Applied Computer Systems
ISSN 2255-8691 (online)
ISSN 2255-8683 (print)
2025, vol. 30, no. 1, pp. 1–11
https://doi.org/10.2478/acss-2025-0001
https://content.sciendo.com

Method for Creating Domain-Specific Dataset Ontologies from Text in Uncontrolled English

Shokoufeh Salem Minab[1*], Erika Nazaruka[2]
[1,2] *Institute of Applied Computer Systems, Riga Technical University, Riga, Latvia*

***Abstract* – Automated understanding of activities in enterprises is challenging due to a lack of domain specifications and a lack of domain ontologies. The goal of this research is to develop a method to extract elements of domain-specific processes from textual documents in unstructured English and form domain dataset ontologies. In order to achieve the goal, the related work on discourse analysis and business process modelling have been considered. The prominent technologies for implementation of the proposed method are machine learning, including classification algorithms and natural language processing using a large language model. The first experimental results are presented, and further research is discussed. Potentially, the method proposed can be implemented as a part of some assisting tool for system analysts and can support an analysis of the domain-specific information by providing contextual information from this and potentially related domains.**

***Keywords* – Business process modelling, domain analysis, natural language processing, ontology.**

I. Introduction

Business process modelling is one of the means that help in analysing enterprises' processes and reflecting activities of enterprises. It helps in achieving the common understanding of business processes by showing all the details of activities that enterprises use [1]. Usually, this process is performed manually due to the insufficient automated support of analytical tasks of information system developers and business (or system) analysts and leads to increased costs in the information system development [2], ends in poor quality of results [3], [4], and takes up to 60 % of overall time in the project workflow [5].

Natural languages are used to communicate information orally and in documents. Therefore, a piece of information on the enterprise business processes can be understood by analysing its documents [6]. Text can be represented in any form of written text, such as descriptions, emails, feedback, posts, etc. Finding the knowledge becomes challenging due to ambiguity of descriptions and different writing styles; however, the use of machine learning and natural language processing techniques to determine the domain and domain elements from the text of different types is promising [7].

There are different modelling languages for specifying business processes to show who is doing the actions, what actions are to be taken [8], what current situation is and what will be the next action [9], which represent business processes in various situations and make more similarity in illustrating different situations in confronting the real world. However, not any machine learning (ML) model can infer such elements, and experts' knowledge to derive these elements is needed. Experts should analyse the text to overcome difficulties of automated processing of elements such as complex relationships among concepts, incomplete natural language rules for conceptual model extraction, complex semantic relationships in natural language text, different solutions to the same problem, and natural language specification problems [3]. The goal of this research is to develop a method to process elements of domain-specific processes from textual documents in unstructured English and create corresponding domain dataset ontologies. This research integrates natural language processing techniques to acquire the domain dataset ontology from unstructured written text that can be used further to assist in developing business process models.

The rest of this paper is organised as follows. Overview of the related work is presented in Section 2. Section 3 introduces the proposed method in detail. Section 4 presents the results of experiments, and concluding remarks are presented in Section 5.

II. Related Work

Our work is mainly related to discourse analysis and business process modelling.

A. Discourse Analysis

Discourse analysis is one of the most important tools in text processing because it allows people to understand meaning from the context and create proper communication. It helps perceive the goal of speakers and create the beliefs and actions of the listeners. Discourse analysis is needed to classify different sentences according to the domain and expand a domain-specific ontology with elements that come from the sentence.

Discourse occurs in different domains, and it is possible to analyse it from different perspectives [10] and different factors like historical or situational context and vocabulary choice [11]. There are some subtasks performed during the process of text classification. Various machine learning algorithms have been

* Corresponding author's e-mail: shokufe.salem@gmail.com
Article received 2024-12-09; accepted 2025-01-03

applied to text classification, but the process is often time-consuming and complicated by challenges in understanding the domain.

Finding the relationships in unstructured text requires using different natural language processing techniques; however, there are no high quality method, and all of the existing ones are semi-supervised [12]–[15]. Role of discourse analysis is important in management decision-making [16].

Different methods are used to classify the text and understand its features [17]–[19]. In [20] and [21], the importance of textual data in enterprises and companies is discussed to improve project management and the quality of services, and the authors' mentioned standard text mining processes are repeatable. The processes employ three parts, which are text pre-processing, text processing, and text analysis.

In [22], the connection between text classification and business intelligence is considered. To build a successful business, it is important to perceive that the data in the business are close to what is expected from them [23]. The authors propose a solution to combine text classification and business intelligence for labelling the text and helping with the management tasks in the enterprise. In [24], text matching according to similarity in two texts is considered, and the authors consider that semantics can be involved in any part of text processing and improve text classification effectiveness. As a result, text can be perceived, hidden meaning can be clarified, and, finally, it helps the process inside the business.

B. Business Process Models

A business process is a structured and repeatable set of activities designed to achieve a specific business goal. It involves the coordination of people, processes, information, and technology to deliver value to customers, stakeholders, or the organisation itself. According to definition in [25]: "Business process modelling describes the manual task of identification and specification of business processes". To ensure effective communication within the organisation, stakeholders can use comprehensive business process models in the enterprises. In [26], the authors propose a new method to extract the business process model from the document. The proposed method is based on syntactic analysis and extracts the "subject-verb-object" construct (SVO). The proposed method saves time and effort in analysing the document, and, because of syntactic and semantic analysis, it can generate a more accurate model. However, this method heavily depends on the structure of documents, and an unclear and complex structure may not work correctly. Secondly, it needs to be compared with other methods to ensure its effectiveness. Thirdly, it needs a higher level of expertise to interpret and analyse the document. In [27], the authors use machine learning algorithms to predict the path in the business process model. The prediction suggests what will be the next event and what elements affect it. The authors use deep learning to train the ML model from event logs.

There are different business process modelling languages; however, Business Process Model and Notation (BPMN) is accepted as the standard to demonstrate the understanding of companies that shows the activities and connections inside the businesses. BPMN is understandable for stakeholders. It improves the comprehensive process of business better than other types of process modelling languages, such as Petri Net, Unified Modelling Language (UML), and Event Process Chains (EPC) [28]. As a result, in this research BPMN has been chosen.

In [29], the authors added the simulation parameters to the Business Process Modelling and Notation (BPMN) diagrams because they believed the use of BPMN in theory was different from its application in a real-world environment. According to [30], using a business process model in business process management is essential to lead the enterprise to success. However, there are some differences in the business process models used in the various enterprises according to the number of physical processes, like transportation of some materials. Some papers consider the use of business process management in a real-word environment [31] and adaptability of it in a real-word environment and show the importance of flexibility and readability of business process modelling in the business situation.

Considering all findings, providing a unique solution that can handle changes in practical business situations will save time and money and retain customers. Within the present research, first, a discourse analysis is performed in order to find out how people understand, embed, classify, and evaluate the text, which leads to an understanding of how to identify the domain. Second, it is evaluated how people use business process modelling, what features are important in business process modelling, and what method can be used to extract the knowledge from the text for a better analysis. The obtained information can result in choosing a suitable platform for business process modelling and gathering the elements of business processes.

III. Proposed Method

In this section, the proposed method is described in detail. The benefits of the classification task of a machine learning model and dataset ontology creation, which involves using ChatGPT and human interaction, are leveraged by a method for creating domain-specific process element ontologies.

This section aims at demonstrating the operation of the model. In brief, the method starts by checking the sentences to understand whether they are related to the text or not. Then, if they are related to text, it checks whether it is possible to find the elements in the ontology to suggest them to the user. Otherwise, the method asks ChatGPT about the elements, suggest them to the user, and asks them to be added to the ontology.

Figure 1 shows the flowchart diagram of the proposed method. The input data are unstructured English text (1). In the first part, some techniques such as pre-processing, vectorisation, and stacking classifier from machine learning are used to detect the classes. The output is corresponding clean sentences that are labelled to specify their class (2). These labelled sentences are input data for the next part, search through ontology to find elements of a business process model

(3), and the output of it is an expansion of the corresponding domain ontology, which specifies elements of business process models (4) (in this case, BPMN elements). Due to conceptual relationships and semantic richness, WordNet is considered a lexical ontology. It serves as a bridge between linguistic resources and formal ontologies; hypernym structure provides a hierarchy like a taxonomy, which is an essential part of an ontology. This research uses WordNet.

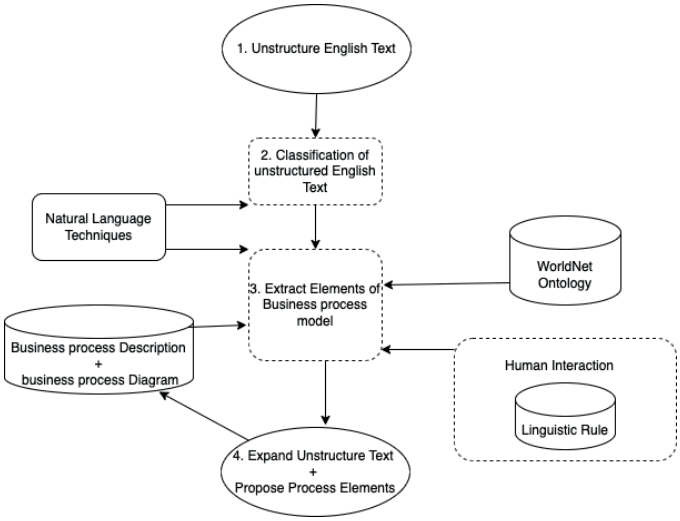

Fig. 1. Flowchart diagram of the proposed method (in general).

WordNet defines rich semantics, which resembles the properties found in ontologies. Also, it is based on shared linguistic knowledge in businesses, aligning with the idea of a common understanding.

This part uses the business process description, the business process diagram, and techniques of natural language processing. If the proposed method cannot match the input with any record according to the proposed elements, the method asks the human to propose a decision and add data to the dataset. As a result, during that time, the ontology will grow, and the ability to extend it will increase for a wide range of specific domains.

A. Dataset

To achieve the goal, a reliable dataset is required. Also, it should be noted that the dataset will be used for a further process, which is the expansion of the ontology with the elements of the proposed business process model.

"The AG's news topic classification dataset"[2] is used. This dataset is collected from more than 2000 news sources and more than one million news articles, which were gathered by an academic news search engine called ComeToMyHead. It consists of original text that is related to the 'world', 'sport', 'business', and 'science and technology' categories in two different files that are used for training and testing. In the train file, there are 120 000 sample records in four classes, and in each class, there are 30 000 samples. In the test file, there are 7600 records in four classes, and in each class, there are 1900 samples. As a result, the dataset is completely normalised.

B. Classification of Unstructured English Text

Figure 2 illustrates the classification of unstructured English text, i.e., Step 2 of the proposed method. In the beginning, input data are unstructured English text (1). To extract meaningful information and uncover the hidden patterns in text data, the next step of the method after collecting data will be the pre-processing of unstructured English text (2). Text data often contain noise, inconsistencies, and irrelevant elements that can affect the accuracy and effectiveness of the subsequent analysis. In this part, pre-processing consists of cleaning and transforming the raw text to give the dataset the necessary structure and coherence for further processing. By applying several pre-processing techniques, the accuracy of the ML model will improve. The pre-processing techniques that are used consist of the conversion to the lower case, tokenization, removal of stop words and stemming.

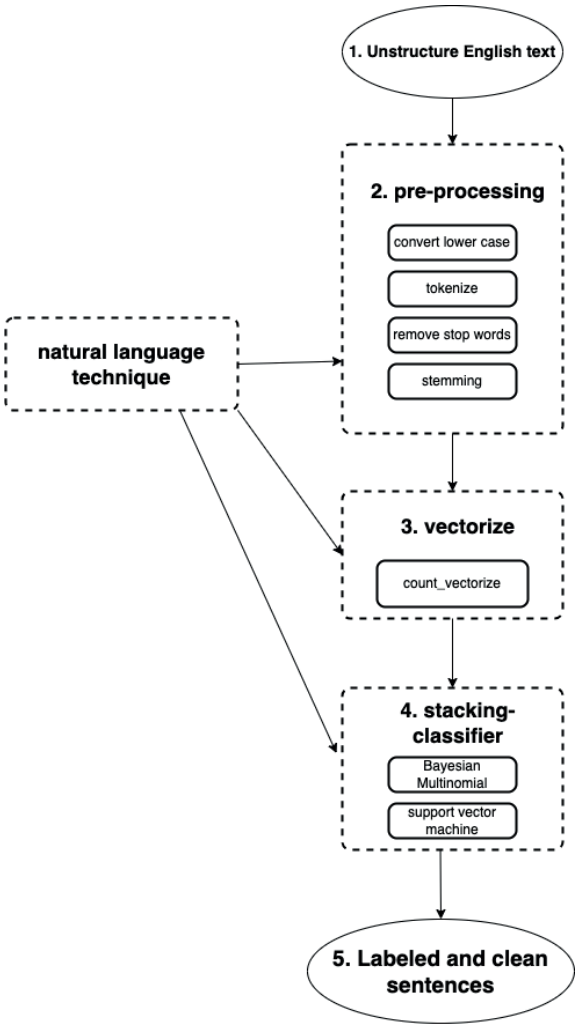

Fig. 2. Flowchart diagram of "classification of unstructured English text", i.e., Step (2) of the proposed method.

[2] The data are accessible at https://github.com/mhjabreel/CharCnn_Keras/tree/master/data/ag_news_csv.

Vectorisation is another important part after pre-processing (3). CountVectorizer is a class used to convert text data into numerical vectors, which is a basic form of word representation. This method is a way to convert a collection of text sentences into a matrix of token counts. Each sentence is shown by a row in the matrix, and each vocabulary is shown by a column. This part focuses on the process of generating vectors, where pre-processed text is transformed into meaningful numerical representations. The aim is to encode the text with rich semantic information that enables a deeper understanding of the underlying data in the form of text and expresses the semantic meaning and contextual relationships between words in the text [32].

Classification of data is the next step in extracting knowledge from unstructured data (4). In the proposed method, a stacking classifier, a machine learning approach for multi-classification task, is used. The stacking classifier is an ensemble learning technique that incorporates multiple base classifiers to improve predictive accuracy [33]. It is a type of stacking models where the predictions from individual base classifiers are used as features to train a higher-level "meta-model" that makes the final predictions [34]. The output will be cleaned and labelled data (5). Figure 3 shows the diagram of the stacking classification task.

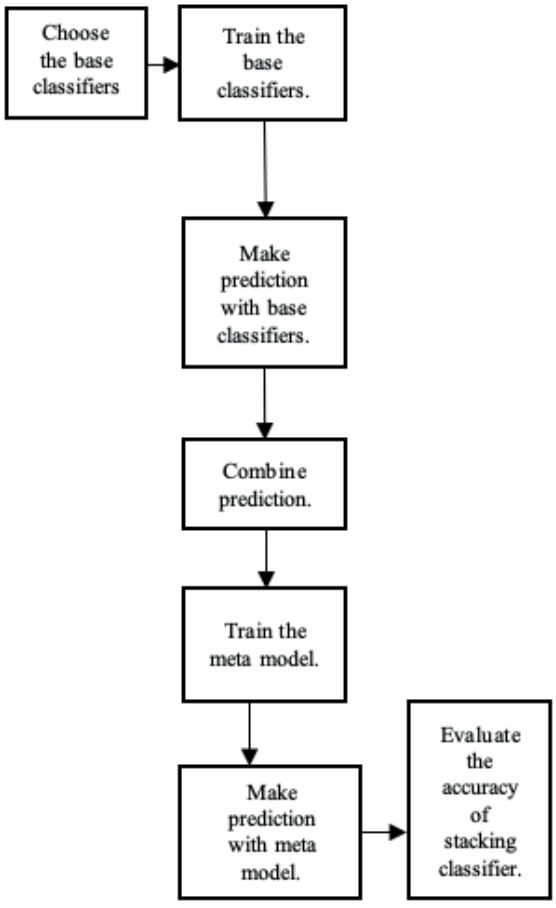

Fig. 3. Flowchart diagram of the stacking classification task.

The task that applies the stacking classifier starts with choosing the classifier, before using the dataset. In this part, dataset is split into three subsets: one subset is used to train the base classifiers, the second subset is used to validate the data, and the final subset is used to test the data. Two subsets of datasets, which are train and validate, are used in the base classifier to train and validate the model. Then a new dataset that contains the predictions from both base classifiers is created to validate the data. Features in the next phases involve training the meta-model on the new dataset and using the predictions from two base classifiers. Then, it is the final prediction, which means classifying the new, unseen text, combining prediction of base classifiers into a new dataset, and feeding it into a trained meta-learner. The meta-learner uses the combined predictions to make the final prediction.

It is important to choose the correct classifier to increase accuracy. In this work, different kinds of algorithms such as k-nearest neighbour, random forest, multinomial naive Bayes, and linear SVM have been applied, and the highest accuracy has been achieved through multinomial naive Bayes and linear SVMs. Figure 4 shows the results of the comparison between different algorithms on unstructured English dataset (AG's news topic).

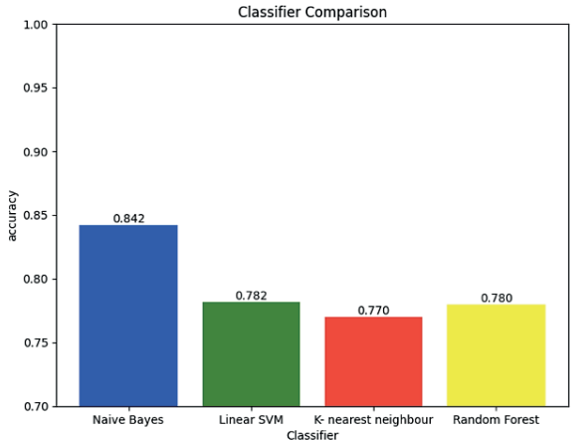

Fig. 4. Results of applying different algorithms to the dataset.

As a result, two classification algorithms, multinomial naive Bayes [33], [35], [36] and linear support vector machines [37], [38], [39], will be used as the base classifiers. By combining these two classifiers as base classifiers in the stacking classifier, the program leverages the strengths of both algorithms. Naive Bayes is likely to capture some unique patterns in the text data, while SVM can handle complex decision boundaries and capture additional patterns that may be missed by naive Bayes. Stacking the predictions of these two classifiers and training a meta-model on the stacked features allows for the exploitation of complementary information from the base classifiers, potentially improving the overall predictive performance on the business category dataset.

By applying pre-processing techniques and then a classification algorithm to the input data, a cleaned and labelled output will be achieved that is suitable for further processing.

The data consist of some features that are essential for achieving the goal, which is to process the text and extract elements of business processes in a specific domain and expand the ontology.

C. Extraction of Elements of Business Processes

The next step is the expansion of the dataset ontology with business process elements. Figure 5 illustrates the flowchart diagram for the next phase "Extract elements of business process model" (Fig. 1), i.e., Phase (3). In this part, the class identified and all sentences are related to the classes that is need according to Fig. 2. If a new sentence contains a concept that is in the ontology, the model searches for it and suggests it. If there is a new concept, then the sentence is parsed using prompt to ChatGPT to find the required BPMN elements, and it is suggested to a human to locate the elements in the ontology. Otherwise, the next sentence is taken, and the first phase repeats.

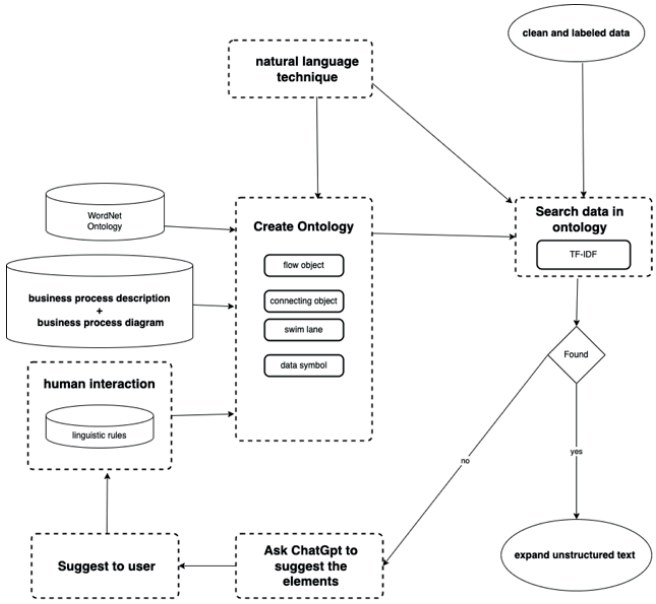

Fig. 5. Flowchart diagram of the "Extract elements of business process model" phase.

To propose the elements of BPMN, an ontology is needed. This ontology explains the concepts and relations within an industry and involves all definitions and knowledge related to the business environment within the domain.

Each company in the domain may have company-specific concepts. Therefore, the developed ontology will contain elements that may correspond to the company and may not. This ontology will give no benefit to a business analyst without supporting functionality. This functionality could be implemented in an assisting tool and would suggest possible elements of a process based on the context similarity.

This part of the method needs different inputs and techniques, such as business description, WordNet ontology, human interaction, and natural language processing.

In this example, BPMN is used as a standard industry notation because of the need to make decisions and understand the process without any constraints of the problem domain, which is the power of BPMN as indicated in [40].

There are four element types to show the business process model in BPMN [40]:

- Flow objects – they show the flow of the process, which consists of activities, events, and gateways:
  – An activity is a special task performed by a person or group of actors;
  – An event is a modification of the process like message, link, or compensation;
  – A gateway is the decision node according to events.
- Connecting objects (data objects and data stores) – they show how tasks are connected.
- Swimlanes (participants) – they show the major participants (an actor or a group of actors) in the process.
- Artefacts (text annotations and groups) – additional information for the process.

Another important input in this part is the WordNet ontology. WordNet is not a formal ontology in the strict sense. However, because of conceptual relationship it is regarded as a lexical ontology and lexical database that is used for linguistic knowledge. It has a set of synonyms with specific concepts for nouns, verbs, adjectives, and adverbs. WordNet shows the related words by creating a lexical relationship as well as proposing the meaning [25], [41], [42].

Despite the acceptable WordNet ontology and business description database, there is not any automated way to extract business process elements, which constitutes the ontology to be created. The reason is the ambiguity of words and that this is not enough to satisfy the syntax variables in the requirement specification [3].

In the last part of the method, the input data can lead to extension of the previously created ontology. The input is a set of labelled and cleaned sentences from the first phase and the ontology that was created in this part. If there is a sentence with similarity, the elements are to be shown to the user as an extension of business process elements; otherwise, ChatGPT is to be asked to suggest the elements, and then the user is asked about their correctness and, in case of the positive answer, elements from the sentence are to be added to the dataset ontology.

It is essential to keep in mind that because ChatGPT depends on probabilistic language modelling rather than domain-specific rules, it has limits that may result in suggestions that are inaccurate or irrelevant. Consequently, human involvement is vital for ensuring output accuracy and control. Users must evaluate the recommended features and determine if they match the planned business processes. Implementing rule-based filtering of the output could be a future solution to improve dependability and decrease the necessity for manual verification. With this method, recommendations might be automatically filtered according to pre-set criteria, increasing ontology extension's accuracy and efficiency.

Matching, in general, is determined by the size of the document, the complexity of the task, and the computational resources available for the task. TF-IDF (Term Frequency-

Inverse Document Frequency) is used in this work to match sentences based on their importance and relevance. It is a numerical statistic used in the natural language processing and information retrieval to measure the importance or relevance of a term (word) within a document or a collection of documents [43].

This method can be implemented by developers of an assisting tool for system analysts as a core component that allows extending dataset ontologies for concrete domains. The developed domain-specific ontologies could be used in this tool for implementing functionality that allows comparing elements of business process models created by a business analyst with the same category elements in an ontology of a specific domain. The result of the comparison could be indication of some lacking or additional actors / participants, activities, events, or objects and, therefore, could be a source for further discussion between a business analyst and a domain expert.

D. Implementation of the Proposed Method

In this part, the steps to implement and evaluate the proposed method are considered. The steps contain detailed information, outcome analysis, and descriptions regarding the tools and technologies that are used to implement the method. This section presents:

- Tools and technologies;
- Implementation of the code;
- Evaluation metrics;
- Evaluation and validation of classifier.

Tools and technologies: In this research, there are some parts comprised of pre-processing, classification of text, choosing the correct class, finding the similarity between each sentence and sentence from the document related to a business process, and proposing the correct elements of business process models, which lead to implementing and completing research in one domain. As a result, the Python language has been chosen for this research, and all details have been implemented from scratch in PyCharm.

Implementation of the code: According to the proposed method, with the intention of pre-processing the unstructured English sentences, classifying them, and then expanding them according to a business process diagram, the code generation is divided into three parts. The first part of the code is related to the pre-processing of the text and identifying the valuable features in each sentence in order to classify them. Then search through the ontology takes places to find the most similar data and suggest the elements if there are similar sentences inside the ontology. In the last part, if there are no any similar sentences in the ontology, it is required to find out from ChatGPT about them and ask the user to check and add them to the ontology.

Evaluation metrics: To assess the performance, five evaluation metrics have been selected to be calculated:

- Accuracy;
- Precision;
- Recall;
- F1 score;
- Run time.

The accuracy indicates the percentage of correctly predicted class labels out of the total number of samples in the test dataset. The accuracy provides an indication of how well the stacking classifier performs on the test dataset. Precision measures the accuracy of positive predictions. Recall measures the proportion of actual positive instances that are correctly predicted. F1 score is the equal means of precision and recall, providing a single metric that balances both precision and recall. Time is the last measure to consider, which refers to the actual time the method takes for an algorithm to complete a specific task [44].

Evaluation and validation of classifier: To validate the result, it is necessary to evaluate if the proposed method works as intended and to identify the weaknesses and limitations of the proposed method that need to be improved. Figure 6 shows the comparison of the accuracy of two base algorithms with the stacking classifier. Figure 7 shows the comparison of the precision of two base algorithms (naive Bayes, SVM) with the stacking classifier and visualises it. Figure 8 illustrates the comparison of the recall of the two base algorithms with the stacking classifier. Figure 9 shows the comparison of the F1 score of the two base algorithms with the stacking classifier.

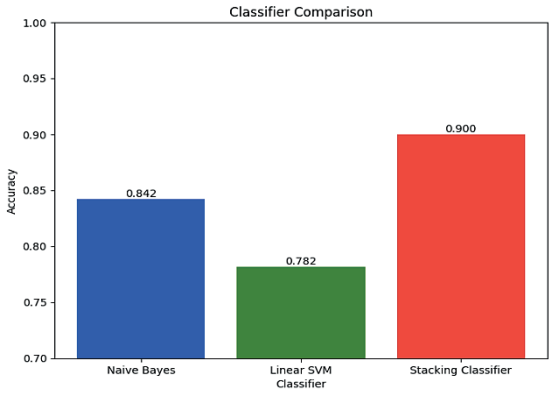

Fig. 6. Comparison of accuracy between stacking classifier and base classifier.

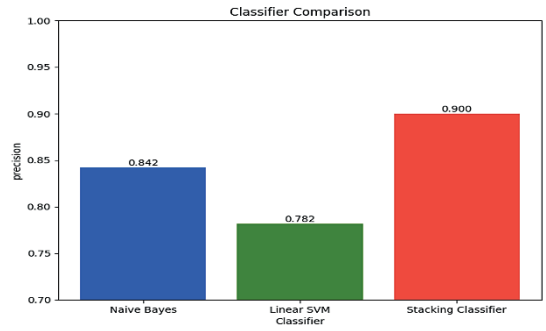

Fig. 7. Comparison of precision between stacking classifier and base classifier.

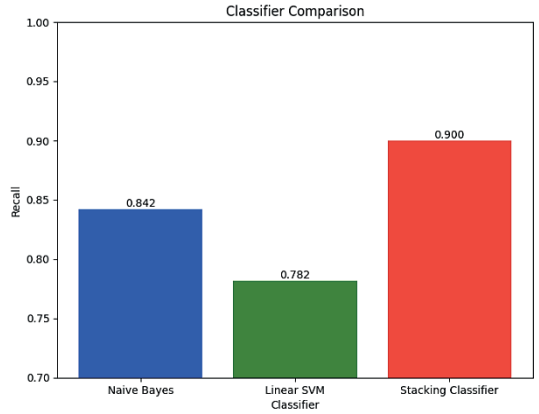

Fig. 8. Comparison of recall between stacking classifier and base classifier.

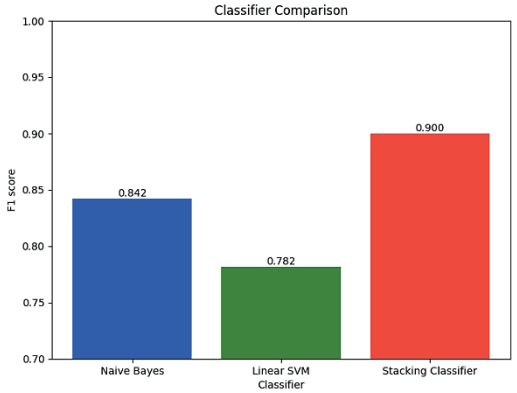

Fig. 9. Comparison of F1 score between stacking classifier and base classifier.

Processing time is the last metric in the classification task for this research, and the measure can be different in different contexts. Time often refers to the computational time or run time required for a model to make predictions on a dataset. The size of the dataset for training is 28.1 MB and for test it is 1.77 MB. Table I shows the comparison of time in different algorithms on the test dataset.

TABLE I
TIME OF DIFFERENT ALGORITHMS

Algorithm	Time (s)	Time (s per 1 MB)
Naive Bayes	0.03	0.001
Linear SVM	41.47	1.39
Stacking classifier	0.138	4.12

The results of classification in the proposed method highlight its superiority over the baseline algorithms. However, it is worth noting that the uniformity of results across accuracy, recall, and precision metrics indicates that the data normalisation process was highly effective. While this demonstrates the robustness of normalisation, it may reduce the informativeness of the evaluation, as it masks potential distinctions in classifier performance. Although the proposed method does not achieve the best time efficiency, priority was given to accuracy due to the nature of the data and the task offline processing requirements.

IV. EMPIRICAL RESULTS

This section provides the explanation of creating the ontology and presents how it can be used during the process. The results are also illustrated through an example.

A. Creation of Dataset Ontology

The business process description that used in this study represents a realistic case from a healthcare institution, specifically focusing on hospital registration. According to the analysis of the document, all elements of the BPMN diagram, which are the flow objects (events, activities, and gateways), connecting objects (sequence flows, message flows, and associations), swimlanes (the activity for a specific role), and data symbols (a certain type of data), are checked separately. ChatGPT was incorporated to generate the elements from the documents because of the reasons presented below [45]–[47].

- This part of research needed to augment the dataset to improve the robustness and generalisation of the work.
- ChatGPT has a better processing of the semantics within the context.
- ChatGPT analyses exploratory data, which is useful for discovering patterns or relationships in the text that might not be immediately apparent.
- ChatGPT can simulate human-like interactions, providing a more natural and intuitive way to collect information or feedback on the elements of interest, especially when there are interactions.

In certain industries or organisations, there may be specific compliance requirements that need to be considered. Checking all elements helps ensure that the business process model corresponds to these regulations and standards. It helps identify the issues in the modelling phase, prevent errors during the process, and find opportunities for process optimisation, which creates a comprehensive and reliable reference for process documentation and future decision-making. It also helps facilitate communication between business analysts, stakeholders, and developers. In general, checking all elements of a business process diagram is a crucial step to ensure the correctness, consistency, compliance, and overall quality of the process model. Table II shows the comparison of elements of BPMN in the sample diagram and extracted by ChatGPT.

TABLE II
COMPARISON OF ELEMENTS OF BPMN FROM DIAGRAM AND CHATGPT

Elements of BPMN	Elements from diagram	Elements by ChatGPT
Activity	18	30
Events	7	5
Gateways	7	2
SequenceFlows	31	26
Associations	3	1
MessageFlows	1	1
SwimLanes	0	8
Data Symbol	2	27

Analysing the elements of business process modelling is expert's duty, and different experts have different ideas for specifying the elements. According to the analysis of the results in Table II, ChatGPT suggests more elements in most of the text because it clarifies more details, and the expert suggests elements at the higher level of abstraction.

For example, one part of text is [48]: "The Intake process starts with a notice by telephone at the secretarial office of the healthcare institute. This notice is done by the family doctor of the person who is in need of mental treatment. The secretarial worker inquires after the name and residence of the patient. On basis of this information, the doctor is put through to the nursing officer responsible for the part of the region that the patient lives in. The nursing officer makes a full inquiry into the mental, health, and social status of the patient in question. This information is recorded on a registration form."

Thus, according to expert's opinion, there are three activities. However, according to ChatGPT suggestion, there are nine activities as shown below.

Expert's activities are:

- Answer notice;
- Record notice;
- Store and Print notice.

ChatGPT activities are:

- Notice by telephone at the secretarial office of the healthcare institute;
- Inquiry by secretarial worker for patient's name and residence;
- Transfer of call to nursing officer responsible for the patient's region;
- Full inquiry into mental, health, and social status by nursing officer;
- Recording information on a registration form;
- Handing in the registration form at the secretarial office;
- Storing form information in the information system and printing it;
- Creating a patient file for new patients;
- Storing the registration form and printout in the patient file.

As it is clear, ChatGPT suggests more details in every expert's activity. Although the result must be checked by experts, it will be possible to choose elements from the suggestions, which will make the process easier for experts.

B. Searching through Dataset Ontology

For processing the domain process elements in the text, two main previous phases are pre-requirement. First, it is necessary to classify the sentences to a correct group according to the business domain that is a category. Second, it is necessary to create an ontology for a specific business domain. After completing these two phases, it will be possible to choose different methods to match the output of them.

For our illustrating example, to create an ontology, there is a need for a business process diagram and its description, which is used from [48]. A business process description is needed to infer the elements of BPMN, and a diagram is needed to prove the elements. This book mentions business process diagrams and their descriptions for realistic cases of healthcare institutes in the Eindhoven region for elderly patients with problems, all of which relate to the registration of patients in the hospital. For each sentence that is received from the business description and diagram, all elements should be separated according to specific modelling concepts that are used in the real world.

If the new sentence matches any existing sentence in the ontology, the method suggests the sentence. Otherwise, if there is a new concept, then the sentence is parsed by using ChatGPT and it is suggested to a human to locate an element in the ontology where it is needed.

To use ChatGPT, GPT-3 API is used for natural language processing tasks. After that by formulating the question inside the function in program one can manually analyse the answer in output. Function is designed to interact with the OpenAI GPT-3 model (specifically the text-davinci-002 engine) to generate a response that explains the elements of the business process model and notation based on a given sentence.

C. Sample Experiments

The first part of the code is related to the pre-processing of the text and identifying the valuable features in each sentence to classify them and to propose the elements of BPMN to fill the gap about. Input sentences can be unstructured sentences of any type. For example:

"The patient's first meeting with the specialist left them feeling hopeful about their treatment options."

After finishing pre-processing, clean data are available for use as below.

"Patient first meet special left feel hope treatment option."

At this part, it will be possible to classify the sentences, with the possibility of 0.9; this sentence is classified to the third group, which is related to "business".

After classifying the sentence, searching through the ontology is started. One of the sentences in the ontology is: "The first intaker plans a meeting with the patient as soon as this is possible" and elements related to sentence are as below.

Flow object:

1. Start Event: This represents the initiation of the process.
2. Activity (Task): "Plans a meeting with the patient as soon as this is possible." This is the main task or activity performed by the first intaker.
3. End Event: Represents the completion of the process.

Connecting object:

1. Sequence Flow: Connects the Start Event to the Task and the Task to the End Event.

Swimlanes:

1. First intaker;
2. Patient;
3. Scheduling.

Another example can be a sentence like:

"Implementing a new patient scheduling system could help reduce wait times and improve efficiency at the hospital".

After pre-processing the sentence is changed to:

"Implement patient schedule system help reduce wait time improve efficient hospital".

After classifying the sentence, searching through ontology is started. However, there are no sentences with high similarity to this sentence. As a result, according to the method, it asks ChatGPT and suggests the user to check and add the sentence inside the ontology and expand the ontology.

Overall, the evaluation of the proposed method demonstrated its innovative integration of classification models and NLP techniques to expand domain dataset ontologies from unstructured English textual documents. Unlike previous studies, which often focus solely on text classification [20], [21] or specific aspects of business processes [22], [23], this method uniquely combines these elements into a cohesive framework, offering a more comprehensive approach to ontology development. Despite these advancements, some limitations warrant further investigation. The scalability of the approach to larger datasets or more complex business domains has not yet been extensively tested, leaving its performance in real-world scenarios as an avenue for future exploration. Additionally, reliance on machine learning models, such as ChatGPT, raises concerns about potential biases in training data and long-term viability due to external dependencies. While human verification ensures the accuracy of outputs, it introduces a potential bottleneck in large-scale applications. The multi-step nature of the method, though innovative, may pose challenges for organisations with limited technical expertise. However, these limitations are outweighed by the method's holistic approach and potential to advance the field, offering a solid foundation for future enhancements that could make the method even more accessible, scalable, and efficient.

V. Results and Conclusions

The proposed method leverages a stacking classifier, incorporating two primary base classifiers – multinomial naive Bayes and linear SVM. Through careful experimentation and evaluation, the results have demonstrated a significant improvement in accuracy compared to existing methods. The effectiveness of the stacking classifier demonstrates its potential applicability in real-world business cases. The heightened accuracy obtained through the integration of diverse classifiers, specifically multinomial naive Bayes and linear SVM, attests to the robustness of the proposed method.

Moreover, in the second part, the application of the classified sentences to match ontology represents an important step towards practical implementation. The utilisation of similarity metrics facilitates the integration of the proposed method into the business process model, offering a systematic and intelligent approach to handling unstructured text within this domain.

Also, utilising ChatGPT in the semi-automated method to suggest elements for a BPMN based on sentences offers an efficient path to developing a comprehensive BPMN and assistance to the expert. By utilising ChatGPT, experts can efficiently identify key elements, improve the efficiency of systems, optimise workflows, and enhance productivity and decision-making processes.

In summary, the integration of the stacking classifier, coupled with the matching of classified sentences to business process documents, not only validates the completeness and correctness of the proposed solution but also positions it as a viable tool for enhancing the efficiency.

In the context of future research, several possibilities can be explored to further improve the understanding and application of the method. Here are some suggestions, such as test the proposed method on a broader range of business domains to assess its generalizability and identify potential domain-specific challenges or optimisations, research dynamic adaptability of the proposed method to evolving requirements over time. This could involve the development of adaptive models that continuously learn from new data and adjust their predictions accordingly and investigate the potential of transfer learning techniques to leverage knowledge gained from one domain or industry for application in another. This can enhance the adaptability of the method across different business contexts.

References

[1] I. Moreno-Montes de Oca, M. Snoeck, H. A. Reijers, and A. Rodríguez-Morffi, "A systematic literature review of studies on business process modeling quality," *Inf. Softw. Technol.*, vol. 58, pp. 187–205, Feb. 2015. https://doi.org/10.1016/j.infsof.2014.07.011

[2] D. Orlovskyi and A. Kopp, "An approach to business process model structuredness analysis: Errors detection and cost-saving estimation," in *ICTERI 2021 Workshops*, O. Ignatenko *et al.*, Eds., in *ICTERI 2021 Workshops. ICTERI 2021. Communications in Computer and Information Science*, vol 1635. Springer, Cham, Sep. 2022, pp. 23–39. https://doi.org/10.1007/978-3-031-14841-5_2

[3] M. Omar and G. Baryannis, "Semi-automated development of conceptual models from natural language text," *Data Knowl. Eng.*, vol. 127, May 2020, Art. no. 101796. https://doi.org/10.1016/j.datak.2020.101796

[4] T. Rozman, G. Polancic, and R. V. Horvat, "Analysis of most common process modeling mistakes in BPMN process models," *EuroSPI'2007*, Sep. 2007. [Online]. Available: https://conference.eurospi.net/images/proceedings/EuroSPI2007-ISBN-978-3-9809145-6-7.pdf#page=30. Accessed on: Dec. 12, 2023.

[5] J. Herbst and D. Karagiannis, "An inductive approach to the acquisition and adaptation of workflow models," in *Proceedings of the IJCAI*, 1999, pp. 52–57. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=03d2b48ef05859066187782952977 0de93f30d51. Accessed: Dec. 13, 2023.

[6] M. Osborne and C. K. MacNish, "Processing natural language software requirement specifications," in *Proceedings of the Second International Conference on Requirements Engineering*, Colorado Springs, CO, USA, Apr. 1996, pp. 229–236. https://doi.org/10.1109/ICRE.1996.491451

[7] K. R. Chowdhary, "Natural language processing," in *Fundamentals of Artificial Intelligence*, K. R. Chowdhary, Ed. New Delhi: Springer India, Apr. 2020, pp. 603–649. https://doi.org/10.1007/978-81-322-3972-7_19

[8] F. Friedrich, J. Mendling, and F. Puhlmann, "Process model generation from natural language text," in *Advanced Information Systems Engineering. CAiSE 2011. Lecture Notes in Computer Science*, H. Mouratidis and C. Rolland, Eds., vol 6741. Springer, Berlin, Heidelberg, 2011, pp. 482–496. https://doi.org/10.1007/978-3-642-21640-4_36

[9] F. Lin, M. Yang, and Y. Pai, "A generic structure for business process modeling," *Bus. Process Manag. J.*, vol. 8, no. 1, pp. 19–41, Mar. 2002. https://doi.org/10.1108/14637150210418610
[10] F. Bargiela-Chiappini, C. Nickerson, and B. Planken, "What is business discourse?" in *Business Discourse*. *Research and Practice in Applied Linguistics*, F. Bargiela-Chiappini, C. Nickerson, and B. Planken, Eds. London: Palgrave Macmillan UK, 2013, pp. 3–44. https://doi.org/10.1057/9781137024930_1
[11] T. Jacobs and R. Tschötschel, "Topic models meet discourse analysis: a quantitative tool for a qualitative approach," *Int. J. Soc. Res. Methodol.*, vol. 22, no. 5, pp. 469–485, Sep. 2019. https://doi.org/10.1080/13645579.2019.1576317
[12] V. Dogra *et al.*, "A complete process of text classification system using state-of-the-art NLP models," *Comput. Intell. Neurosci.*, vol. 2022, Jun. 2022, Art. no. e1883698. https://doi.org/10.1155/2022/1883698
[13] A. Bhardwaj, Y. Narayan, and M. Dutta, "Sentiment analysis for Indian stock market prediction using Sensex and Nifty," *Procedia Comput. Sci.*, vol. 70, pp. 85–91, 2015. https://doi.org/10.1016/j.procs.2015.10.043
[14] A. Ittoo and A. van den Bosch, "Text analytics in industry: Challenges, desiderata and trends," *Comput. Ind.*, vol. 78, pp. 96–107, May 2016. https://doi.org/10.1016/j.compind.2015.12.001
[15] V. Prokhorov, M. T. Pilehvar, D. Kartsaklis, P. Lio, and N. Collier, "Unseen word representation by aligning heterogeneous lexical semantic spaces," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 6900–6907. https://doi.org/10.1609/aaai.v33i01.33016900
[16] E. Darics and J. Clifton, "Making applied linguistics applicable to business practice. Discourse analysis as a management tool," *Appl. Linguist.*, vol. 40, no. 6, pp. 917–936, Dec. 2019. https://doi.org/10.1093/applin/amy040
[17] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, Feb. 2018. https://doi.org/10.1177/0165551516677946
[18] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," *ArXiv Prepr. ArXiv13024964*, Feb. 2013. https://doi.org/10.48550/arXiv.1302.4964
[19] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alex. Eng. J.*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021. https://doi.org/10.1016/j.aej.2021.02.009
[20] N. Ur-Rahman and J. A. Harding, "Textual data mining for industrial knowledge management and text classification: A business oriented approach," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 4729–4739, Apr. 2012. https://doi.org/10.1016/j.eswa.2011.09.124
[21] P. Sunagar, A. Kanavalli, S. S. Nayak, S. R. Mahan, S. Prasad, and S. Prasad, "News topic classification using machine learning techniques," in *International Conference on Communication, Computing and Electronics Systems*, *Lecture Notes in Electrical Engineering*, V. Bindhu, J. M. R. S. Tavares, A.-A. A. Boulogeorgos, and C. Vuppalapati, Eds., vol. 733. Singapore: Springer Singapore, 2021, pp. 461–474. https://doi.org/10.1007/978-981-33-4909-4_35
[22] S. Godbole and S. Roy, "Text classification, business intelligence, and interactivity: automating C-Sat analysis for services industry," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Las Vegas Nevada USA, Aug. 2008, pp. 911–919. https://doi.org/10.1145/1401890.1401999
[23] V. A. Zeithaml, M. J. Bitner, and D. D. Gremler, "The gaps model of service quality," *Serv. Mark.*, pp. 37–49, 1996.
[24] S. Albitar, S. Fournier, and B. Espinasse, "An effective TF/IDF-based text-to-text semantic similarity measure for text classification," in *Web Information Systems Engineering – WISE 2014*, *Lecture Notes in Computer Science*, vol. 8786, B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali, and Y. Zhang, Eds., Cham: Springer International Publishing, 2014, pp. 105–114. https://doi.org/10.1007/978-3-319-11749-2_8
[25] H. Leopold, Ed., *Natural Language in Business Process Models: Theoretical Foundations, Techniques, and Applications*, LNBIP, vol. 168. Cham: Springer International Publishing, 2013. https://doi.org/10.1007/978-3-319-04175-9
[26] K. Honkisz, K. Kluza, and P. Wiśniewski, "A concept for generating business process models from natural language description," in *Knowledge Science, Engineering and Management: 11th International Conference, KSEM 2018,* Part I 11, Changchun, China, Aug. 2018, pp. 91–103. https://doi.org/10.1007/978-3-319-99365-2_8
[27] M. Camargo, M. Dumas, and O. González-Rojas, "Learning accurate LSTM models of business processes," in *Business Process Management*, *Lecture Notes in Computer Science,* T. Hildebrandt, B. F. Van Dongen, M. Röglinger, and J. Mendling, Eds., vol. 11675. Cham: Springer International Publishing, July 2019, pp. 286–302. https://doi.org/10.1007/978-3-030-26619-6_19
[28] M. Kocbek, G. Jost, M. Hericko, and G. Polancic, "Business process model and notation: The current state of affairs," *Comput. Sci. Inf. Syst.*, vol. 12, no. 2, pp. 509–539, 2015. https://doi.org/10.2298/CSIS140610006K
[29] M. El Kassis, F. Trousset, N. Daclin, and G. Zacharewicz, "Business process web-based platform for multi modeling and simulation," in *JFMS 2022-Les Journées Francophones de la Modélisation et de la Simulation: Ingénierie Dirigée par les Modèles pour la Théorie de la Modélisation et de la Simulation et les Systèmes Multi-Agents*, cepadues, 2022.
[30] J. Erasmus, I. Vanderfeesten, K. Traganos, and P. Grefen, "Using business process models for the specification of manufacturing operations," *Comput. Ind.*, vol. 123, Dec. 2020, Art. no. 103297. https://doi.org/10.1016/j.compind.2020.103297
[31] M. Wiemuth, D. Junger, M. A. Leitritz, J. Neumann, T. Neumuth, and O. Burgert, "Application fields for the new object management group (OMG) standards case management model and notation (CMMN) and decision management notation (DMN) in the perioperative field," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, pp. 1439–1449, 2017. https://doi.org/10.1007/s11548-017-1608-3
[32] R. Kumar, K. S. R. Murthy, J. Ramesh Babu, and A. Shaik, "Live text analyzer to detect unsolicited messages using count vectorizer," *J. Eng. Sci.*, vol. 14, no. 06, 2023. [Online]. Available: https://jespublication.com/uploads/2023-V14I60100.pdf. Accessed on: Jan. 02, 2024.
[33] R. Odegua, "An empirical study of ensemble techniques (bagging, boosting and stacking)," in *Proc. Conf.: Deep Learn. IndabaXAt*, 2019. [Online]. Available: https://www.researchgate.net/profile/Rising-Odegua/publication/338681864_An_Empirical_Study_of_Ensemble_Techniques_Bagging_Boosting_and_Stacking/links/5e2417c2458515ba2092f5cf/An-Empirical-Study-of-Ensemble-Techniques-Bagging-Boosting-and-Stacking.pdf. Accessed on: Nov. 12, 2023.
[34] G. Kunapuli, *Ensemble Methods for Machine Learning*. Simon and Schuster, 2023.
[35] T.-T. Wong and H.-C. Tsai, "Multinomial naïve Bayesian classifier with generalized Dirichlet priors for high-dimensional imbalanced data," *Knowledge-Based Systems*, vol. 228, Sep. 2021, Art. no. 107288. https://doi.org/10.1016/j.knosys.2021.107288
[36] L. Jiang, S. Wang, C. Li, and L. Zhang, "Structure extended multinomial naive Bayes," *Inf. Sci.*, vol. 329, pp. 346–356, Feb. 2016. https://doi.org/10.1016/j.ins.2015.09.037
[37] E. Mayoraz and E. Alpaydin, "Support vector machines for multi-class classification," in *Engineering Applications of Bio-Inspired Artificial Neural Networks*, *Lecture Notes in Computer Science*, J. Mira and J. V. Sánchez-Andrés, Eds. Springer Berlin Heidelberg, 1999, pp. 833–842. https://doi.org/10.1007/BFb0100551
[38] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition.," in *Esann*, 1999, pp. 219–224. [Online]. Available: https://www.esann.org/sites/default/files/proceedings/legacy/es1999-461.pdf. Accessed on: Nov. 12, 2023.
[39] W. S. Noble, "What is a support vector machine?" *Nat. Biotechnol.*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006. https://doi.org/10.1038/nbt1206-1565
[40] R. Flowers and C. Edeki, "Business process modeling notation," *Int. J. Comput. Sci. Mob. Comput.*, vol. 2, no. 3, pp. 35–40, Mar. 2013. https://ijcsmc.com/docs/papers/March2013/V2I3201305.pdf
[41] I. Mel'čuk and A. Polguere, "A formal lexicon in meaning-text theory (or how to do lexica with words)," *Comput. Linguist.*, vol. 13, pp. 261–275, 1987.
[42] J. Mendling, *Metrics for Process Models: Empirical Foundations of Verification, Error Prediction, and Guidelines for Correctness*, vol. 6. Springer Science & Business Media, 2008. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=LDUh05-Z-kkC&oi=fnd&pg=PA1&dq=Mendling,J.:MetricsforProcessModels:EmpiricalFoundationsofVerification,Error+Prediction,+and+Guidelines+for+Correctness.+LNBIP,+vol.+6.+Springer,+Heidelberg+(1974)&ots=7ZlqcB4No9&sig=SNjDw5vugvHaH1F22Lk3aRnGrJ8. Accessed: Dec. 19, 2023.

[43] S. Qaiser and R. Ali, "Text mining: use of TF-IDF to examine the relevance of words to documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, July 2018. https://doi.org/10.5120/ijca2018917395
[44] N. Japkowicz, "Why question machine learning evaluation methods," in *AAAI workshop on evaluation methods for machine learning*, 2006, pp. 6–11. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=46ecd8f4708764a9c5efb4916395f5d013b970cb. Accessed: Nov. 12, 2023.
[45] S. S. Biswas, "Role of Chat GPT in public health," *Ann. Biomed. Eng.*, vol. 51, no. 5, pp. 868–869, Mar. 2023. https://doi.org/10.1007/s10439-023-03172-7
[46] I. Adeshola and A. P. Adepoju, "The opportunities and challenges of ChatGPT in education," *Interact. Learn. Environ.*, vol. 32, no. 10, pp. 1–14, Sep. 2023. https://doi.org/10.1080/10494820.2023.2253858
[47] B. D. Lund and T. Wang, "Chatting about ChatGPT: how may AI and GPT impact academia and libraries?" *Libr. Hi Tech News*, vol. 40, no. 3, pp. 26–29, Feb. 2023. https://doi.org/10.1108/LHTN-01-2023-0009
[48] M. Dumas, M. La Rosa, J. Mendling, and H. A Reijers, *Fundamentals of Business Process Management*. Springer, Mar. 2013. https://doi.org/10.1007/978-3-642-33143-5

Shokoufeh Salem Minab received her Bachelor and Master degrees in Computer Engineering from the Islamic Azad University of Mashhad, Iran. She worked for five years as a Lecturer at the same university, teaching various computer science courses. In 2024, she completed her second Master degree in Computer Systems at Riga Technical University, Latvia.
Her research interests include natural language processing, business process modelling, text mining, and data analysis. After graduation, she joined the Big Data in Biomedical Research Group at the Technical University of Munich, where she worked on projects involving large-scale data analysis and biomedical applications.
E-mail: shokoufeh.salem-minab@edu.rtu.lv
ORCID iD: https://orcid.org/0009-0006-0148-3552

Erika Nazaruka received M. Sc. in Computer Systems in 2003 and PhD degree (Dr. sc. ing.) in Information Technology with specialisation in system analysis, modelling and design from Riga Technical University in 2006. She has been an Associate Professor at the Institute of Applied Computer Systems in Riga Technical University since 2013. She is the author of more than 70 conference papers, four book chapters and one book. Her research interests include topological functioning modelling, software quality assurance, model-driven and object-oriented software development, and natural language processing. She and her co-author, Janis Osis, were awarded by the Latvian Academy of Sciences for the book "Model-Driven Software Development: Architectures and Functions", which was recognized as one of the most significant theoretical achievements of Latvian Science in 2011.
E-mail: erika.nazaruka@rtu.lv
ORCID iD: https://orcid.org/0000-0002-1731-989X