operation is in Iran's oil and gas sector, which makes the company's procurement processes include both international and local.

This research focuses on local procurement, conducted with local suppliers in Iran through the Persian Language. This research has two main contributions. First, the proposed method is the introduction of the supervised text classification method as an activity discovery method, and second, the experimentation of the method with real-life data, which tests the feasibility of the methodology's practicality at the practice and industry level.

2 RELATED WORKS

Business process discovery is one of the most widespread problems among researchers in the business process management domain. However, the challenge of extracting business processes is to make them explicitly visible and comprehensible, and data-driven solutions like process mining can facilitate the process by providing algorithms to synthesize a process model from data, event logs (Van Dongen *et al.*, 2009).

Process mining deals with techniques designed to extract knowledge from event logs (W. Van Der Aalst *et al.*, 2012) and illustrates it as a process model. These techniques provide new tools for a wide range of applications, Process Discovery, Monitoring, and Enhancement, with the goal to facilitate the process alignment and bottleneck analysis while predicting the problems in the execution of processes (Turner *et al.*, 2012).

Van der Aalst & Nikolov (2007) presented a tool for ProM (Process Mining Tool), EmailAnalyzer, to analyze and transform email messages in MS-Outlook into a format used in process mining tools. This research's main goal was to create a social network map from email logs.

Business processes that are rather mental than physical are harder to discover and are executed by "knowledge workers" (Di Ciccio *et al.*, 2012). Di Ciccio *et al.* (2012) addressed such processes to discover automatically and mine the corresponding implicit process model inside them. First, email messages were extracted in their presented methodology, and communication threads were structured upon them. In the second step, essential parts, like activities and tasks, were identified. Finally, the consequence data was mined with Process Describing Grammar (PDG). This approach was presented as "MAILOFMine.".

Extracting workflows using Natural Language Processing (NLP) and sequence mining techniques is another area that researchers explored with unstructured texts from emails like the research conducted by Shing *et al.* (2019). They used latent semantic indexing, an unsupervised technique, and density-based spatial clustering of applications with noise to determine the number of clusters and label events.

The method of email analysis and classification is a well-discovered area. Corston-Oliver *et al.* (2004) from Microsoft Research demonstrated a use case for the email classification method as a summarizer of emails to create a "to-do list". In this research, a dataset of 15,741 email messages was collected. With the help of human annotators, 146 messages were tagged independently. The prepared dataset was used to train a Linear Support Vector Machine (SVM) Model to classify emails to tasks in the to-do list. Borg *et al.* (2021) also tried to propose an approach to classify customer support emails to increase service speed. The method they found with the best performance in F1-Score was a Long-Short Term Memory (LSTM) Network to classify emails into 33 different classes.

Banziger *et al.* (2019) investigated a similar problem using unsupervised machine learning to automatically detect and assign activity labels to messages in a CRM (Customer Relationship Management) tool. Jlailaty *et al.* (2017) also tried to address the same problem with the unsupervised clustering machine learning technique to automatically label emails with the related activity and mine the respected process model.

Mavaddat *et al.* (2011) proposed a three-step method to extract business processes from emails. They created a model to demonstrate the interactions among different role instances in a process: email categorization, conversation network fining, and conversation network tagging. In the first step, they divided emails into two different categories: business process-related and non-business process-related. They tried different algorithms like Naïve Bayes and Support Vector Machine (SVM) to find the best text mining algorithm. The authors used WEKA (Waikato Environment for Knowledge Analysis) tool to automatically categorize emails through learning from the previously prepared training dataset (manually labeled emails dataset). The binary emails classification output allowed them to connect threads and conversations about a similar topic through semantic similarity measurement of each email to the other emails. Finally, to label the interactions between each role instance, authors used the Speech Act Theory (Koller *et al.*, 1970) to classify instances to one of the labels: "assertive, directive, commissive, expressive, and declaration" (illocutionary speech acts). The authors suggested that it is possible to discover business process fragments by exploring the resulting conversation networks and patterns.

Mavaddat *et al.* (2011) and Jlailaty *et al.* (2019) tried to address the challenge of non-related emails and filter the related emails to be mined for activities and related information. After data preprocessing and cleaning, the sentences of each email were classified into two categories: process-oriented and non-process oriented (a binary classification). This classification is conducted through creating a vocabulary dictionary from a process model repositories and ontologies (Jlailaty *et al.*, 2019). For the next step, the business-related emails were clustered with clustering techniques. A label was defined in a "semi-automatic way." Finally, each activity instance's metadata was extracted through the linked information in a cluster, like the recipients'

[4] The companies' primary focus area is the production of industrial turbo-machineries for Iran's oil and gas market. The manufacturing process of these machines is very complicated, which demands high precision in every activity. Different parts, raw materials, and services are needed to feed the production lines. A strong procurement team supplies these inputs to manage purchasing processes and suppliers' communications.

organizational role. Researchers tested the proposed approach on the Enron dataset.

Laga *et al.* (2019) presented an approach to facilitate labeling emails and automatically classify them into process instances, activity IDs, and actors (process-related items) using machine learning techniques. In this approach, the authors created a platform that helps users collaborate to create annotated data and increase data volume through time. They also created a proof of concept with a dataset containing 1026 emails that used logistic regression classifier as a multiclass predictor to predict process-related tags to validate their approach.

This literature review focused on providing a brief overview of research concerning business process discovery from emails. The methods in the previous studies mainly were investigated unsupervised machine learning, like clustering techniques. On the other hand, some studies experimented with the supervised email classification methods to classify emails into non-related and related emails (Binary Classification) or predict a set of process-related labels (Multiclass Classification). One of the main implicit assumptions in the literature is to associate each email to only one activity. In contrast, this assumption cannot be valid in most real-life situations. There might be occasions when emails mention multiple processes.

In this study, we focused on a similar problem. However, we addressed it with a different methodology, discussed in the following section.

3 METHODS

The main objective of this research is to address the potential insight inside daily communication of organizations through emails and to explore how the discovered process model from emails can be used for process improvement.

To achieve the research objective, we proposed a framework consisting of three steps: Data Preprocessing, Activity Discovery, and Process Mining.

Figure 1 presents the methodology that will be used. The input of this methodology is email log extracted from communications, and the output is a process model discovered from the result of mining the final event log.

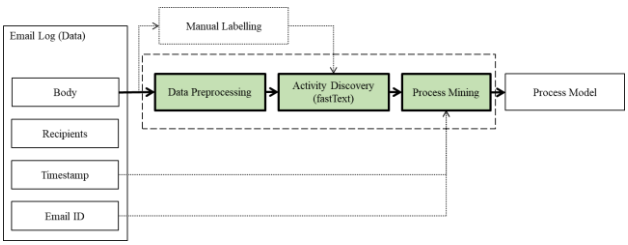


Figure 1: the proposed methodology

3.1 Preprocessing

Email logs contain four main attributes: Body, Subject, Recipients, and Timestamp. While email attachments can also be considered an attribute, including them in the analysis is challenging. At the same time, attachments can be considered redundant information without contribution to the process. So, we excluded them from the scope of this research.

An email body carries the message and intention of the sender to the receivers, recipients. A timestamp is a digital record of the time of occurrence of sending the emails.

Like any other human-generated data, the email's body text contains noise, irrelevant characters, symbols, and punctuations in this case. Therefore, preprocessing is needed to reduce noise and its effects by removing or changing them with more appropriate characters. Punctuations and symbols should be removed before training or testing the model. This process can vary depending on the emails' language and the domain, as there may be cases that special symbols can have special meanings. For instance, our data was in Persian, which does not require words to be transformed to capital or lowercase. At the same time, numbers contributed the least to our goal.

In parallel with the data cleaning step, a random sample of data from email bodies was explored by an expert agent to assign a proper label to each email body based on the workflow's steps and activities defined earlier. At the same time, the agent must be aware of the organizational environment that emails are communicated. For instance, the emails from the finance department contain contents, words, and acronyms that understanding them by a non-expert agent would be impossible. The consequent labeled dataset is used to train the classifier and fed to the model as the training data.

The other challenge with email data is the "reply," "reply-all," and "forward." These actions in email-based communication can make the dataset more complicated. In this project, the email database was designed to log every instance as a separate event and store them separately while connecting them through a LetterID, shared among a loop of emails. For situations where emails were stored in a different logic, appropriate approaches can handle this challenge and create an identification method to connect emails into a unique case number. Most email services add the sequence of such actions in the email thread. A log of previous communications is stored in the latest email message.
Additionally, the challenge of mixed messages in emails is widespread and can affect similar research. While this issue is prevalent, some researchers have investigated how to extract each sentence in an email and classify each into relevant classes. In this research, each email was related to only one activity due to our case study's communication style to avoid ambiguity in their organizational communication.

3.2 Email Classification

Text classification is one of the methods that have many applications in real life, like spam detection, and research. In this research, we use fastText to classify emails. fastText is an open-source, free, linear-based model and lightweight library that allows users to create text representations and text classifiers (Joulin *et al.*, 2017).

fastText uses a hierarchal softmax function that reduces the computational complexity, leading to a faster search for the predicted class (Joulin *et al.*, 2016) and works on standard, generic hardware (Alessa *et al.*, 2018). Unlike deep learning models, fastText is a standard tool that helps the results' transparency and reproducibility.

3.3 Process Mining

Process mining aims to extract information about event logs (Van Der Aalst *et al.*, 2004). Process Mining is a relatively new research discipline. It can be described as a bridge between data

mining and business process modeling (W. Van Der Aalst *et al.*, 2012).

In this step, we use the generated event log to discover the process models with the fuzzy Miner, an algorithm capable of mining unstructured behaviors in large event logs, with configurable output to reach the desired level of abstraction and proper visualization. Fuzzy Miner is one of the most valuable tools in case study applications which the data is from real-life and contains noises and complexities which must be handled through the miner's algorithm (Günther *et al.*, 2007).

In this research, we chose fuzzy Miner because of its ability to deal with unstructured processes due to abstraction and clustering techniques to produce understandable models from unstructured processes (Van Der Aalst *et al.*, 2007).

4 EXPERIMENTS

This section validated the designed methodology with the data exported from the case study company. The results and the discovered business process from the dataset of the procurement department have also been provided.

In this project, we focused on the procurement department's processes due to several reasons. First, the procurement processes were one of the most critical processes for the company to manage and improve. There were a few bottlenecks in the process that affected the flow of goods and services, and managers were interested in knowing more about them. Second, focusing on only a few processes was crucial to our research as communications in an organization with more than 100 employees can be challenging and complex. Finally, there were enough email threads to be extracted from the database. Employees had to follow up the process through emails from two years before this research.

The email log used in this research was extracted from the primary database of organizational correspondence. The dataset contained 100,000 rows of data across all the departments, each line representing an email sent from a sender to one or more recipients. To narrow down the emails to the procurement department's emails and, consequently, the related process, we filtered out emails concerning the procurement department by searching relevant emails that purchase experts were involved in at least one email, leading to a dataset of 1933 email instances. After the filtration, A dataset containing 1087 rows (56.2% of the total rows of the main dataset) was randomly sampled and labeled based on the existing workflow's activities. Figure 2 demonstrates the organization's purchasing process workflow that each purchase requester must follow to procure the materials through the procurement department.

The workflow depicts the process with ten main activities and two decision points. However, the activities in the workflow of the procurement department could not represent all the communications in emails. So, we had to expand the labels to 19 activities to cover them all. The labels' extension helped us to better capture the as-is state of the process in the log.

In the process of labelling, it is crucial to define each label clearly to maintain the alignment of the labels to the activities. Labels and their definitions are provided in Table 1.

While manual labeling was in progress, we also checked for cases that emails can include multiple activities. However, we could not find any email carrying two or more activities simultaneously.

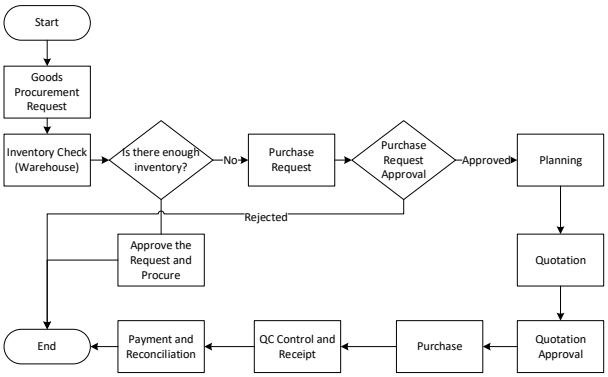


Figure 2: The process of procurement

Although the emails were only in Persian, we labeled them in English equivalent words. The labeling in English does not affect the process as the machine learning techniques can only process numbers, not words, in practice. This transformation from characters to numbers is done through the word processing libraries. In most cases, like this research, there is no need for translation, as the fastText supports Persian.

The dataset used had six main attributes, and their list with its description is provided in Table 2. A sample of the dataset is provided in Table 3.

As we studied the data, emails in the case study organization were only used to facilitate internal communication and process paperwork. So, it can be assumed that each email was only related to one activity.

To prepare the data for activity classification, we had to start with preprocessing of the data. All the steps from data preprocessing to label prediction were followed in the Python[5] environment in Google Colaboratory[6]. The steps of the preprocessing process were as follows:

1. Removing redundant symbols and punctuations ("!" #$%& ()*+,-;<=>?@[\]^_`{|}~<.*?>|&(['')
2. Removing redundant spacings
3. Removing numbers
4. Removing HTML codes

After the preprocessing step, the data needed to be prepared for supervised classification. fastText works with a specific configuration of data. Labels must be added to each email message of the training dataset as a prefix to the starting point of messages and the input (__label__ <Email Body>).

To build the classifier, we experimented with 15 different feature settings of fastText: Learning rate, embedding dimension, n-grams, and epoch. The best configuration was found as below:

1. lr (learning rate) = 0.9
2. embedding dimension=150

[5] https://www.python.org/

[6] https://colab.research.google.com/

3. epoch=50
4. n-grams= (2,10)

Table 1: Labels and their Definitions

#	Labels	Definitions
1	Comment	Informing others about comments on the process or activities conducted
2	Approval	Approval is given when direct managers are agreed to purchasing of a product or service
3	FYI	(For-your-information) Informing others in an email loop about an event (non-related to the primary process)
4	Approval Request	Requesting for Approval from direct managers
5	Prioritization Request	Requesting for prioritization for the previous activity
6	Payment Request Approval	Approval of a payment request
7	Payment Approval	Approving the initiation of the payment process
8	Payment	Finalization of the payment
9	Follow up	Follow up notification to ask for the latest state of the process
10	Purchase Request	Sending the purchase request to procurement officers
11	Out-Sourcing	Initiation of the out-sourcing process
12	Payment Request	Request for payment
13	Quotation	Quoting process from the market from a similar product
14	Planning	Reference for the planning process
15	Documentation (Receipt)	Documentation of the Goods/Service Receipts
16	QC approval request	Requesting for Quality Control Approval
17	Documentation (Invoice)	Documentation of the Invoices
18	QC Ok	Receiving the Approval of the QC team
19	Product Approval Request	Requesting for Product Approval from stakeholders

Table 2: dataset attributes and description

#	Attribute	Description
1	LetterID (LID)	A case number that connects an email loop
2	Employee ID Sender (EIDS)	A unique number that is assigned to each employee and is logged when he/she sends an email
3	Employee ID Receiver (EIDR)	A unique number that is assigned to each employee and is logged when he/she receives an email
4	Receive Date (RD)	The sent date of the email
5	Email Body (EB)	The text of the email body

Table 3: A sample of the data

LID	EIDS	EIDR	RD	EB
81177	7557	6448	2018-05-28 15:44:58	باسلام احتراما جهت استحضار
81177	8448	8532	2018-05-29 09:37:26	باسلام لطفا پس از مذاکره اقدام شود باسلام احتراما مراتب مورد تایید
81177	8448	7428	2018-05-29 11:50:25	است خواهشمند است دستور مقتضی صادر فرمایید

With this setting, the model was trained with a training dataset containing 900 rows of data and then tested with a dataset consisting of 187 rows. The model could achieve the best training accuracy of 98.8% and validation accuracy of 85.8%.

Then this model was used to predict (label the unlabelled data) the remaining dataset to create the final event log. To better understand the generated event log and its distribution in the dataset, we provide the distribution of the labels in the final dataset in Figure 3. This figure shows that most of the emails are only with the "Comment" label, meaning that senders tried to inform others about their comments on the process or activities conducted on the process. The second most repeated emails were about "Approvals," demonstrating another main email application in this organization.

The next step in our methodology is process mining, discovering the process model from the labeled dataset. We used Disco to apply the fuzzy Mining algorithm to discover the process model. The output of this step is illustrated in Figure 4.

We preferred to mine the event log with Disco; it has a user-friendly interface. Disco is a complete process mining toolkit from Fluxicon that facilitates process mining (Günther & Rozinat, 2012). We tried fuzzy Miner because of its ability to deal with unstructured processes due to the abstraction and clustering techniques and makes understandable models from unstructured processes (W. M.P. Van Der Aalst & Günther, 2007). Datasets can be imported in Disco in CSV or Excel, and the processes automatically mined through fuzzy Miner.

Based on the Disco's statistics, this dataset contained 1933 events in 204 cases, meaning 204 unique LetterIDs that were employed communicated with each other. The median duration of the processes was 18.2 days, and it took 54.1 days for the most extended case to complete. Seventy people were involved in generating this dataset, and only 81 cases could be captured from their start to the end.

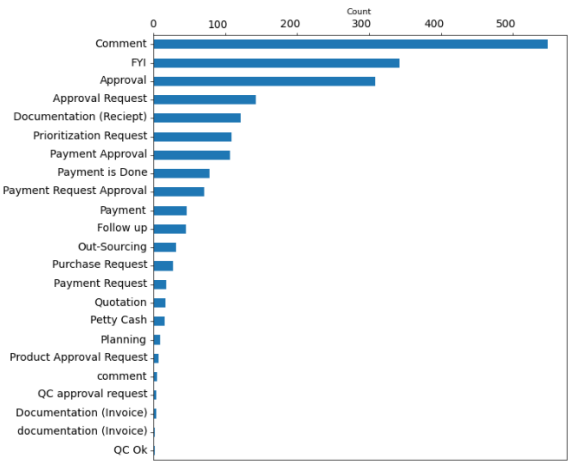


Figure 3: the distribution of the Activities after Activity classification

The discovered model, the output of fuzzy Miner in Disco, is presented in Figure 4 (appendix 1). This model has two primary components: activities in rectangle boxes and arrows as a flow between activities. The bolder and darker the boxes and arrows, the higher the frequency of execution in the process.

At first glance, the model is very complicated, hard to understand and follow, which depicts how complex was the communication in the case study department. Statistically, 57.27% of activities were not directly connected to the procurement process (figure 2). However, their role was crucial to transfer information. The frequency of such activities is illustrated through the darkness of colors in the model.

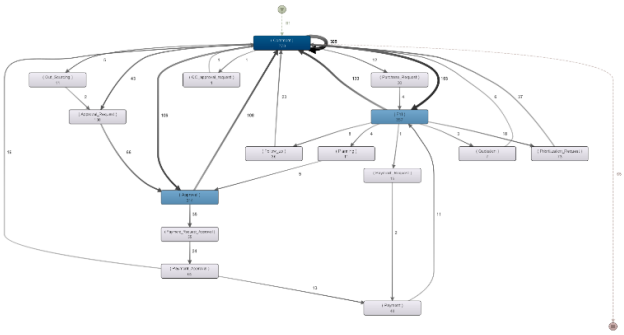


Figure 4: The Discovered process model based on the labeled email log

With more exploration, we figured out the procurement process was not aligned with the standard workflow in some cases, or only people tended to communicate at the start or a purchase request, and the rest were followed offline, without any digital record.

This discovery was beneficial for the procurement managers to highlight the role of such a system in their process regarding transparency and auditability of the procurement process.

The fuzzy Miner kept the patterns with higher frequency and omitted the less frequent ones with increased abstraction levels. Figure 5 is the consequent model from a higher abstraction of the process model with a cut-off of 55%. (Higher quality picture of the process model is provided in appendix 2.)

In this model, the "Approvals" activity was the critical activity that employees followed through the procurement process in their emails. At the same time, the rest of the workflow was not captured as frequently as this activity. So, there is a clear gap between the discovered process model and the designed workflow. Therefore, it can be inferred that the primary application of their email-based communication is only to communicate managerial approvals and decisions about the procurement process.

5 CONCLUSIONS

In this paper, we proposed a methodology to discover business processes from email logs. We combined the supervised text classification technique with fastText and the fuzzy mining to mine processes inside the consequent event log.

We tested the methodology with real-life data of a case study organization. To show its practicality, we focused on procurement processes.

In this case study, we were able to show that there were gaps in the process, from what was followed in the emails and the designed procurement process. On the other side, as the transparency of the process was not also evident for the procurement managers, this method helped them have a better understanding of how the procurement process is followed in emails.

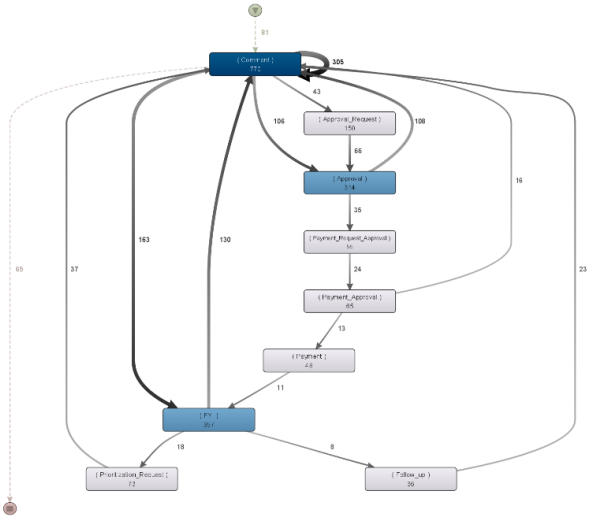


Figure 5: Higher abstraction of the discovered process model with the cutoff of 55%

Additionally, we could discover that emails were mostly about receiving approvals from the managers, as a matter of officiality, and not following the rest of the procurement process completely via emails.

With such a methodology, managers can be aware of the processes in unstructured communication environments and avoid possible process deviations. At the same time, they can monitor business processes and solve real-time bottlenecks. In our case, "Approval" activity could be considered a bottleneck, as the process's progress was highly dependent on its smooth execution.

In the future, we will try to experiment with more data and implement semi-supervised methodologies to classify emails better. We also suggest experimenting with pre-trained word embeddings like Glove[7] or any other word embedding tools. Text feature extraction techniques can also increase classification accuracy, which we suggest for further research.

Our proposed methodology is primarily limited in inflexibility to change the activities, which is the most critical drawback of supervised machine learning. The model can be trained through time intervals with new datasets to solve this challenge. Like Laga *et al.* (2019), there can be a platform to facilitate the labeling process with the help of users.

During this research, we figured out that many factors affect the organization's communications, making process discovery very challenging through the emails generated by the employees, so this topic can also be investigated in future research

[7] https://nlp.stanford.edu/projects/glove/

REFERENCES

[1]. Alessa, A., Faezipour, M., & Alhassan, Z. (2018). Text classification of flu-related tweets using FastText with sentiment and keyword features. *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 366–367. https://doi.org/10.1109/ICHI.2018.00058

[2]. Banziger, R., Basukoski, A., & Chaussalet, T. (2019). Discovering Business Processes in CRM Systems by Leveraging Unstructured Text Data. *Proceedings - 20th International Conference on High Performance Computing and Communications, 16th International Conference on Smart City and 4th International Conference on Data Science and Systems, HPCC/SmartCity/DSS 2018*, 1571–1577. https://doi.org/10.1109/HPCC/SmartCity/DSS.2018.00257

[3]. Borg, A., Boldt, M., Rosander, O., & Ahlstrand, J. (2021). E-mail classification with machine learning and word embeddings for improved customer support. In *Neural Computing and Applications* (Vol. 33, Issue 6). https://doi.org/10.1007/s00521-020-05058-4

[4]. Corston-Oliver, S., Ringger, E. K., Gamon, M., & Campbell, R. (2004). Task-Focused Summarization of E-mail, BT - ACL-WS2004A. *Proc. of ACL Workshop'04*, 1–8. https://aclanthology.org/W04-1008.pdf

[5]. Di Ciccio, C., Mecella, M., Scannapieco, M., Zardetto, D., & Catarci, T. (2012). MailOfMine - Analyzing mail messages for mining artful collaborative processes. *Lecture Notes in Business Information Processing*, *116 LNBIP*, 55–81. https://doi.org/10.1007/978-3-642-34044-4_4

[6]. Günther, C. W., & Rozinat, A. (2012). Disco: Discover your processes. *CEUR Workshop Proceedings*, *936*, 40–44.

[7]. Günther, C. W., & Van Der Aalst, W. M. P. (2007). Fuzzy mining - Adaptive process simplification based on multi-perspective metrics. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4714 LNCS*, 328–343. https://doi.org/10.1007/978-3-540-75183-0_24

[8]. Jlailaty, D., Grigori, D., & Belhajjame, K. (2017). Business Process Instances Discovery from Email Logs. *Proceedings - 2017 IEEE 14th International Conference on Services Computing, SCC 2017*, 19–26. https://doi.org/10.1109/SCC.2017.12

[9]. Jlailaty, D., Grigori, D., & Belhajjame, K. (2019). On the elicitation and annotation of business activities based on emails. *Proceedings of the ACM Symposium on Applied Computing, Part F1477*, 101–103. https://doi.org/10.1145/3297280.3297534

[10]. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models. *ArXiv:1612.03651 [Cs]*, 1–13. http://arxiv.org/abs/1612.03651

[11]. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, *2*, 427–431. https://doi.org/10.18653/v1/e17-2068

[12]. Koller, A., & Searle, J. R. (1970). Speech Acts: An Essay in the Philosophy of Language. *Language*, *46*(1), 217. https://doi.org/10.2307/412428

[13]. Laga, N., Elleuch, M., Gaaloul, W., & Smaili, O. A. (2019). Emails analysis for business process discovery. *CEUR Workshop Proceedings*, *2371*(2), 54–70.

[14]. Mavaddat, M., Beeson, I., Green, S., & Sa, J. (2011). Facilitating Business Process Discovery using Email Analysis. *BUSTECH 2011 : The First International Conference on Business Intelligence and Technology*, *c*, 40–44. http://www.thinkmind.org/index.php?view=article&articleid=bustech_2011_2_30_90043

[15]. Shing, L., Wollaber, A., Chikkagoudar, S., Yuen, J., Alvino, P., Chambers, A., & Allard, T. (2019). Extracting Workflows from Natural Language Documents: A First Step. In *Lecture Notes in Business Information Processing* (Vol. 342, pp. 294–300). Springer International Publishing. https://doi.org/10.1007/978-3-030-11641-5_23

[16]. Turner, C. J., Tiwari, A., Olaiya, R., & Xu, Y. (2012). Process mining: From theory to practice. *Business Process Management Journal*, *18*(3), 493–512. https://doi.org/10.1108/14637151211232669

[17]. Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., Van Den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., De Leoni, M., … Wynn, M. (2012). Process mining manifesto. *Lecture Notes in Business Information Processing*, *99 LNBIP*(PART 1), 169–194. https://doi.org/10.1007/978-3-642-28108-2_19

[18]. Van Der Aalst, W. M.P., & Günther, C. W. (2007). Finding Structure in Unstructured Processes: The Case for Process Mining. *Proceedings - 7th International Conference on Application of Concurrency to System Design, ACSD 2007*, 3–12. https://doi.org/10.1109/ACSD.2007.50

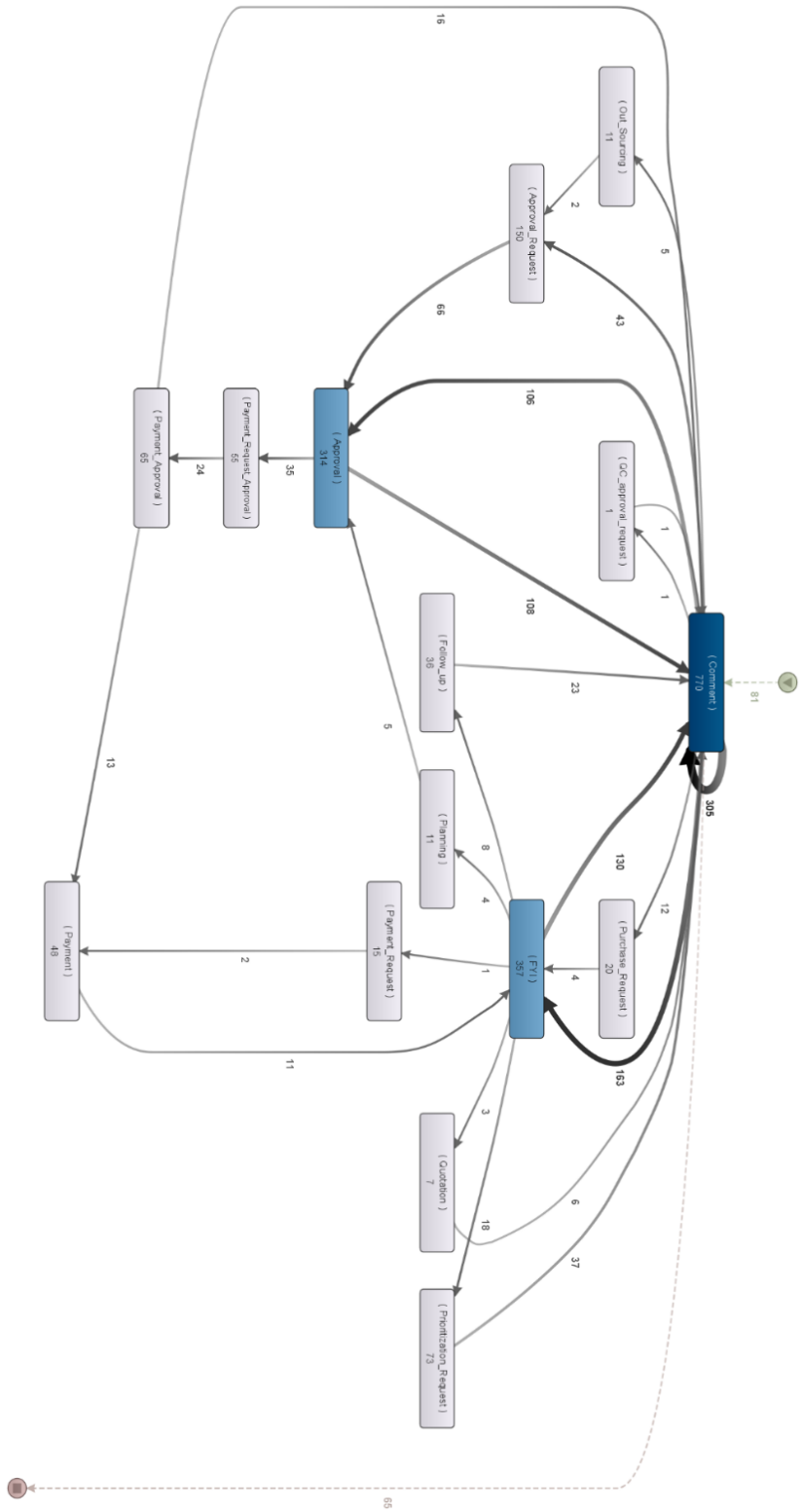
[19]. Van Der Aalst, W., Weijters, T., & Maruster, L. (2004). Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1128–1142. https://doi.org/10.1109/TKDE.2004.47

[20]. van der Aalst, Wil M P, & Nikolov, A. (2007). EMailAnalyzer: An E-Mail Mining Plug-in for the ProM Framework. *BPM Center Report BPM-07-16*, *August*, 1–26.

[21]. Van Dongen, B. F., Alves De Medeiros, A. K., & Wen, L. (2009). Process mining: Overview and outlook of Petri net discovery algorithms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 5460 LNCS* (pp. 225–242). Springer. https://doi.org/10.1007/978-3-642-00899-3_13

APPENDIX 1

The Discovered process model based on the labeled email log (Figure 4)



APPENDIX 2

Higher abstraction of the discovered process model with the cut-off of 55% (Figure 5)