Discovering Business Processes from Email Logs using fastText and Process Mining

Yaghoub Rashnavadi[1], Sina Behzadifard[2], Reza Farzadnia[3], Sina Zamani[4]
Kharazmi University[1,2,4], Oil Turbo Compressors[3]
March 2020

I. Abstract

Communication has never been more accessible than today. With the help of Instant messengers and Email Services, millions of people can transfer information with ease, and this trend has affected organizations as well. There are billions of organizational emails sent or received daily, and their main goal is to facilitate the daily operation of organizations. Behind this vast corpus of human-generated content, there is much implicit information that can be mined and used to improve or optimize the organizations' operations. Business processes are one of those implicit knowledge areas that can be discovered from Email logs of an Organization, as most of the communications are followed inside Emails. The purpose of this research is to propose an approach to discover the process models in the Email log. In this approach, we combine two tools, supervised machine learning and process mining. With the help of supervised machine learning, fastText classifier, we classify the body text of emails to the activity-related. Then the generated log will be mined with process mining techniques to find process models. We illustrate the approach with a case study company from the oil and gas sector.

Keywords: *Process Mining, Business Process, fastText, Email logs, Process Model Discovery*

II. Introduction

Email is playing a crucial role in our daily lives with over 124.5[1] Billion of business emails sent and received daily, the importance of such a tool cannot be ignored, especially in the context of business. Although the use of Email is mostly for personal cases, many organizations use this convenient tool to facilitate both internal and external communications and even manage complex projects through a large team.

While organizations are executing their business processes, employees are generating a significant amount of unstructured natural language documents [1]. With a simple analysis, we can understand that they are following a business process that is not explicitly mentioned.

The lack of invisibility over the process that has followed can cause intentional or unintentional deviation from the central business processes. It may result in losing the competency of the organization.

So, we hypothesize that the implicit business process and workflows followed in an Email loop can be discovered with mining the unstructured textual data that employees have generated through their daily communication. The objective of this paper is to propose an approach to discover underlying business processes by combining a fastText classifier and process mining and test this approach on a real dataset from an organization.

This paper is organized as follows. In Section 2, related works and literature will be reviewed, and the main differences discussed. In Section 3, we will introduce the method and the approach proposed and Section 4, we continue with a real experiment with the data from the case study organization, and the result will be demonstrated in section 5.

III. Related Works

The problem of process discovery is one of the popular topics among researchers in the business process management domain. To solve this issue, they are trying to find solutions with algorithms that can synthesize a process model from event logs [2]. The research area of process mining deals with techniques that are designed to extract knowledge from Event Logs [3], and this information can be illustrated as process models. These techniques provide new tools for a wide range of usage, Process Discovery, Monitoring, and Enhancement. These tools have many applications, and the most common use case is to facilitate the process alignment and bottlenecks analysis while predicting the problems in the execution of processes [4].

Extracting workflows with the use of NLP and sequence mining techniques is also another area that researchers tried to explore with unstructured texts from Emails like Shing et al. [1]. However, few research papers are targeting this subject.

In 2007, Aalst et al. [5] presented a tool for ProM (Process Mining Tool), EmainAnalyzer. This tool analyzes and transforms email messages in MS-Outlook into a format that can be used in process mining tools. The main goal of this research was to create a social network map from Email logs.

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

doi:10.20944/preprints202005.0007.v1

Banziger et al. [6] have investigated the use of unsupervised machine learning to detect and assign labels to activities automatically. The unsupervised machine learning algorithm had been tested over the data from a CRM (Customer Relationship Management) System.

Jlailaty et al. [7] have also tried to address the same problem with unsupervised clustering machine learning techniques to automatically label Emails with a related activity and mine the respected process model.

In machine learning, many approaches can be used to classify text content. Rosander and Ahlstrand [8] investigated the email classification with Long-Short term Memory Networks in Customer Service to improve the matching process of Emails and Customer Service Agents.

IV. Methods and Approach

The main objective of this research is to propose a framework that can extract activities from Email Corpus and mine the process models behind them, which can help stakeholders to understand the dynamics between them. This framework consists of 3 main steps: Data Preprocessing, Activity Discovery, and Process Mining.

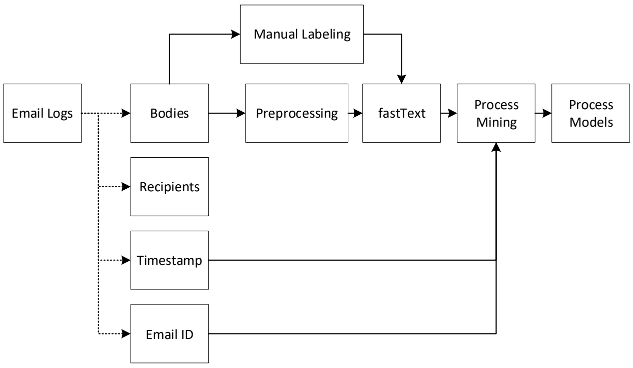

Figure 1: the proposed framework in 3 steps

A. Data collection and preprocessing

Email logs contain four main attributes, Body, Subject, Recipients, and timestamp. While attachments of the Email can also be considered as the main attributes, this research, we assume attachments are redundant. Body of Email is a text data that carries the main message and intention of the sender to the receivers. Recipients are a list of people involved in receiving or sending the Emails. The timestamp is a digital record of the time of occurrence of sending the Emails.

The body text of the Email, like any other Human Generated Data, contains many noises and irrelevant information. In order to remove such noises, a structured process is needed to reduce their effect by removing or changing them with the desired character. Punctuations and Symbols are most of these noisy data, and all such data should be removed before training or testing the model.

In this step, a sample data from Email Bodies were investigated by an expert human agent that his/her primary responsibility is to assign a proper label to each email body based on the workflow's steps and activities defined earlier. This dataset will be used to create the classifier.

B. Email Classification

Text Classification is one of the challenges that have many applications in real life. In this research, we use fastText[2] to classify Emails. fastText is an open-source, free, linear based model, lightweight library that allows users to create text representations and text classifiers. fastText uses a hierarchal softmax function that reduces the computational complexity, leading to a faster search for the predicted class [9]. It works on standard, generic hardware with better accuracy and faster [10] than similar deep neural networks.

C. Process Mining

The goal of process mining is to extract information about processes from event logs [11]. Process Mining is a relatively new research discipline and can be assumed as a bridge between data mining and business process modeling [3]. In this research, we use the classified email log to discover the process models inside the log with the fuzzy Miner. Fuzzy Miner is a flexible process mining technique that can adaptively simplify mined process models, and It is one of the most useful tools in case study applications [12]. The process of process mining has followed in Disco, which is a complete process mining toolkit from Fluxicon that facilitates process mining [13]. The other reason that the fuzzy Miner is suitable [14] for this approach is the ability to deal with unstructured processes due to the use of abstraction and clustering techniques and attempts to make more understandable models from unstructured processes. Datasets can be imported in Disco in CSV or Excel, and the processes will be automatically discovered through fuzzy Miner.

V. Experimental Design

In this section, we present the designed experiment over the case study organization[3].

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. So, in order to limit the Emails to a specific department and consequently the related process, we filtered out Emails concerning the "Procurement Department" by

[2] https://fasttext.cc

[3] Pars Investment Casting, https://www.otec-co.com/

searching relevant emails that Purchase experts who were involved in at least one Email. A sample of 1087 rows of data was labeled based on the main designed activities in the existing workflow.

The below figure demonstrates the workflow of the purchasing process in the organization that each purchase requester needs to follow to procure the material needed through the procurement department:

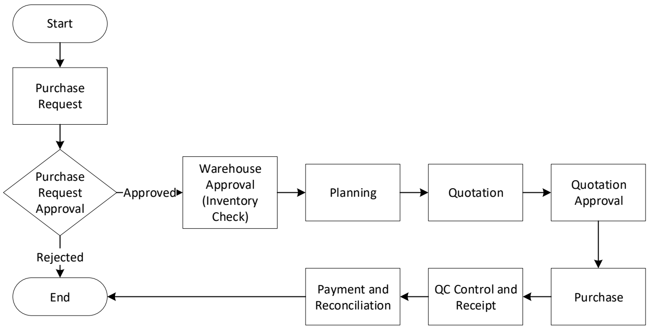

Figure 2: The process of procurement

The workflow models the process with nine main activities. However, as the above activities could not represent all the communications carried in Email, we had to expand the activities to 18, and this expansion helped us to follow the as-is state of processes in the log.

Then we preprocessed the final log to be fed to the classifier. To build the classifier model, we experimented with 15 different feature settings of fastText, with the parameter values of lr (learning rate) = 0.9, epoch=50, n-grams= (2,10). The preprocessed dataset was trained with 900 rows of data, tested with 176 rows with an accuracy of 98.5%, and validation of 85.22%.

After classification, all the dataset was labeled with the activities, and the log is ready for the next step, process mining.

In this step, we used Disco to apply the fuzzy Miner to the log and extract process models. The output of this step is illustrated as below:

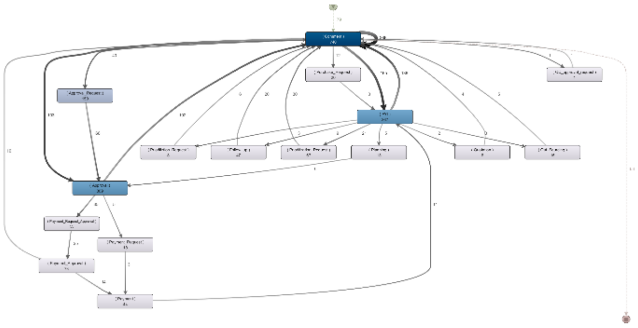

Figure 3: The Discovered process model based on Email Log with the fuzzy Miner in Disco

VI. Results & Discussion

In this dataset, there were 1933 events in 204 cases. The median duration was 18.2 days, and it took 54.1 days for each case to complete. Sixty-nine people were involved in generating this dataset, and all the bodies of Emails were in Persian.

At first glance to fig. 3, the discovered model seems to be complicated, which shows how complex communication can be in an organization. Based on the information mined from the email log, 57.27% of activities were not directly connected to the procurement process, but their role is crucial to transfer information through the process. The frequency of such activities is illustrated through the darkness of colors.

With more exploration in the results of the mining, we figured out that there are other processes that people follow, which are not followed the standard workflow or only communicated to initiate or close a process while following the process offline.

With an increase in abstraction by the fuzzy Miner, we can see that “Approvals” are the central part of the workflow that employees follow in this organization, and the rest of the workflow is not logged as much these activities.

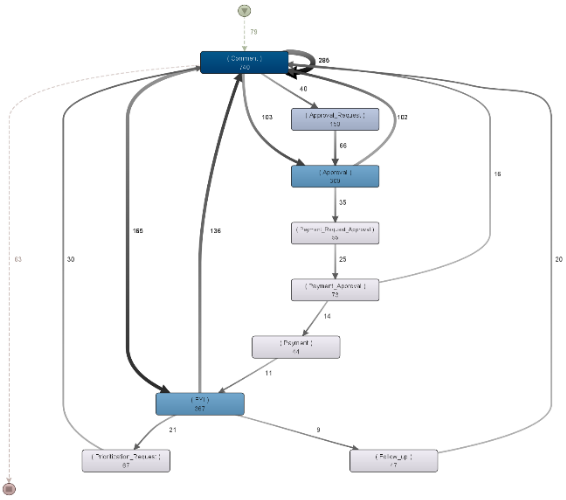

Figure 4: Higher abstraction of the discovered process model with the cutoff of 53%

VII. Conclusion

In this paper, we proposed a novel framework for business process discovery from the Email logs of an organization. To accomplish this goal, we combined the supervised text classification technique with fastText and the fuzzy Miner for process mining. The main objective of this research was to investigate the feasibility of extracting the implicit process models from Emails using the supervised text classification. We experimented with the framework over a case study to show its practicality, and the information mined showed it could be beneficial for organizations with Emailing infrastructure to monitor the process and improve them based on the discovered knowledge.

With this framework, managers can be aware of the real processes that are conducted and avoid possible deviations from the designed processes, while they can monitor the business processes and solve real-time bottlenecks. This capability can help them to take measures, solve pain-points, and increase the level of business processes agility.[15]

In the future, we will try to experiment with more data and try to implement semi-supervised methodologies to classify Emails.

During this research, we figured out that many factors are affecting the communications in the organization, which makes process discovery very challenging with analyzing contents generated by them, so this topic can also be investigated with enthusiastic researchers.

VIII. References

[1] L. Shing *et al.*, “Extracting Workflows from Natural Language Documents: A First Step,” in *Business Process Management Workshops*, vol. 342, F. Daniel, Q. Z. Sheng, and H. Motahari, Eds. Cham: Springer International Publishing, 2019, pp. 294–300.

[2] B. F. Van Dongen, A. A. De Medeiros, and L. Wen, “Process Mining: Overview and outlook of Petri net discovery algorithms,” in *Transactions on Petri Nets and Other Models of Concurrency II*, Springer, 2009, pp. 225–242.

[3] W. Van Der Aalst *et al.*, “Process mining manifesto,” in *International Conference on Business Process Management*, 2011, pp. 169–194.

[4] “Process mining: from theory to practice | Business Process Management Journal | Vol 18, No 3.” [Online]. Available: https://www.emeraldinsight.com/doi/abs/10.1108/14637151211232669. [Accessed: 20-Apr-2019].

[5] W. M. van der Aalst and A. Nikolov, “EMailAnalyzer: an email mining plug-in for the ProM framework,” *BPM Center Report BPM-07-16, BPMCenter. org*, 2007.

[6] R. B. Banziger, A. Basukoski, and T. J. Chaussalet, “Discovering Business Processes in CRM Systems by leveraging unstructured text data,” presented at The 4th IEEE International Conference on Data Science and Systems (DSS-2018), Exeter, UK, 2019.

[7] D. Jlailaty, D. Grigori, and K. Belhajjame, “Business Process Instances Discovery from Email Logs,” in *2017 IEEE International Conference on Services Computing (SCC)*, 2017, pp. 19–26, DOI: 10.1109/SCC.2017.12.

[8] O. Rosander and J. Ahlstrand, *Email Classification with Machine Learning and Word Embeddings for Improved Customer Support*. 2018.

[9] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, “FastText.zip: Compressing text classification models,” *arXiv:1612.03651 [cs]*, Dec. 2016.

[10] A. Alessa, M. Faezipour, and Z. Alhassan, “Text Classification of Flu-Related Tweets Using FastText with Sentiment and Keyword Features,” in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 2018, pp. 366–367, DOI: 10.1109/ICHI.2018.00058.

[11] W. Van der Aalst, T. Weijters, and L. Maruster, “Workflow mining: Discovering process models from event logs,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1128–1142, 2004.

[12] C. W. Günther and W. M. Van Der Aalst, “Fuzzy mining–adaptive process simplification based on multi-perspective metrics,” in *International conference on business process management*, 2007, pp. 328–343.

[13] C. W. Günther and A. Rozinat, “Disco: Discover Your Processes.,” *BPM (Demos)*, vol. 940, pp. 40–44, 2012.

[14] W. M. van der Aalst and C. W. Gunther, “Finding structure in unstructured processes: The case for process mining,” in *Seventh International Conference on Application of Concurrency to System Design (ACSD 2007)*, 2007, pp. 3–12.

[15] R. L. Raschke and J. S. David, “Business Process Agility,” p. 7, 2005.