Sampo Pietikäinen

DISCOVERING BUSINESS PROCESSES FROM UNSTRUCTURED TEXT

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY
2020

ABSTRACT

Pietikäinen, Sampo
Discovering Business Processes from Unstructured Text
Jyväskylä: University of Jyväskylä, 2020, 109 pp.
Information Systems Science, Master's Thesis
Supervisor: Pulkkinen, Mirja

Manual processing of the documents can be a time-taking task for a knowledge worker. This workload can be familiar to Business Process Management professionals who may have to go through multiple process descriptions in their work. This thesis attempts to find a way to mitigate the workload of the knowledge worker by proposing a natural language processing solution for discovering Business Processes from Business Process description documents. The research applied the design science research method and took several steps to produce the solution. The named entity recognition solution provided weak results, and instead of improving the solution, the research utilized genre analysis methods to seek an alternative approach. The classification of the headings of the document was deemed as a possibly viable solution. Four classification pipelines were built for classification of the headings and evaluated with cross-validation. The results of the first pipeline were somewhat promising; however, the cross-validation that was supposed to evaluate the ability to retrieve processes with previously unknown words had a poor performance. The following pipelines were created to improve from the baseline set up by the first pipeline. The second pipeline used part-of-speech tagging, the third used list of verbs relevant to business processes and the fourth pipeline used the context where process names appeared. These pipelines did not, however, make substantial improvements

Keywords: Business Process, Business Process Management, Natural Language Processing, Information Extraction, Information Retrieval, Design Science Research

TIIVISTELMÄ

Pietikäinen, Sampo
Liiketoimintaprosessien löytäminen strukturoimattomasta tekstistä
Jyväskylä: Jyväskylän yliopisto, 2020, 109 s.
Tietojärjestelmätiede, pro gradu -tutkielma
Ohjaaja: Pulkkinen, Mirja

Asiakirjojen käsittely manuaalisesti kuluttaa paljon tietotyöntekijän resursseja. Tämä koskee myös liiketoimintaprossien johtamisen asiantuntijoita, joiden työ voi vaatia useiden liiketoimintaprosessien kuvausten lukemista. Tämän tutkielman tavoitteena oli löytää ratkaisuja, jotka vähentävät tietotyöläisen asiakirjojen lukemiseen käyttämää aikaa soveltamalla luonnollisen kielen käsittelyn menetelmiä liiketoimintaprosessien etsimiseen asiakirjoista. Tutkimusmenetelmänä oli suunnittelutieteellinen tutkimus, joka sisälsi useita iteratiivisia vaiheita. Nimetyn kohteen tunnistamista käytettiin ensimmäisen ratkaisun suunnittelemiseen. Se ei kuitenkaan tuottanut toivottuja tuloksia, joten tutkimus siirtyi arvioimaan parempia mahdollisia ratkaisuja genre-teoriaa soveltavalla analyysillä. Tämän analyysin perusteella kehitettiin neljä asiakirjojen otsikkojen luokittelevaa ratkaisua tunnistamaan liiketoimintaprosesseja. Luokitteluratkaisut arvioitiin ristiinvalidoinnilla. Ensimmäinen luokitteluratkaisu suoriutui sattumanvaraisesti jaetusta ristiinvalidoinnista lupaavasti. Validoinnissa, jossa arvioitiin prosessien tunnistamista uusista asiakirjoista, ratkaisu ei kuitenkaan suoriutunut hyvin. Toinen luokitteluratkaisu sovelsi luokittelussa sanaluokkien tunnistamista. Kolmas luokitteluratkaisu hyödynsi listaa joka sisälsi liiketoimintaprosesseissa käytettäviä verbejä. Neljäs luokitteluratkaisu käytti syötteenä otsikon lisäksi kontekstia eli lauseita joissa otsikot esiintyivät asiakirjan tekstissä. Nämä luokitteluratkaisut eivät kuitenkaan tuottaneet merkittävästi ensimmäistä ratkaisua parempia tuloksia.

Asiasanat: Liiketoimintaprosessi, liiketoimintaprosessien johtaminen, luonnollisen kielen käsittely, tiedon eristäminen, tiedonhaku, suunnittelutieteellinen tutkimus

FIGURES

TABLES

TABLE OF CONTENTS

1 BACKGROUND

In Business Process Management, the creation of the process model requires comprehensive understanding from the process engineer, and usually, needed process information exists in the organization's documents. However, this information is not always easily retrievable since the information might be dispersed in unstructured text. This research will search for solutions to discover process-related information from unstructured text to mitigate this problem.

This chapter delves further into the concepts used in this research. The main concepts are Business Process Management, structurality of the text, and Natural Language Processing. Since the purpose of this research is to contribute to the automatic discovery of business process components, it is reasonable to explain what is meant by 'business process components' and 'business process elements,' that this thesis gratuitously uses interchangeably. Although structurality of the text is not a common academic concept, this section attempts to clarify the meaning of it since it does matter from the machine learning aspect, if the text is unstructured, semi-structured or structured. Finally, the chapter goes through the basic concepts of the natural language processing, which requires some clarification in the context of this research, since Natural Language Processing consists of all the human language use such as speaking and handwriting. The definition of Natural Language Processing is approached through text-related concepts like information extraction and text mining.

1.1 Business Process Management

Since the context of this thesis is heavily in the Business Process Management domain, it is necessary to illuminate the concept of Business Process Management. This section will look into the definition of Business Process Management and present the viewpoints of different authors. For example, Strnadl defines a process as "a structured and measured sequence of activities designed to produce a specific output based on defined input” and a Business Process as ”a complete, dynamically coordinated set of collaboration and transactional activities” (2006).

Hammer, on the other hand, describes the Business Process Management as the union of the Business Process Re-engineering and the quality movement. According to Hammer, BPM is an "integrated system for managing business performance by managing end-to-end business processes." (2015) Hammer's definition highlights the value creation ("business performance") and means for this value creation (end-to-end business processes). We can probably safely assume that this integrated system does not equal to an information system like a Business Process Management Suite but the way to work that has been integrated into an organization. Hammer also lists principles of Process Management:

- All work is process work
- One process is better than many
- Even good process must be performed effectively
- Even good process can be made better
- Every good process eventually becomes a bad process (2015).

The last one of these principles raises a question in the context of this thesis. When does the process documentation become stale? Do we need to find the documentation of stale processes? There could be an argument that most of the process documentation could be too old to be useful, and there are so few relevant documented processes that they are easy to retrieve manually. This might be true in some cases; however, “any process is better than no process,” and if the stale documentation is all that there is, it still might provide a starting point for process re-design.

Swenson & von Rosing and Weske have more precise definitions than Hammer. According to Weske: "Business process management includes concepts, methods, Business Process Management, and techniques to support the design, administration, configuration, enactment, and analysis of business processes." (p5., 2012).

Swenson and von Rosing have a somewhat similar definition:" Business process management (BPM) is a discipline involving any combination of modelling, automation, execution, control, measurement, and optimization of business activity flows in applicable combination to support enterprise goals, spanning organizational and system boundaries and involving employees, customers, and partners within and beyond the enterprise boundaries" (Swenson & von Rosing, 2015).

Both Swenson & von Rosing and Weske thus imply that in Business Process Management, the management part is a combination of specific actions. Swenson & von Rosing also define the actors, context, and the reason why BPM is done (to support enterprise goals) (2015).

Swenson & von Rosing attempt to clarify common misconceptions of Business Process Management. They distance BPM from Information Technology; basically, they are implying that just having and using Business Process

Management System is not Business Process Management. Swenson and von Rosing additionally try to steer the Business Process Management definition away from commercialization and marketing hype, i.e., BPM is not a product or a market segment. (2015.)

Dumas, La Rosa, Mendling & Reijers write that "Business Process Management (BPM) is the art and science of overseeing how work is performed in an organization to ensure consistent outcomes and to take advantage of improvement opportunities." (p. 1, 2018) This definition is criticized by Swenson & von Rosing since they interpret it meaning as merely an improvement of a single process step would be accounted as Business Process Management (2015). Dumas et al., however, later explicitly state that "BPM is not about improving the way individual activities are performed" (2018).

The most appealing of these definitions is Hammer's definition. It tells the purpose and methods briefly. Although Swenson & von Rosing's list gives us some indication of what are the actions when managing business processes, it raises a question: are these actions exhaustive? Are these definitions susceptible to change if innovation in the Business Process Management context appears?

Rather than finding the most accurate definition for the business process management, the more interesting question is where does the possible contribution of this research stand in Business Process Management. The following sections will look into the BPM lifecycle and principles of Business Process Management from the view of this research.

1.1.1 Business Process Management Lifecycle

The predecessor of the business process management was the Business process re-engineering. It did have a similar goal to change the organizational work into more productive and competitive. However, its greatest weakness, according to Hammer, was that redesigns were meant to be substantial one time efforts (2015). In hindsight, it certainly seems evident that no process would remain competitive and effective indefinitely after the redesign. If the redesign happens after competitors have developed far better processes, the process redesign would be just a reactive effort. Analogues can be found from the Information Technology where agile methods and continuous improvement are confronting the massive one-time development investments.

In Business Process Management processes should be improved continuously. For this continuous effort, the process management should follow an iterative lifecycle. There are multiple BPM and process lifecycles (see, for example, Reichert, Hallerbach, & Bauer, 2015 and von Rosing, Foldager, Hove, von Scheel, & Bøgebjerg, 2015). The Natural Language Processing could make some of the steps of these lifecycles less time-consuming. This thesis will briefly inspect the lifecycle presented by Dumas et al. (2018) in the next table (TABLE 1) and align it with this research.

TABLE 1 Business Process Management Lifecycle steps and activities (Dumas et al., 2018)

Lifecycle step	Activities (Dumas et al., 2018)
Process identification	”The outcome of process identification is a new or updated process architecture, which provides an overall picture of the processes in an organization and their relationships.”
Process discovery	Documenting the current state of the process (As-is modeling).
Process analysis	Purpose of the process analysis is to measure and analyse the as-is models from process discovery to create insight of the processes.
Process redesign	Produces the to-be process models. This step may produce multiple options that are analysed. Thus process redesign may be executed in parallel with the process analysis.
Process implementation.	Includes organizational change management and automation. In the organizational management change the process participants change the way of working and in the process automation IT systems are used to automate the processes. This produces the executable new process.
Process monitoring	Includes measuring the running processes and discovering possible bottlenecks and errors.

The outcome of the design that this research proposes could be useful in all the life cycle steps after process documentation, i.e., there should be some documentation where to retrieve processes. The proposed design could save time if there is breakage or a longer pause in the BPM lifecycle.

1.1.2 Maturity Models

At which point would an organization be in need of the process discovery solution that this research is examining? The presumption would be that the organization should have at least some interest in Business Process Management if the organization wants to know what processes are in its documents. Can an organization be a Business Process Management powerhouse that knows which processes it is dealing with? Maturity models can be used to evaluate Business Process Management in the organization. There are also several maturity models in the BPM context (see, for example, Rosemann & Vom Brocke, 2015). The purpose of the maturity model is to assess the maturity of processes and Business Process Management in an organization. The next table (TABLE 2) presents the BPM maturity model that applies to the levels of the Capability Maturity Model Integration (CMMI) framework.

TABLE 2 Capability Maturity Model Integration framework (Dumas et al., 2018)

Levels	Description (Dumas et al., 2018)
Level 1 (Initial)	BPM is non-existent (There is no process documentation)
Level 2 (Managed)	The first processes are documented but the methods and tools have not been internalized.
Level 3 (Defined)	Organization enables the collaboration process development. Inhouse knowledge increases.
Level 4 (Quantitatively managed)	BPM Center of Excellence is established. BPM is integrated in every project.
Level 5 (Optimizing)	BPM has organization wide acceptance in strategic and operational level BPM Center of Excellence diminishes.

Dumas lists six success factors for the BPM maturity: Strategic Alignment, Governance, Methods, IT, People, and Culture. Each of these factors includes five capabilities that can be used as individual measures of the BPM maturity of the organization (2018, p. 478).

Based on the maturity levels, the natural language processing of Business Process documents would be most useful in level 3, but it would also be relevant in levels 2-4. The first level might not have any documentation to process. On the fifth level, the knowledge of the processes is organized so well that the knowledge worker does not have to rely on unstructured text documentation (Although this might be an overly optimistic view of the organization on the highest maturation level).

Rummler and Ramias present more organizational categorization for the Business Process Management context. Their Value Creation Hierarchy (VCH) has five levels. The topmost level is the Enterprise level, which presents the whole organization, which Rummler & Ramias see as a large processing system. This level encompasses every single one of the organization's processes. Value Creation Level is the second level, which depicts how the organization creates, sells, and delivers products and services in a Value Creation System. Large organizations may have multiple Value Creation Systems. In the third level, this Value Creation system is divided into three sub-processing systems: Launched, Sold, and Delivered. Rummler & Ramias refer to these sub-processing systems also as processes, although they clarify that these are still just ”bundle of processes.”

The fourth level is the process level. The processes here are end-to-end processes, i.e., they start from market input or another value chain, and their output is a service or a product to market or input to another value chain. This level also includes the management processes and the support processes along with the Launched, Sold and Delivered. Although Rummler & Ramias do have a definition and place in the hierarchy for the process, they do not strictly reserve the use of the word ”process” to this level.

The last level is the subprocess/task/subtask level, which contains, as the name implies, all the more detailed work processes like tasks, subprocesses, and procedures. Rummler & Ramias bundle these lower levels to one level and are not setting strict categories or definitions to sub-processes or tasks. Rummler and Ramias are also not strictly defining the process nor its components but providing an Enterprise Architecture framework for Business Process Management. They do, however, present the diagram and maps that are used to document each level. This framework may, however, provide a category for business processes in the information extraction on machine learning classification context. For example, if a knowledge worker would like to know if the process is in the support, management, launched, sold, or delivered category. (2010.)

In this research, it is presumed that the design will contribute to the following areas in the Business Process Management domains: Lower levels in the value hierarchy, middle levels in the maturity model, and all the other levels except process identification in BPM lifecycle (TABLE 3).

TABLE 3 BPM Maturity Model levels, Value Hierarchy levels, and Lifecycle steps. The relevant levels and steps for this research's solution are bolded.

Maturity Model	Value Hierarchy	Lifecycle step
Level 1 (Initial) BPM is non-existent (There is no process documentation)	Enterprise Level	Process identification
Level 2 (Managed)	Value Creation Level	**Process discovery**
Level 3 (Defined)	Processing Sub-Systems Level	**Process analysis.**
Level 4 (Quantitatively managed)	**Process Level**	**Process redesign**
Level 5 (Optimizing)	**Subprocess/Task/Subtask Level**	**Process implementation.**
		Process monitoring

1.1.3 Definition of a Process

The word "process" has its origins in the Latin word "procede," go forward. It has gained multiple meanings throughout its existence in biology, law, computing, and chemistry. When looking at the definitions of "process" from dictionaries, most relevant definitions that fit in the Business Process Management context would be "a series of actions or operations conducing to an end. Especially: a continuous operation or treatment especially in manufacture" from Merriam and Webster's online dictionary (Process, 2019b) and "A series of actions or steps taken in order to achieve a particular end." from Lexico's online dictionary (Process, 2019a). So, according to dictionaries, a process contains actions and has an end.

In the business process management literature, some definitions of" process" and" business process" are merely paraphrasing the dictionary definitions in a manner more suitable for business discipline. For example, Weske writes that "a business process consists of a set of activities that are performed in coordination in an organizational and technical environment. These activities jointly realize a business goal."(2012) Weske bases this definition on Davenport's view on "process." This view includes the description: "a process is simply a structured, measured set of activities designed to produce a specified output for a particular customer or market" (Weske, p5.2012).

Strnadl conduces from previous literature that: "A business process is a complete, dynamically coordinated set of collaborational and transactional activities that (1) delivers value to customers or (2) fulfils other strategic goals of the enterprise" (2006). Dumas et al. restrain to delve into the deep end of the process definition. In their introduction, a process is a chain of events, activities, and decisions that "ultimately add value to the organizations and its customers" (2018, p. 1). These definitions add the organizational context and business value of the outcome to the meaning of "process".

Rummler, Ramias & Rummler, however, have a bolder definition: " process is a construct for organizing the value-adding work to achieve a business valued milestone". They further lay down three criteria for a process. According to them the process:

> " Can be performed effectively and efficiently. Can be managed effectively. Offers the potential for a competitive advantage."

They argue that treating just any work in a sequence as the process does not contribute to enhanced organizational performance. (2009.)

Rummler & Ramias specify process scope in the Value Creation Hierarchy as an end-to-end process. In this scope, the process begins from a market or another value chain and produces an output to market or another value chain (2010). Hammer offers us more criticism for the ambiguity of process definition. Hammer criticizes the description of the process by the quality approach (e.g., Six Sigma) as "any sequence of work activities". Hammer argues that this would bring almost any work, also the work that is not strategically relevant included in the business process management. Management of such a large amount of small scale processes would become difficult. (2015.)

Hammer defends the definition of the process in Business Process Reengineering, of which proponent Hammer was. The description of a process as: "end-to-end work across an enterprise that creates customer value", according, to Hammer gives the organization focus on the meaningful aspects of the operations. Hammer names, especially fragmentation as one of the "evils" that implementing this definition could alleviate. (2015.)

Von Rosing et al. present a class model to describe levels of work. This class model includes from highest to lowest level, process area, process group, process, process step, activity and procedure. With this model, von Rosing et al. challenge the definitions of the compositions and decompositions of a process as merely a set of activities or steps. Since process as a bundle of steps only tells that there is a relationship between process and step or activity. Von Rosing et al. argues that these different levels are inherently different concepts. They further criticise the definition of a process as a set of smaller units of work will lead processes being analogous to Matryoshka dolls. Von Rosing et al. implicitly shun the use of the word process outside its categorization context. According to von Rosing et al., ambiguity in the hierarchy makes recognition of the level where the process belongs impossible without additional context. This lack of information will make the process documentation difficult. Von Rosing et al. does not, however, give concrete examples or cases where not using the levels of work has led to problems. Nor do they tell where does this need for level categorization emerges. (2015.)

Thus von Rosing et al. refrain from defining a process as a set of activities or steps at all. They describe a process as a "member, along with other processes, of the chain of dependent work within the complete Process Life Cycle". The output of a process is a "single, complete, and meaningful result that contributes to the completion of the valued output necessary for the conclusion of the work of a Process Group." A process group is a set of processes that produces an output that has value to specific stakeholders. It has only one process that produces the final output and the other processes ensure that the main process can produce the expected value. Process group contains all the processes to plan, prepare and deliver its output. (2015.)

Although von Rosing et al.'s criticism is certainly valid. Their definition is not much better since it requires to explain the Process Life Cycle and the process group. Unlike the other authors, von Rosing et al. Also give instructions on process naming. The process name should be derived from the process goal. A process goal is "to change one or more inputs into a specific output each time it is executed". Also, von Rosing et al. states that the verb, combined with the process goal in the name, should express the sense of completion. (2015.)

1.1.4 The Conclusion to Process Definition

Although Hammer, Rummler & Ramias and von Rosing et al. do not claim that process does not contain any smaller units of work they do attempt to steer the definition of the process into context of management. As if the work cannot be managed, it is not a process. This is a sound argument, especially when talking about the "process" in Business Process Management.

Von Rosing et al. do give instructions on how to identify and name a process. Also, von Rosing et al. provide a list of verbs to be used in process documentation. (2015.) This verb list could have some use in the Natural Language Processing context. However, if the document explicitly claims that an entity is a process, it might be out of this research' reach to evaluate through Natural Language Processing if the entity is indeed a process that complies with Weske's, Hammer's or Rummler et al. 's definition. Even though the natural language processing methods would not be able to recognize if a process has a competitive edge or if it is easily managed, the outcome of this research might be able to make the documented work more manageable. Also, it could provide a more transparent glance through organizational knowledge.

1.1.5 The Documentation and Presentation of the Process

Business processes can be documented with natural languages, but there are specific notations and diagrams to make the communication of the process easier. IDEF0 in the business process context describes the process as a function or activity that has input, controls, mechanism and output. IDEF0 modelling presents activity as a box and input, output, mechanism and controls as one or more arrows pointing at the box (FIGURE 1).

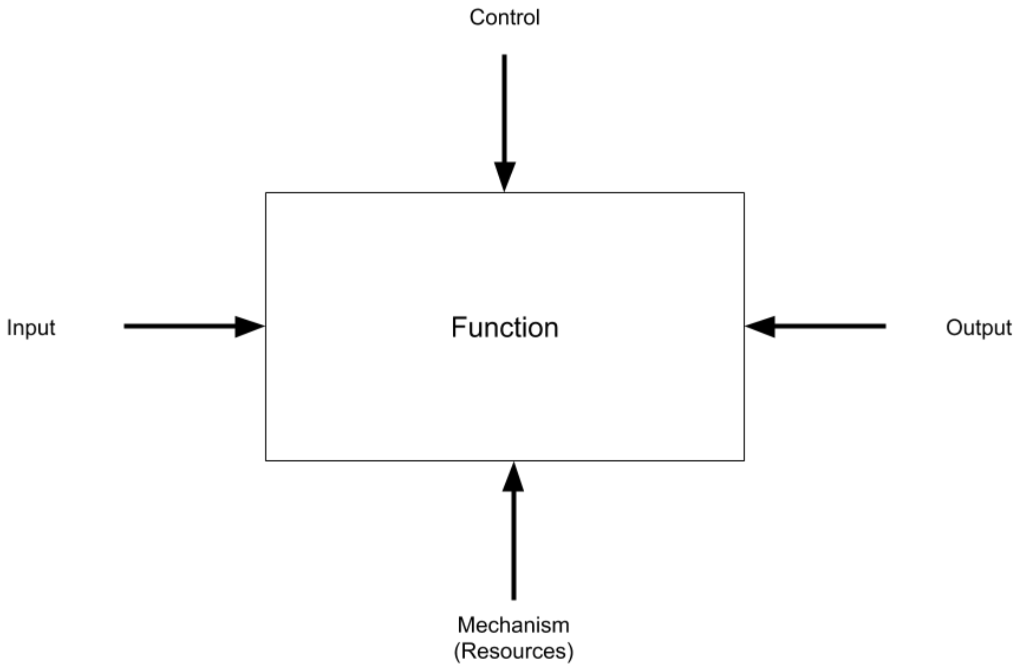


FIGURE 1 IDEF0 presentation of a function

IDEF0 allows to present activities hierarchically so that one activity box may include multiple activities and multiple activities can be abstracted to fewer activities. This abstraction will, however, lead to the situation of multiple "matryoshka dolls" as von Rosing et al. described (2015). It does give more information about the process context and the means than, for example, Business Process Modelling Notation.

As an example implementation of IDEF0 is an early warning system development process, by Fortier & Dokas. They mentioned also looking into methodologies like UML and BPMN but chose the IDEF because of the combination of the definition of the engineering process and constraints, it provided (2008). IDEF0 has been a basis for more extensive Process Scope Diagrams, also known as IGOEs (Inputs, Guides, Outputs and Enablers). According to Harmon, IGOE is useful in problem analysis and provides a better emphasis on policies, rules and management than other workflow diagrams (2010).

The purpose of the Business Process Modelling and Notation(BPMN) is to provide a standardized common language for those who work with the business processes. It is also closely connected to the Business Process Execution Language (BPEL), which enables the execution and web server choreography of the business processes. BPMN can be used just to document the business process, or it can be used to create running instances in the process in several Business Process Management systems. The Business Process Modelling and Notation of a process includes activities and other elements relevant elements as well as IDEF. The next section explains more about these activities and elements.

1.1.6 Components of Business Process

As this thesis often refers to business process components and business process elements, it is essential to clarify what are these elements and components. Dumas et al. condense these into business process ingredients (2018) which are presented in the following figure (FIGURE 2) and explained in the table after the figure (TABLE 4).

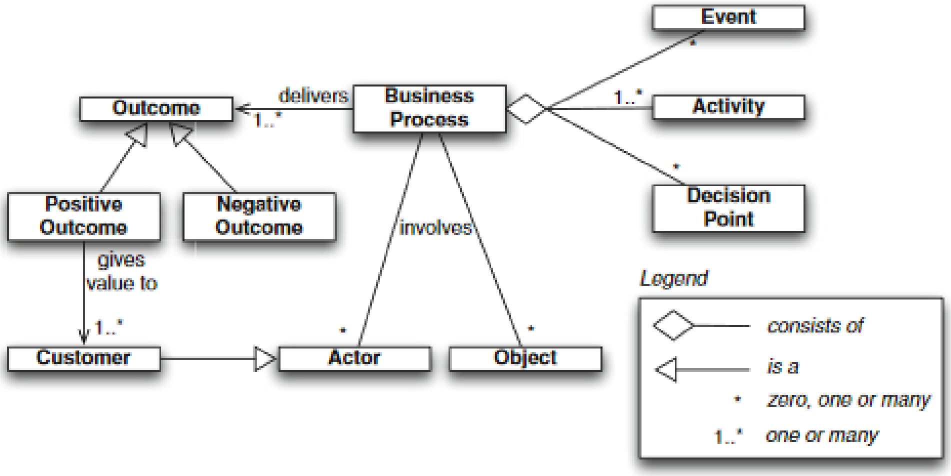


FIGURE 2 Business Process Ingredients (Dumas et al., 2018)

TABLE 4 Business Process ingredients and their definitions.

Component	Definitions by Dumas et al.
Event	Thing that happen "atomically, which means that they have no duration"
Activity	"fine-grained or coarse grained units of work"
Decision Point	"points in time when a decision is made that affects the way the process is executed"
Actors	"human actors, organizations, or software systems acting on behalf of human actors or organizations"
Objects	"Physical objects, such as equipment, materials, products, paper documents" and "Informational objects, such as electronic documents and electronic records"
Outcome	Negative or Positive

Dumas et al. Use these elements to define a business process as "a collection of interrelated events, activities, and decision points that involve a number of actors and objects, which collectively lead to an outcome that is of value to at least one customer" (2018).

Another categorization of business process components is by von Rosing, Laurier & Polovina (2015). They present the decomposed Process Meta-Objects. These meta-objects consist of:

- Process area (categorization)
- Process group (categorization)
- Business process
- Process step
- Process activity
- Event
- Gateway
- Process rule
- Process measurement (process performance indicator)
- Process owner
- Process flow (including input/output)
- Process role (von Rosing et al., 2015).

Components in the Business Processes are not drastically different in the literature as the BPMN standardizes the documentation. The table in the appendix (Appendix 1) shows the comparison between Dumas et al.'s and von Rosings et al.'s definitions of process components.

Von Rosen et al. does not have an equivalent to an actor. Closest meta-object would be a Process Role or Process owner. From this comparison, it can be agreed that a process can be divided into smaller steps, i.e. Activities. It also includes events that change the state of the process and an outcome that should, according to the process definitions, produce business value to the organization. The process may also include gateways that affect paths of the sequence flow objects that are relevant to the process.

1.1.7 Process Discovery

The goal of this research is to find a solution to discover business processes, which is an actual step in the Business Process lifecycle. According to Dumas et al. "Process discovery is defined as the act of gathering information about an existing process and organizing it in terms of an as-is process model. This definition emphasizes gathering and organizing information" (2018).

The purpose of the solution suggested in this thesis could support this definition of Process discovery, and would provide useful tools for it. In the preliminary stage, the discovery of the process elements from natural language text would be agnostic on the "as-is" or "to-be" aspect of the process, and the organization of the extracted information is not in this research scope.

Dumas et al. name three Process Discovery Methods: evidence-based discovery, interview-based discovery, and workshop-based discovery. The evidence-based discovery is divided into three sub-methods: Document Analysis, Observation and Automated process discovery. Document analysis means using existing documentation related to business processes as a source to discover business processes and Automated process analysis refers to mining event logs to discover processes. (2018.) The first steps this research is taking towards the automated process discovery is automatization of the document analysis from the ideal case where the process description exists and is explicit.

1.2 Unstructured, Semi-structured and Structured Text

'Unstructured' and 'structured' may not be the most definitive modifiers for the 'text'. It is not that clear when the unstructured becomes structure or vice-versa. Unstructured data refers to data of which content is not structured for computer consumption (Gandomi & Haider, 2015). This can be, for example, text or images. In other words, the consumer of the text is human rather than a computer. Unstructured text can convey very explicit knowledge for the reader and can be the best way to inform humans. Thus it should not be mixed with ambiguous text, although ambiguity can be one of the traits of the natural language. The structured text would be information, such as a database table that is readily usable for automatic processing purposes. Between these extremes lies semi-structured text. Abiteboul lists common aspects that make the text or data semi-structured:

- The structure is irregular
- The structure is implicit

- The structure is partial
- Indicative structure vs constraining structure
- A-priori schema vs a-posteriori data guide
- The schema is very large
- The schema is ignored
- The schema is rapidly evolving
- The type of data elements is eclectic
- The distinction between schema and data is blurred (Abiteboul, 1997).

Abiteboul mentions HTML and SGML (a precursor of HTML and XML) multiple times in the Semi-Structured context. HTML's main function is to describe how the text content is viewed to the consumer of the content, although it does provide opportunities for some machine readability. (1997.)

The Portable Document Format (PDF) veers more towards human readability than HTML at the cost of machine readability. The evident problem is that PDF's layout may seem visually similar to XML and HTML formats. Yet it mainly contains metadata of how to visually layout the elements in the document and does not contain rich semantics as the markup languages. So even if, for example, a table in PDF and HTML may look similar to a human, to a machine HTML is more semi-structured and a PDF is more unstructured than HTML. For example, extracting text from tables in PDF may result in text that is not machine-readable (Van Auken, Jaffery, Chan, Müller, & Sternberg, 2009). Spacy, a natural language processing library that is used in chapter 3 of this research, takes only unstructured text as an input. The process descriptions that this research uses are in PDF format. Most basic PDF parsers do lose the relevant information that the PDF provides. There are layout-aware text extraction tools such as LA-PDF Text (Salloum, Al-Emran, Monem, & Shaalan, 2018; Ramakrishnan, Patnia, Hovy, & Burns, 2012) and proprietary tools like iText (Kreuzthaler, Schulz, & Berghold, 2015)

1.3 Natural Language Processing

"Natural language processing employs computational techniques for the purpose of learning, understanding, and producing human language content." Natural Language Processing may include machine translation, speech recognition, and producing artificial speech and also more advanced applications such as sentiment analysis and conversational agents. (Hirschberg & Manning, 2015.) Natural Language Processing can be seen as an application of information technology, statistical and linguistic theories.

Other related terms to Natural Language Processing that concentrate more on the automatically consuming unstructured text would be Text Mining, Information Extraction and Information Retrieval.

Text Mining refers to discovering knowledge from textual data through Natural Language Processing methods (Hotho, Nürnberger & Paaß, 2005) and could be considered as a hyponym of Data Mining. Information Extraction refers to methods used to get the embedded information in unstructured text into a structured format. (Piskorski & Yangarber, 2013).

Information Extraction is close to this research since the research problem is how to find the process component in the unstructured text. At the lowest level information extraction can be Part-of Speech assignment and named entity recognition, and at the highest level semantically parsing the sentences so that for example, a location of the organization can be derived.

1.4 Motivation

Business Process Modeling creates explicit process knowledge from tacit process knowledge. Usually, a domain expert creates the informal description of the process which is then consumed into a formal process model by a process modeller. Both the domain expert and the process modeller should work closely together and have a high level of expertise. Neither of these is always reality. (Pinggera et al., 2010) Also, Dumas et al. point out fragmented process knowledge as one of the challenges in Process Discovery (p. 162, 2018). Information on the processes may be distributed to different domain experts. One of the strengths of the process documentation, when discovering processes, is, however, that it does not require reaching the experts involved in the process. Although arguably, the participation of these experts is very useful if possible.

In the Business Process Management lifecycle, 60% of the time is spent on the creation of process models. This time consumption is a paradox because of the availability of textual process descriptions in organizations (Friedrich, Mendling, & Puhlmann, 2011). Even though process descriptions might be available, they might not be easily retrievable. The volume of unstructured data, including unstructured text, is continuously growing (Dhar, 2013). Also, the organizational information has been distributed to multiple formats documents, wikis, spreadsheets, emails, instant messages and memorandums. To find relevant information and extracting knowledge from this is a challenge for a knowledge worker. Manually searching correct information consumes resources and the information should be at hand at the correct time.

Manual information retrieval with inadequate tools may lead to Information overload. Bawden & Robinson assert that: ”information overload occurs when information received becomes a hindrance rather than a help, even though the information is potentially useful” (2009). Even when the psychological effects on knowledge worker are ignored, the time taken by the information retrieval is a definite hindrance. Text analytics and natural language processing have been suggested as a tool to harness the value of the unstructured text (Müller, Junglas, Debortoli & vom Brocke, 2016).

The previous research on Natural Language Processing and Business Process Modeling and Management has mainly concentrated on the generation of process models from a preselected text that is known to contain process descriptions. For example, Ferreira, Thom & Fantinato developed a system that generated process models from short sentences that included only text that was relevant to the process model (2017). These studies have mainly handled the unstructured process descriptions as a pseudo-code.

Process description also differs from existing Natural Language Processing implementations in that Business Process descriptions are more ambiguous. Business processes steps do have recommended predefined verbs. However, they are also common in the English language, unlike in medical and biological, where information extraction can be enhanced simply by comparing the frequency of the words in the English language.

The problem is how to find process components from the unstructured text where they may or may not exist. This research attempts to use existing Natural Language Processing solutions on extracting Business Processes from unstructured texts but also broaden the unstructured text where to extract process elements. Fast retrieval from documents would further enhance the existing solution for process extraction.

As it was stated in the section on Business Process Management (section 1.1), the document analysis as a process discovery method is an analysis of process-related documents and the automated process discovery is the process mining from the event logs (Dumas et al., 2018). The existence of large amounts of process documentation and automated process mining provides motivational questions for this research: "Why not automate the document analysis". It could be of course, even more, efficient when combined with process mining, although that is not in this research' scope.

1.5 Research Method

The main research method of the thesis will be design science research method. To implement the design science research method, the research will follow Design Science Research Methodology Process (Peffers, Tuunanen, Rothenberger & Chatterjee, 2007). Design Science Research Methodology in this research is presented in the next figure (FIGURE 3).

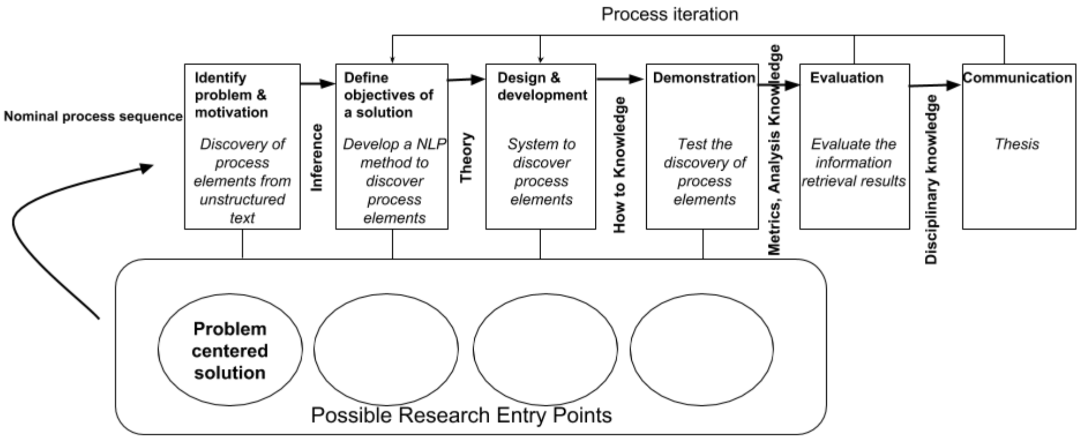

Process iteration
Nominal process sequence
Identify problem & motivation
Discovery of process elements from unstructured text
Inference
Define objectives of a solution
Develop a NLP method to discover process elements
Theory
Design & development
System to discover process elements
How to Knowledge
Demonstration
Test the discovery of process elements
Metrics, Analysis Knowledge
Evaluation
Evaluate the information retrieval results
Disciplinary knowledge
Communication
Thesis
Problem centered solution
Possible Research Entry Points

FIGURE 3 Design Science Research Methodology Process (Adapted from Peffers, Tuunanen, Rothen-berger & Chatterjee, 2007).

Definition of the objectives of a solution step will include evaluation of the literature research results for the possible tools and algorithms to be used in the design and development step. Also, the evaluation will use the analysis from the evaluation results seen as successful in the previous literature. Since the Design Science Research Method is an iterative process, it can take steps back to re-adjust the research towards the solution. This also applies to this research.

After literature research, a design using the named entity recognition was developed, but it performed poorly in the evaluation. After this, the research iterated back to defining the objectives for a solution that utilized genre theory methods. The observations of this analysis were used as a basis for the design that recognizes process names from document headings. The results were promising, and the last design step added more context for the classification to produce the final solution in the scope of this research. The following figure presents the iterative process of this Design Science Research (FIGURE 4).

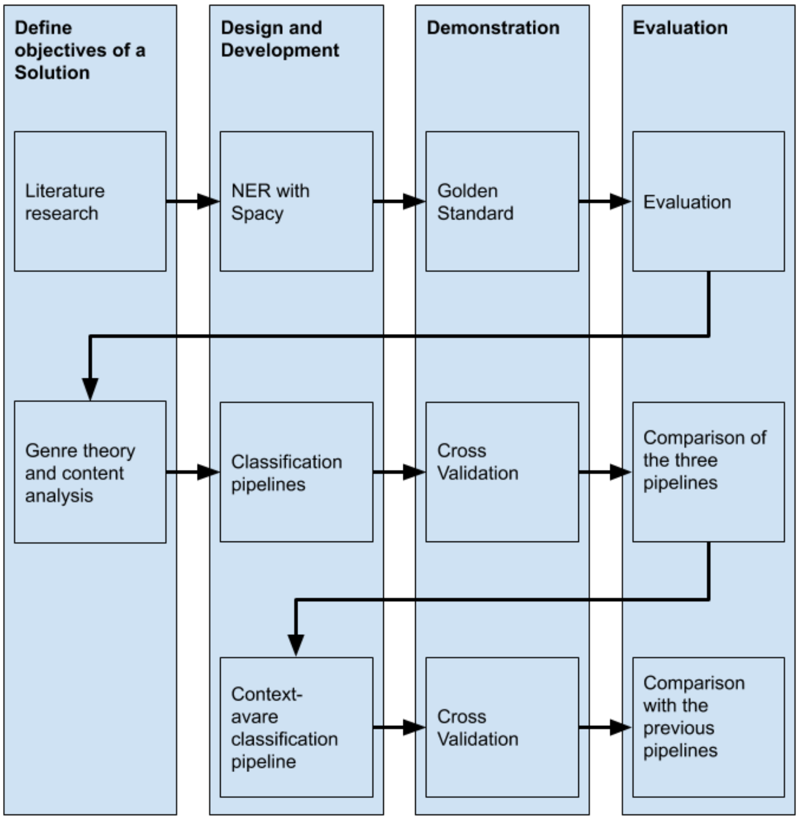

Define objectives of a Solution
Design and Development
Demonstration
Evaluation
Literature research
NER with Spacy
Golden Standard
Evaluation
Genre theory and content analysis
Classification pipelines
Cross Validation
Comparison of the three pipelines
Context-avare classification pipeline
Cross Validation
Comparison with the previous pipelines

FIGURE 4 The iterative process of this Design Science Research

The research will also use the design science research method's guidelines (Hevner, March, Park & Ram, 2004). The way Hevner et al.'s guidelines align with this research is presented in the following table (TABLE 5).

TABLE 5 Design Science guidelines by Hevner et al. (2004)

Guideline	Guideline description (Hevner et al. 2004)	How guidelines are implemented in this research.
Guideline 1	"Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation."	The result of the research is an instantiation that demonstrates ability to recognize business process elements.
Guideline 2	"The objective of design-science research is to develop technology-based solutions to important and relevant business problems."	The design provides a solution for Information Extraction in the Business Process Management domain.

(to be continued)

TABLE 5 (to be continued)

Guideline 3	"The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods."	The solution is evaluated with dynamical analysis methods by using Information retrieval metrics: Precision, Recall, F-score, ROC and AUC.
Guideline 4	"Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies."	The contribution will be the methods that can be used to recognize business process elements from unstructured text.
Guideline 5	"Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact."	After the development step, the methods used will be compared and evaluated.
Guideline 6	"The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment."	Due to the maturity of the NLP research, the possible methods will be researched in the literature review and then selected through evaluation for implementation. Also, the research will include a content analysis of the Business Process documents.
Guideline 7	"Communication of Research Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences."	Thesis

2 LITERATURE REVIEW

This literature review looks into the state of the art methods, tools and measures that could be applicable for the solution of the research. The purpose of this literature review is to support the definition of the solution and the design and development step. The observations from the measures used in the reviewed papers' results will support the evaluation of the design.

2.1 Literature Review Scope

The most relevant part in mining business process parts from unstructured text is recognition of sufficient recall and precision so the information can be further presented to the user. Business process management provides a holistic view of processes and management that has been divided into Strategic Alignment, Governance, Methods, Information Systems, People, and Culture (Rosemann & vom Brocke, 2015). These core elements have many different layers of documentation, and as whole produce, much interesting, unstructured text but the purpose of this literature review is to concentrate on the process part in the Business Process Management and more precisely on the explicit process definition in unstructured text.

This research does aim to contribute to extracting relevant process components such as activities, roles and actor, but the scope of the prototype is to concentrate on the process names. Process names are viable starting point since the processes are more likely explicitly named in the process descriptions than, for example, roles or activities. The assumption here is that if there are no process names in the document, there is probably no other Business Process component.

Although process mining might support recognition of the process components, it will be ignored in the literature review. Also, the generation of Business Process Model Notation is not going to be the end result of this research. Other processes, workflow models and procedures will also be counted out from the final artefact, but this will be briefly examined in the literature research if any possible applicable solution or methods are found.

2.2 Research Methods

The literature review will be conducted as a descriptive review as defined by Paré, Trudel, Jaana & Kitsiou (2015) and the literature research will follow the process of Information Systems' literature research (TABLE 6) divided into eight steps as defined by (Okoli & Schabram, 2010). The purpose was to find the Natural Language Processing methods, tools and measures that would be relevant for discovering process elements from documents containing unstructured process descriptions.

TABLE 6 Steps of the literature research of the thesis

Literature Review guideline steps (Okoli & Schabram, 2010)	Implementing the steps in the Literature Review in the Thesis
1. Purpose of the literature review	Purpose of the literature research is to discover and evaluate Natural Language Process methods and their applicability in the Business Process Management domain
2. Protocol and training	
3. Searching for the literature	The search of the literature will use Google Scholar, ACM, IEEE and EbscoHOST to find the articles from journals and conference papers.
4. Practical screening	In Process Discovery the titles that refer for example to the process of the Natural Language Processing itself rather than discovering processes with these methods, the paper will be excluded.
5. Quality appraisal	The Descriptive Review does not require a quality appraisal (Paré et al., 2015)
6. Data extraction	Tools (frameworks, systems and applications) and methods (statistical models, algorithms), and measures will be extracted from the studies.
7. Synthesis of studies	The data of these studies will be qualitatively analysed and presented as a table. Content analysis/ frequency analysis.
8. Writing the review	The results and the review process will be reported in literature research and in the thesis.

The literature review did not have the protocol and training step since only one person took part in the research. The sources for the research papers were Google Scholar[1], ACM[2], IEEE[3] and EbscoHost Business Source Elite[4]. These search platforms were selected since they all provided access to collections of the journal and conference articles related to information systems, business and information technology topics.

The search was done in the aforementioned databases with combinations of phrases related to business process, natural language processing and phrases that were specific to this research problem (TABLE 7). Process Description was selected as a search phrase to cover possible existing papers on other than business processes. “Process” itself was deemed as too vague as a single search phrase.

[1] https://scholar.google.com

[2] https://dl.acm.org

[3] https://ieeexplore.ieee.org

[4] https://www.ebsco.com/products/research-databases/business-source-elite

These search phrases were in quotes, to ensure that only the whole phrase was accepted, and combined with "AND" operator.

TABLE 7 Grouping of the Search Phrases

Natural Language Processing	**Business Process**	**Problem-Specific Phrases**
Information Extraction	Business Process Management	Unstructured text
Natural Language Processing	Business Process	Process Description
Information Extraction		
Text Mining		

The search result counts and the different combinations are in the appendix (Appendix 1). From these results, the papers that contained extraction of information from semi-structured or unstructured text by Natural Language Processing methods. The search included only papers that were published since the year 2014 to ensure that the methods and the tools were state-of-the-art. Reasons for exclusion included:

- Preliminary paper (no evaluation of the solution included)
- The research is not in English
- The research is about translating natural language text
- The text processed are not English language
- The paper concentrates mainly on process mining
- The paper was not accessible.

The natural language processing task in papers varied and the selected papers included for example Text Mining from emails (Jlailaty, Grigori, & Belhajjame, 2017a, Jlailaty, Grigori, & Belhajjame, 2017b), Classifying Websites by Industry Sector and Business Process specific solutions like identifying process tasks from process descriptions (Leopold, van der Aa, & Reijers, 2018). Since Google Scholar had the most numerous results, the browsing was ended when there were not any interesting papers in two consecutive pages. The 18 papers selected for further inspection are in the appendix (Appendix 2).

The main research method of the thesis will be design science research method. The research will follow Design science Research Methodology Process (Peffers et al., 2007) to implement the design science research method. Definition of the objectives of a solution step will include evaluation of the literature research results for the possible tools and algorithms to be used in the design and development step.

2.3 Results of the Literature Review

This chapter will shortly reveal the most popular methods, the applications and tools where these methods were used and measures found in the literature review. The tables presenting the results in each section include only those measures, tools and methods that appear more than once. This does not mean that for example, the tool that appeared more than once is inherently more inadequate than a tool that has appeared in five papers. The purpose of this literature review is to shed light upon state of the art in the natural language solutions as it is one of the knowledge resources that are needed in defining the objectives for a solution step in the Design Science Research Method (Peffers et al., 2007).

2.4 Natural Language Processing Methods

The methods include linguistic methods like named entity recognition, statistical methods and classifiers like Support Vector Machines. The following table (TABLE 8) presents the methods that appeared in more than one paper.

TABLE 8 Natural Language Processing Methods that appeared more than once in the literature

Method	Appearances	Papers
Naïve Bayes	3	· De Medio, Gasparetti, Limongelli & Sciarrone, 2017 · Berardi, Esuli, Fagni, & Sebastiani, 2015 · Annervaz, George, & Sengupta, 2015
SVM	3	· Annervaz, George, & Sengupta, 2015 · Liu, Javed & Mcnair, 2016 · Leopold, van der Aa & Reijers, 2018
Decision Trees	2	· Annervaz, George, & Sengupta, 2015 · De Medio, Gasparetti, Limongelli & Sciarrone, 2017
NER	2	· Niboonkit, Krathu & Padungweang, 2017 · Iren & Reijers, 2017
Hierarchical clustering	2	· Jlailaty, Grigori & Belhajjame, 2017a · Jlailaty, Grigori, & Belhajjame, 2017b
word2vec	2	· Jlailaty, Grigori & Belhajjame, 2017a · Jlailaty, Grigori, & Belhajjame, 2017b

(to be continued)

TABLE 8 (to be continued)

tf-idf	5	· Jlailaty, Grigori & Belhajjame, 2017a · Jlailaty, Grigori, & Belhajjame, 2017b · De Medio, Gasparetti, Limongelli & Sciarrone, 2017 · Iren & Reijers, 2017 · Berardi, Esuli, Fagni, & Sebastiani, 2015
Latent Semantic Indexing	3	· Jlailaty, Grigori & Belhajjame, 2017a · Revindasari, Sarno & Solichah, 2016 · De Medio, Gasparetti, Limongelli & Sciarrone, 2017
Pos-tagging	5	· Ferreira, Thom, & Fantinato, 2017 · Leopold, van der Aa & Reijers, 2018 · Lindsay, Read, Ferreira, Hayton, Porteous, & Gregory, 2017 · Sawant, Roy, Parachuri, Plesse & Bhattacharya, 2014 · Mehmood, Iftikhar & Iftikhar, 2016
Cosine Similarity	2	· Sarno & Solichah, 2016 · Iren & Reijerss, 2017
Jaccard index	2	· Iren & Reijers, 2017 · Liu, Javed & Mcnair, 2016

The methods presented here are not always exclusive. The natural language processing pipeline may include multiple different methods. For example, the paper by Iren & Reijers included term frequency calculations (tf-idf), named entity recognition, Cosine Similarity, and comparisons with Jaccard index variant in their procedure (2017). Some papers did have usually exclusive methods to compare the results like Annervaz, George, & Sengupta, who compared Support Vector Machines, random forest, decision trees and naïve Bayes classification (2015).

2.4.1 Support Vector Machines

Support vector machines (SVM) are originally binary classifiers. It is an extension of the maximal margin classifier and support vector classifier. The maximal margin classifier's purpose is to find the maximal margin for the separating hyperplane boundary in the space populated by distinctly separated classes of training observations. In the case of where observations are not separable. Also, the changes observations that are closest to the hyperplane (support vectors) can have a too large effect on the hyperplane.

The support vector classifier allows the support vectors to be in the wrong side of the hyperplane (i.e. classified with the wrong label). Support vector classifiers' have only linear hyperplanes as boundaries which may make them inappropriate for many cases. Support vector machines, however, include a kernel, a function that can achieve, for example, binomial or radial boundary. (James, Witten, Hastie & Tibshirani, 2013.)

Annervaz, George, & Sengupta compared the support vector machine performance in extracting structural information from lease documents. Support vector machine outperformed random forest, naïve Bayes and decision tree classifiers (2015). Berardi, Esuli, Fagni, & Sebastiani used SVM to classify websites with multiple labels (2015). Leopold, van der Aa, and Reijers used SVM to classify potential tasks from textual description. By using this method, out of the 424 task instances, 342 were classified correctly. SVM was selected since according to they perform relatively well with small datasets, overfitting is not such a big problem as with the other classifiers and they are scalable (Leopold, van der Aa, and Reijers, 2018).

2.4.2 Hierarchical Clustering

Hierarchical clustering produces clusters by building a tree-like structure. The most common form of hierarchical clustering is bottom-up clustering which starts the clustering from the bottom, i.e. from the leaves where each leaf is an observation. These leaves fuse into branches and finally into one branch or trunk when describing in more arboreal terms. The height of the cut defines the number of clusters cut. Cutting at height 0 means that each of the observations belongs to their own cluster and cutting at the highest point means that there is only one cluster. (James et al., 2013, pp. 390-398.)

Jlailaty, Grigori & Belhajjame selected the Hierarchical clustering over the K-means clustering for three reasons. First, hierarchical clustering, unlike K-means clustering, included cluster and sub-cluster hierarchy similar to process and activity hierarchy. The second reason was "K-means is sensitive to cluster centre initialization." Lastly, K-means requires the number of clusters as an input, and in this case, the number of clusters was unknown. (2017a)

2.4.3 Term Frequency-Inverse Document Frequency

Term frequency-inverse document frequency (tf-idf) is a method originally developed for text retrieval. Term frequency refers to the frequency of the distinct terms in the document. Using term frequency alone to compare documents could, however, lead to high recall and low precision. Thus it is used with inverse document frequency that is an inverse function of how often the term is present in all the documents. (Salton & Buckley, 1988)

For example, the term "research" might be frequent in documents consisting of a variety of research papers, but it would be rather common in all the documents, and the inverse document frequency function would give the term a relatively lower weight than, for example, for the term "Business Process Management". Berardi et al. used tf-idf to classify websites together with other methods (2015).

2.4.4 Named Entity Recognition

Named entity recognition (NER) is a sub-task of information extraction to recognize different information units from unstructured text such as person, location or organization names (Nadeau & Sekine, 2007). In named entity recognition, each word recognized in the text as an information unit is given an entity label. Example of the named entity recognition can be seen in the next table (TABLE 9).

TABLE 9 Example of named entity recognition of the sentence: "Clarke was born in Minehead, Somerset, England in 1917".

Word	Named Entity
Clarke	PERSON
was	
born	
in	
Minehead,	LOCATION
Somerset,	LOCATION
England	LOCATION
in	
1917	DATE

Multiple methods have been implemented on named entity recognition such as SVM (Nadeau & Sekine, 2007). Iren & Reijers used named entity recognition for identifying terms that had specific meanings inside the organization (2017). Named entity recognition was mostly used as a preliminary method for preparing the text for the actual Natural Language Processing task for the research problem.

2.5 Tools

Tools include corpora, libraries, frameworks, software and databases used in the papers for Natural Language Processing purposes. The tools that appeared in more than one research papers can be seen in the following table (TABLE 10). The tools that were used in more than one paper had the same authors like in the papers by Leopold, Pittke, & Mendling, (2015) and Pittke, Leopold, & Mendling, (2015). Most of the tools that appeared more than once were software libraries that were either specialized in machine learning or natural language processing. Besides libraries, there were also two semantic networks, BabelNet and Wordnet. All the papers did not, however, report explicitly the tools that were used (Revindasari, Sarno & Solichah, 2016).

TABLE 10 Tools that appeared more than once in the literature

Tool	Appearances	Papers
BabelNet	2	· Pittke, Leopold, & Mendling, 2015 · Leopold, Pittke, & Mendling, 2015
NLTK	3	· Annervaz, George, & Sengupta, 2015 · Jlailaty, Grigori & Belhajjame, 2017a · Jlailaty, Grigori, & Belhajjame, 2017b
Scikit-learn	4	· Jlailaty, Grigori & Belhajjame, 2017a · Jlailaty, Grigori, & Belhajjame, 2017b · Pustulka-Hunt, Telesko & Hanne, 2018 · Annervaz, George, & Sengupta, 2015
Stanford NLP Tools	7	· Iren & Reijers, 2017 · Sawant, Roy, Parachuri, Plesse & Bhattacharya, 2014 · Leopold, van der Aa & Reijers, 2018 · Niboonkit, Krathu & Padungweang, 2017 · Mehmood, Iftikhar & Iftikhar, 2016 · Gao & Bhiri, 2014 · Lindsay, Read, Ferreira, Hayton, Porteous, & Gregory, 2017
Wordnet	3	· Leopold, van der Aa & Reijers, 2018 · Mehmood, Iftikhar & Iftikhar, 2016 · Gao & Bhiri, 2014
WEKA	2	· De Medio, Gasparetti, Limongelli & Sciarrone, 2017 · Leopold, van der Aa & Reijers, 2018
LibSVM	2	· Liu, Javed & Mcnair, 2016 · Berardi, Esuli, Fagni, & Sebastiani, 2015

2.5.1 Libraries

Libraries included toolkits created for Natural Language Processing such as GATE (Sawant, Roy, Parachuri, Plesse & Bhattacharya, 2014), Spacy (Ferreira, Thom & Fantinato, 2017) and OpenNLP (Annervaz, George, & Sengupta, 2015). Some of the libraries were method or algorithm specific like LibSVM (Liu, Javed & Mcnair, 2016; Berardi, Esuli, Fagni, & Sebastiani, 2015) and TreeSVM (Berardi, Esuli, Fagni, & Sebastiani, 2015).

Two papers had the same authors and used similar tools and methods in both papers to discover Business Process instances (Jlailaty, Grigori, & Belhajjame, 2017b) and Business Process Activities (Jlailaty, Grigori, & Belhajjame, 2017a) from email logs.

Stanford NLP tools in this literature review refer to multiple tools in the Stanford CoreNLP that includes Stanford Parser, Stanford POS-tagger, and Stanford NER. Stanford NLP tools are Java-based and open source. Iren & Reijers used Stanford parser for identifying the grammatical structure and lemmatization of the words (2017). Niboonkit, Krathu & Padungweang used Stanford's named entity recognition for discovering success factors in articles (2017). Stanford's NLP was also used for part-of-speech tagging (Mehmood, Iftikhar & Iftikhar, 2016; Sawant et al., 2014). Majority of the authors used Stanford parser for dependency analysis (Gao and Bhiri, 2014; Lindsay et al., 2017; Sawant et al., 2014; Leopold, van der Aa & Reijers, 2018).

Natural Language Toolkit (NLTK) is as the name implies a toolkit for natural language processing. NLTK is programmed in python and is open source (Bird, Klein, & Loper, 2009). Annervaz, George, & Sengupta used NLTK for ″some basic natural language processing steps″ (2015). Jlailaty, Grigori, & Belhajjame also used the NLTK for basic natural language processing tasks like removing punctuation and numbers (2017a, 2017b). They also used it for extracting named entities (2017b).

Scikit-learn is a machine learning library written in python (Pedregosa et al., 2011). Pustulka-Hunt, Telesko & Hanne, used scikit-learn for automated sentiment analysis of the comments in a Gig work platform (2018). Annervaz, George, & Sengupta applied scikit-learn for classical machine learning models to classify opportunities in legal texts (2015). Scikit-learn was used in hierarchical clustering emails into business process activities (Jlailaty, Grigori, & Belhajjame, 2017a) and business process instances (Jlailaty, Grigori, & Belhajjame, 2017b).

2.5.2 WordNet

Wordnet refers to a lexical database that consists of English language nouns, verbs, adjectives and adverbs and their meanings and relations to each other. WordNet can be used for word sense disambiguation (Jurafsky & Martin, 2009). Word sense disambiguation is useful; for example, when the task is to know does the word "bear" in the text refer to a large omnivorous mammal or an act of carrying something. It can also be used for analysing word similarities (Gao & Bhiri, 2014).

Since Wordnet also includes textual descriptions of each word, it was used for concept identification by Mehmood, Iftikhar & Iftikhar (2016). Similar Lexical databases are multilingual Babelnet and VerbNet. Babel is a multilingual lexical database that utilizes multiple sources like VerbNet and Wordnet. Pittke, Leopold, & Mendling used Babelnet to recognize homonyms and synonyms in activities of process models (2015). WordNet and VerbNet were used by Leopold, van der Aa, & Reijers (2018).

2.6 Measures

In the evaluation of the papers, the most common metrics were precision (12 papers), recall (11 papers) and F-measures (12 papers). The measures that appeared more than once are presented in the following table (TABLE 11).

TABLE 11 Measures that appeared more than once in the literature

Measure	Appearances	Papers
Recall	11	· Annervaz, George, & Sengupta, 2015 · Jlailaty, Grigori & Belhajjame, 2017a · Mehmood, Iftikhar & Iftikhar, 2016 · Revindasari, Sarno & Solichah, 2016 · Gao & Bhiri, 2014 · Jlailaty, Grigori, & Belhajjame, 2017b · De Medio, Gasparetti, Limongelli & Sciarrone, 2017 · Iren & Reijers, 2017 · Sawant, Roy, Parachuri, Plesse & Bhattacharya, 2014 · Leopold, van der Aa & Reijers, 2018 · Ferreira, Thom, & Fantinato, 2017
Precision	12	· Annervaz, George, & Sengupta, 2015 · Jlailaty, Grigori & Belhajjame, 2017a · Mehmood, Iftikhar & Iftikhar, 2016 · Revindasari, Sarno & Solichah, 2016 · Gao & Bhiri, 2014 · Sawant, Roy, Parachuri, Plesse & Bhattacharya, 2014 · Jlailaty, Grigori, & Belhajjame, 2017b · De Medio, Gasparetti, Limongelli & Sciarrone, 2017 · Iren & Reijers, 2017 · Leopold, van der Aa & Reijers, 2018 · Ferreira, Thom, & Fantinato, 2017 · Liu, Javed & Mcnair, 2016
Accuracy	4	· Annervaz, George, & Sengupta, 2015 · De Medio, Gasparetti, Limongelli & Sciarrone, 2017 · Pustulka-Hunt, Telesko & Hanne, 2018 · Ferreira, Thom, & Fantinato, 2017

(to be continued)

TABLE 11 (to be continued)

F-measure	12	· Berardi, Esuli, Fagni, & Sebastiani, 2015 · Jlailaty, Grigori & Belhajjame, 2017a · Mehmood, Iftikhar & Iftikhar, 2016 · Jlailaty, Grigori, & Belhajjame, 2017b · De Medio, Gasparetti, Limongelli & Sciarrone, 2017 · Iren & Reijers, 2017 · Sawant, Roy, Parachuri, Plesse & Bhattacharya, 2014 · Leopold, van der Aa & Reijers, 2018 · Ferreira, Thom, & Fantinato, 2017 · Liu, Javed & Mcnair, 2016 · Pustulka-Hunt, Telesko & Hanne, 2018 · Niboonkit, Krathu & Padungweang, 2017
Rand-index	2	· Jlailaty, Grigori & Belhajjame, 2017a · Jlailaty, Grigori, & Belhajjame, 2017b

Recall refers to the percentage of all the correct, true positives items retrieved. The precision is the percentage of how many of the retrieved items were true positive in the results. F-measure is a combination of recall and precision. F-measure is used because precision and recall alone do not present well the performance.

Liu, Javed & Mcnair used coverage instead of recall in evaluation entity retrieval from an employer knowledge base. Coverage is the percentage of queries where the system returns a non-null result. This was used since recall refers to all the accurately retrieved results, and in this case, all the existing correct answers were not known, and it was difficult to obtain. (2016.) Accuracy is a fraction of true positives and true negative prediction from the total population. Rand Index is a measure similar to accuracy, but it is applicable to evaluate clustering.

2.7 Conclusion

The papers in this literature review included a wide variety of methods and tools. Term frequency-inverse document frequency (tf-idf) and part-of-speech (POS) tagging were the most popular methods used in the papers. Majority of the papers used a method that was not used in the other papers. This was even more highlighted in the used tools. The ambiguous results indicate that there is not a one Natural Language Processing method nor a tool to solve all the problems. Recall, Precision and the F-Score would seem to be a valid measurement for evaluation of the design and for making the results comparable.

Some supporting methods and tools such as WordNet and its extension and named entity recognition are probably valuable in supporting the Design Science Research solution. To select a suitable method for the annotation of the process components from documents would require a comparison of multiple selected tasks. The next chapter presents the process of creating the Machine Learning model for the extraction of the business process names as named entities.

3 DESIGN AND DEVELOPMENT OF THE NAMED ENTITY RECOGNITION SOLUTION

This chapter represents the Design and Development step of the Design Science Research method. The objective is to build an information extraction solution that can recognize the processes as Named Entities. In practice, this solution uses named entity recognition capabilities of Spacy, a Natural Language Processing tool. This step includes gathering the business process documents, parsing the text into the appropriate format, training the model and evaluation of the model (FIGURE 5).

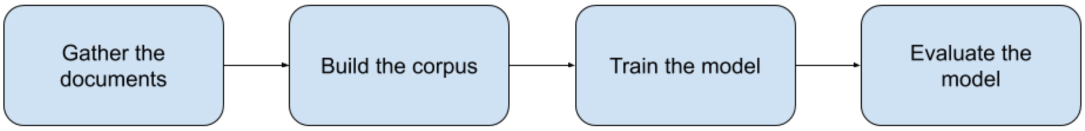


FIGURE 5 Steps in designing the named entity recognition solution

The artefacts that this design attempted to produce were the corpus and the prototype that demonstrate the pipeline. The outcome of this step did not, however, provide the wanted results and the shortcomings will be discussed in the Discussion section.

3.1 Gathering Example Documents

This section presents the documents that were selected for the corpus. The documents were searched from the document collections of the intergovernmental organizations. The organizations were the European Union's institutions. The documents that were selected to be included the corpus for the model contain explicitly process descriptions. Although recognizing documents that explicitly assert containing Business Processes might seem like an easy task, even the search tools provided documents that were mostly irrelevant for this purpose. Non-relevant documents consisted of guides for Business Process management system, how to draw Business Process Modelling Notations or memos that called for describing the business processes. Search phrases that were used to search the documents were "business process", "business processes". Since "business process" gave too many irrelevant results, the documents were also searched with the phrase "bpmn".

Search from the European Union, and its agencies document repositories provided four documents that are presented in the next table (TABLE 12). The code in the table represents an alias for the document. These aliases are used to refer to the Business Process documents in this research for the sake of brevity.

TABLE 12 Documents included in the corpus

Document Name	Code	Organization	Document Version	Year
DLV02.01 – Business processes	DLV	EUROPEAN COMMISSION (EU)	5.0	2018
NSW Prototype System Design Document SafeSeaNet	Ship	European Maritime Safety Agency (EU)	1.91	2015
TARGET Instant Payment Settlement User Requirements	Tips	ECB (EU)	1.0	2017
Customs Decisions Business User Guide	Customs	EUROPEAN COMMISSION (EU)	2.00	2017

3.2 Pre-processing of the Documents

Before named entity recognition could consume the training documents, the PDF files had to be processed into machine-readable plaintext format. The pre-processing of the documents included converting PDF to TXT files, cleaning the text files and splitting the text files and uploading files into the Brat annotation tool[5]. The cleaning, splitting and uploading were done to ensure the compatibility with the Brat annotator. There were some concerns when parsing the PDF to text that will be discussed in the discussion section (3.4).

3.2.1 Building the Corpus

The corpus was to build by extracting the text from PDF documents. The process of creating the corpus from PDF files is presented in the next figure (FIGURE 6).

[5] https://brat.nlplab.org/

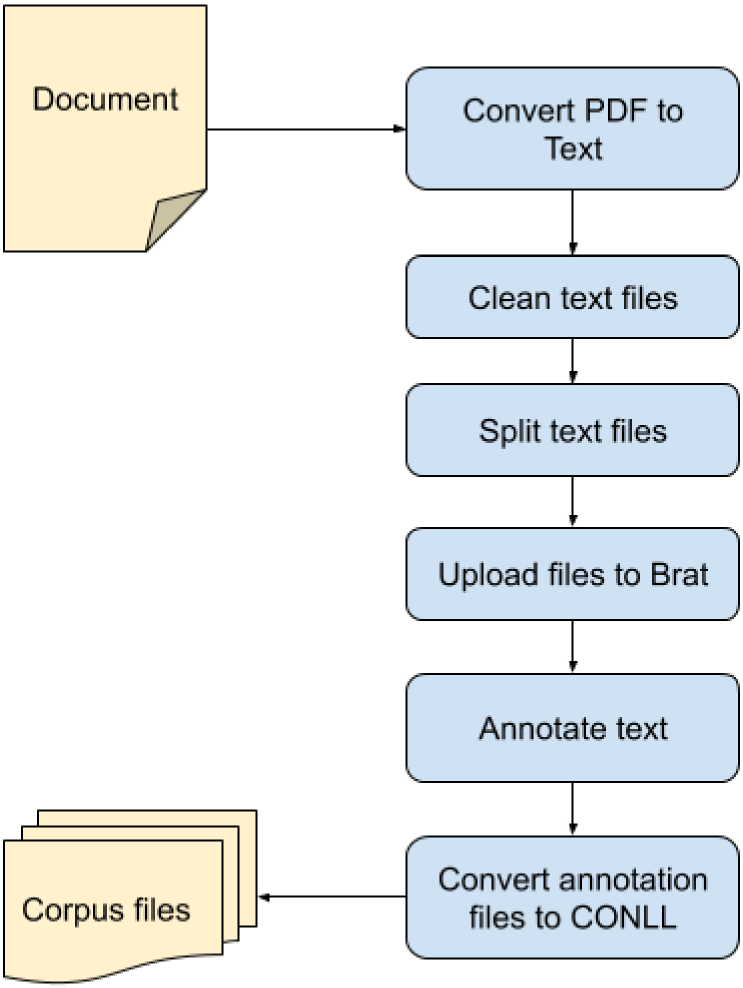


FIGURE 6 The process of converting documents into corpus

The corpus creation included the following steps:

1. **Converting PDF to TXT files**: Command line application pdf2text is used to convert the PDF file into a raw text file.
2. **Cleaning the text files:** Brat annotator had difficulties with white spaces at the beginning and the end of the line. These are cleaned with a script.
3. **Splitting the text files:** Brat annotator is slow with large text files, so the text files are divided into files containing 100 lines at maximum.
4. **Uploading files to Brat:** Since Brat runs as a service, the files are loaded manually into the filesystem along with the configuration and empty annotation files.
5. **Annotation of the files:** The process names appearing in the TXT files are annotated.

6. **Converting annotation files to CoNLL files:** The annotation files are converted to CoNLL format by using a script.

Although the first four steps were rather trivial text processing activities, it should be noted that raw text parsed from the PDF documents did lose relevant information and the raw text became challenging to comprehend. One weakness of this development step may be that documents that were identified as process documentation included process description tables and process diagrams. The reason for this is that a human reader can understand process documentation better. However, information that has been constructed for human reading becomes hard to consume by machines, and the parsed raw text becomes difficult to read for the human annotator.

3.2.2 Annotating

To train the model to recognize the necessary process parts, the model needs an example data of already annotated process parts. From DLV02.01 – Business processes document (DLV) the process and the step names were labelled. However, other documents did not include explicitly labelled steps. In this case, only the process names were annotated to build the first iteration of the prototype quickly. The documents were annotated by using the Brat annotator. After the annotation, files were converted into IOB CoNLL format. Where 'I' denotes word being inside a tag, 'O' that the word is outside the tag and 'B' word being the beginning of the tag. An example IOB format is presented in the next table (TABLE 13). The annotated corpus files were then used to train Spacy's named entity recognition model.

TABLE 13 A sentence in IOB format

Tag	Start Position	End Position	Word
O	2231	2234	The
O	2235	2242	purpose
O	2243	2245	of
O	2246	2249	the
B-Process	2250	2258	Download
I-Process	2259	2269	registered
I-Process	2270	2285	classifications
O	2286	2293	process
O	2294	2296	is
O	2297	2299	to
O	2300	2307	provide
O	2308	2314	access
O	2315	2317	to
O	2318	2321	SDG
O	2322	2337	classifications

3.2.3 Training

Spacy Natural Processing Language library was used as a tool to build a prototype and test pipeline. It was selected since it provided many of the tools out of the box such as Part of speech tagging, lemmatization, semantic relation parser and named entity recognition (NER). The named entity recognition feature was especially interesting since Spacy made it possible to create a custom Named Entities. A custom NER model was created by using this feature with 'PROCESS' as an only Named Entity label. Spacy uses convolutional neural network models for named entity recognition. In the literature research papers, Ferreira et al. used Spacy because of its accuracy, and it filled all their requirements as Natural Language Processing tool (2017).

Training of the Spacy was done according to the documentation[6]. The model contained only one label, "PROCESS". The annotated files were transferred into a format that was consumable by Spacy. Customs, Ship and Tips documents were used for the training as the DLV document was set aside for evaluation.

3.3 Evaluation

The evaluation in this step is analytical and more precisely dynamical analysis as described by Hevner et al. (2004) The design was mathematically evaluated by using information retrieval metrics such as Recall, Precision and F-measure.

DLV document was selected as the gold standard, and other documents were used for training. The manually annotated components in the gold standard will be treated as true positives, and the automatically annotated components will be separated into true positives and false positives. The manually annotated components that are not found in the automatic annotations are treated as false negatives. In other words, the manually annotated process names are the correct answers that the named entity recognition system is trying to find those correct answers. When evaluating the system's ability to retrieve business process components there probably are no definitive causal claims made, since it would mainly present that this system retrieves this well in from these documents. The primary purpose of the named entity recognition has been discovering proper nouns that might not have other meanings, unlike the process names. The further utility and the causal claims might need a field experiment.

As can be seen from the following table (TABLE 14), the results of this evaluation are unacceptably low for any solution. The first evaluation was done by using only the custom document as a training corpus, then Ship document was added, and finally, Tips document was added. Purpose of this was to see how the different combinations and increasing the corpus size affect the named entity recognition.

[6] https://spacy.io/usage/training#ner

TABLE 14 Named entity recognition solution's evaluation results when attempting to extract processes from the DLV document

Training Documents	Precision	Recall	F-measure
customs	25.00	0.60	1.17
customs ship	0.00	0.00	0.00
customs Ship tips	60.00	1.80	3.50

The next table (TABLE 15) presents all the entities that the solution offers as a process have the word "management" or "manage" in them. At maximum, the solution could recall 1.8 % of the correct entities, which is three entities.

TABLE 15 All the entities that the solution recognized as a correctly or incorrectly as a process name

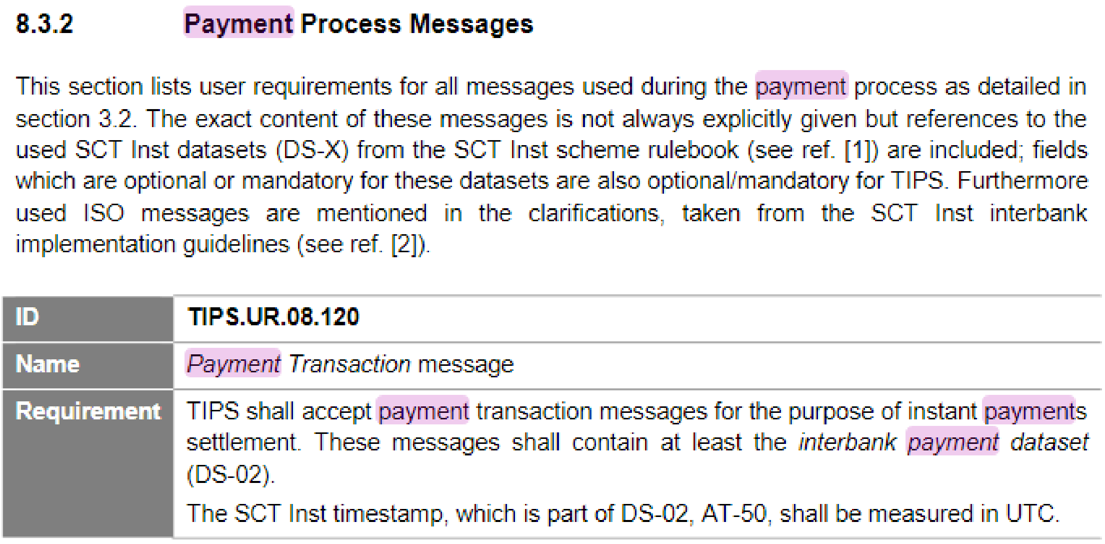
customs	customs Ship	customs Ship tips
Manage link sets	-	Manage link sets
Manage		Manage link sets Description
Manage		Manage link sets
The Information Management business		Manage Registered Classification
		Manage Registered

As the results show, the proposed solution does not currently discover process names. Following discussion section will look into reasons for these results and discuss the next steps of the Design Science Research Method.

3.4 Discussion of the Pipeline as a Solution

The results of this preliminary experiment were subpar, and the larger corpus did not have barely any effect on the results. The recall was exceptionally poor. This, however, is an excellent opportunity to reflect these problems and challenges in Natural Language Processing. There are a few possible reasons why this demonstration does not produce adequate results, even for a prototype.

- **The training corpus is too small:** the training of neural networks may require massive corpora.
- **The annotations were biased:** there was only one annotator, and there was not any evaluation made. This could be solved with ideally multiple annotators and validating the annotation, for example, with Cohens Kappa.
- **Information was lost in parsing:** relevant information was lost when parsing the PDF documents.
- **Variability of the documents:** the processes in the documents were too different from being useful for recognizing the processes from other documents.
- **Process names were too ambiguous:** especially process names with one common word may have homonyms or polysemes (FIGURE 7). Process names are not precisely Named Entities.
- **Machine readability versus human readability:** the purpose of the business process documents is to be human-readable first. Different structures fit for different purposes.



8.3.2 Payment Process Messages

This section lists user requirements for all messages used during the payment process as detailed in section 3.2. The exact content of these messages is not always explicitly given but references to the used SCT Inst datasets (DS-X) from the SCT Inst scheme rulebook (see ref. [1]) are included; fields which are optional or mandatory for these datasets are also optional/mandatory for TIPS. Furthermore used ISO messages are mentioned in the clarifications, taken from the SCT Inst interbank implementation guidelines (see ref. [2]).

ID	**TIPS.UR.08.120**
Name	*Payment Transaction* message
Requirement	TIPS shall accept payment transaction messages for the purpose of instant payments settlement. These messages shall contain at least the *interbank payment dataset* (DS-02). The SCT Inst timestamp, which is part of DS-02, AT-50, shall be measured in UTC.

FIGURE 7 Example of process name ambiguity in a document

Although from these results, it could be possible to iterate back to the design and development step and continue to build a corpus or develop a solution by using other Natural Language Processing methods and tools. Instead, this research will return back to defining objectives for the solution to utilize the theoretical background for the possible solution. Reason for this is the previously mentioned weaknesses in the solution that contributed to the failure. Also, the annotation is a time-consuming task and rather error-prone. How large corpus this would require is uncertain. Since documents containing Business Process descriptions are a somewhat distinct form of communicating knowledge, this research will continue by examining the documents through the lens of genre theory.

4 GENRE ANALYSIS OF THE BUSINESS PROCESS DESCRIPTIONS

Historically genre has been mainly a core concept of literary studies and literary theory, but it has since gained multiple interdisciplinary perspectives (Heikkinen & Voutilainen, 2012, pp. 21-22). Biber, Connor & Upton conclude from inspection of the similar and sometimes interchangeable terms "register" and "genre" that the genre is a "culturally recognized message type". Genres also vary based on context and purpose. (2007)

Yates & Orlikowski similarly define the genre in the organizational context as a form of communication that is characterized by structural, linguistic or substantive conventions (1992). For example, the memo and Curriculum Vitae are widely recognized forms of communication that have a distinct purpose and structure.

In the Business Process Management domain, genre approach can be used in multiple ways: to examine the message created inside a process, or in the management of the Business Processes itself. In this research, the purpose is to design a system capable of recognizing the process components from the documents. Business process descriptions are not however restricted into one genre nor a genre system.

4.1 Multimodality and the Move Analysis

Mikkonen defines a multimodal document as a presentation that may include any linguistic and visual methods (2012, p. 296). Concept of multimodality is implemented in this research since when communicating in the context of Business Process Management using process documentation, multimodal features such as tables and diagrams (especially workflows) are used. In practice, this means that the move analysis will include coding of the titles, tables and figures.

A move is a functional unit in a text that contributes to the purpose of the genre (Connor & Mauranen, 1999; Biber et al., 2007). Swales established the division of the genre into moves and steps (1990 cited through Bateman, 2008, pp. 194). Steps in the move are functions for achieving the purpose of the move (Biber et al., 2007).

As it was previously mentioned, the exact label of the genre is not going to be attached to these documents, and this thesis is not arguing if the process descriptions itself are a separate genre. Viewing the business process description as a genre or a move might help to recognize and classify description by Machine Learning means.

In this research, the documents will be analysed on the terms of structure by applying the guidelines introduced by Delin, Bateman & Allen (2002). Their suggestion includes the following levels presented in the next table (TABLE 16).

TABLE 16 Levels of structure by Delin et al. (2002)

Content structure	the structure of the information to be communicated
Rhetorical structure	the rhetorical relationships between content elements; how the content is 'argued'
Layout structure	the nature, appearance and position of communicative elements on the page
Navigation structure	the ways in which the intended mode(s) of consumption of the document is/are supported
Linguistic structure	the structure of the language used to realise the layout elements

Especially analysis of the layout structure would be useful for this research. Delin et al. used these levels to analyse Wildlife Encyclopedia page on Bengal tiger (2002). Bateman, Delin & Henschel applied these levels of structure to analyse the print version and the web version of The Guardian newspaper (2007).

Although this analysis turns the direction of the research towards a more qualitative analysis, the purpose of this analysis is still to provide valuable information for extracting the Business Process components. Also, the results of this analysis should provide knowledge for further steps of this Design Science Research.

The main level that is looked into in this research is the layout structure. Since PDF documents are mainly intended to be viewed by humans and from the machine reading perspective, the text is unstructured and the semantic data is mainly for layout purposes. However, the layout could give some clues for the Business Process components. Thus, the question for this part of the research is: where are the process names? The next figure (FIGURE 8) presents an example layout of DLV02.01 Business Processes document where layout units that contain Business Process names are highlighted with grey. The left margin of the figure has the codification of each element and notes. The right margin shows the elements and group of the elements that may appear more the once. The repetition is denoted with an ellipsis.

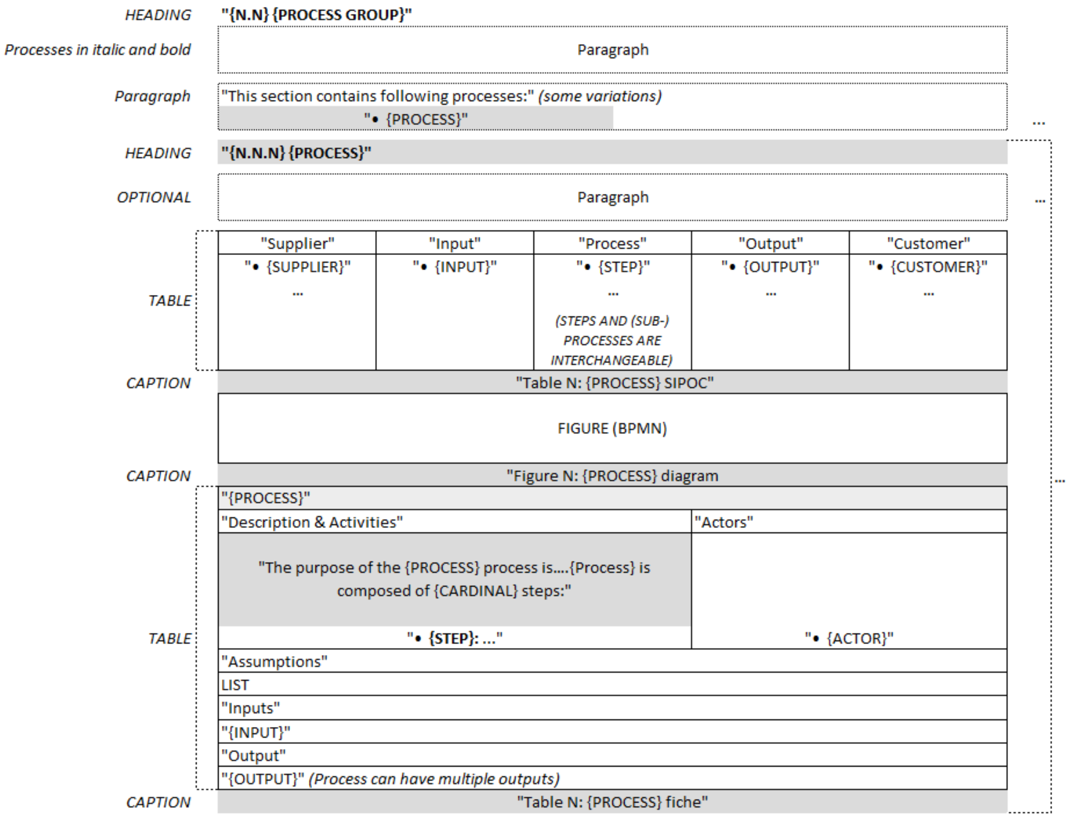

HEADING
"{N.N} {PROCESS GROUP}"
Processes in italic and bold
Paragraph
Paragraph
"This section contains following processes:" (some variations)
"• {PROCESS}"
...
HEADING
"{N.N.N} {PROCESS}"
OPTIONAL
Paragraph
...
TABLE
"Supplier"
"Input"
"Process"
"Output"
"Customer"
"• {SUPPLIER}"
"• {INPUT}"
"• {STEP}"
"• {OUTPUT}"
"• {CUSTOMER}"
...
(STEPS AND (SUB-) PROCESSES ARE INTERCHANGEABLE)
CAPTION
"Table N: {PROCESS} SIPOC"
FIGURE (BPMN)
CAPTION
"Figure N: {PROCESS} diagram
...
"{PROCESS}"
"Description & Activities"
"Actors"
"The purpose of the {PROCESS} process is....{Process} is composed of {CARDINAL} steps:"
TABLE
"• {STEP}: ..."
"• {ACTOR}"
"Assumptions"
LIST
"Inputs"
"{INPUT}"
"Output"
"{OUTPUT}" (Process can have multiple outputs)
CAPTION
"Table N: {PROCESS} fiche"

FIGURE 8 Layout structure of the DLV document as an example

This analysis attempts to answer the following questions: Where are the processes and their components in the sections describing processes? Further questions are: what are the layouts of the descriptions and how does this affect information retrieval? Also, there are document related questions like what are the audiences and purposes of the documents? These are taken into account since the communicative purpose is one factor that affects the language used in communication (Biber et al., 2007, p. 8). We want to see if these also affect the multimodality and the communication of the processes in the documents. Although this analysis utilizes the tools from genre theory, move analysis, and multimodality, this chapter's analysis is not aiming to conduct exhaustive genre research, but it aims to increase the knowledge of the state of problems mentioned in the Design Science Research Method's defining the objectives for a solution activity (Peffers et al., 2007).

The data in this analysis included the European Union's documents that were used in the previous chapter's named entity recognition solution (TABLE 17) and additional United Nation's Business Process Analyses (TABLE 18). These documents belong to the case study series by United Nations Economic and

Social Commission for Asia and the Pacific (UNESCAP)[7]. Business Process Analyses were consistent with each other, so they were all analysed in one section.

TABLE 17 European Union Documents

Document Name	Code	Organization	Document Version	Year
DLV02.01 – Business processes	DLV	EUROPEAN COMMISSION (EU)	5.0	2018
NSW Prototype System Design Document SafeSeaNet	Ship	European Maritime Safety Agency (EU)	1.91	2015
TARGET Instant Payment Settlement User Requirements	tips	ECB (EU)	1.0	2017
Customs Decisions Business User Guide	customs	EUROPEAN COMMISSION (EU)	2.00	2017

TABLE 18 United Nations ESCAP Business Process Analyses

Document Name	Code	Organization	Year
Business Process Analysis of Import of Wool to Nepal	BPA1	UNESCAP (UN)	2017
Business Process Analysis of Import of Light Motor Vehicles from the third Countries to Bhutan via Kolkata Port	BPA2	UNESCAP (UN)	2017
Business Process Analysis of Export of Plastic Kitchenware and Tableware from Bangladesh to Bhutan	BPA3	UNESCAP (UN)	2017
Business Process Analysis of Import of Lentil from Nepal to Bangladesh	BPA4	UNESCAP (UN)	2017
Business process analysis of trade procedures in selected Central Asian countries	BPA5	UNESCAP (UN)	2015

The next sections will present the observations and content structure of the analysed Business Process Documents.

[7] https://unnext.unescap.org/content/business-process-analysis-simplify-trade-procedures-case-studies

4.1.1 Customs Decisions Business User Guide

Customs document's purpose is to inform "the end-users of the Customs Decisions Management System (CDMS) and of the EU Trader Portal (EU TP)" from the business perspective. The document is "intended for readers with various backgrounds and operational roles within the Customs Decisions related system domain". So it is not written primarily for business process management experts.

The customs document differs the most from the other documents in terms of audience and layout. It does not have tables to condense the process information, and it does not use any standardized workflow notation. Also, the page numbering starts after each process chapter. Visually the document uses more colour and other semiotic means than the other documents, like for example, informative colourful boxes that stress the time constraints. The text extracted from these elements is reasonably consistent.

As can be seen from the process descriptions in the figure (FIGURE 9), the chapter has a process name as a chapter heading. After the heading, the stakeholders are listed with bullet points.

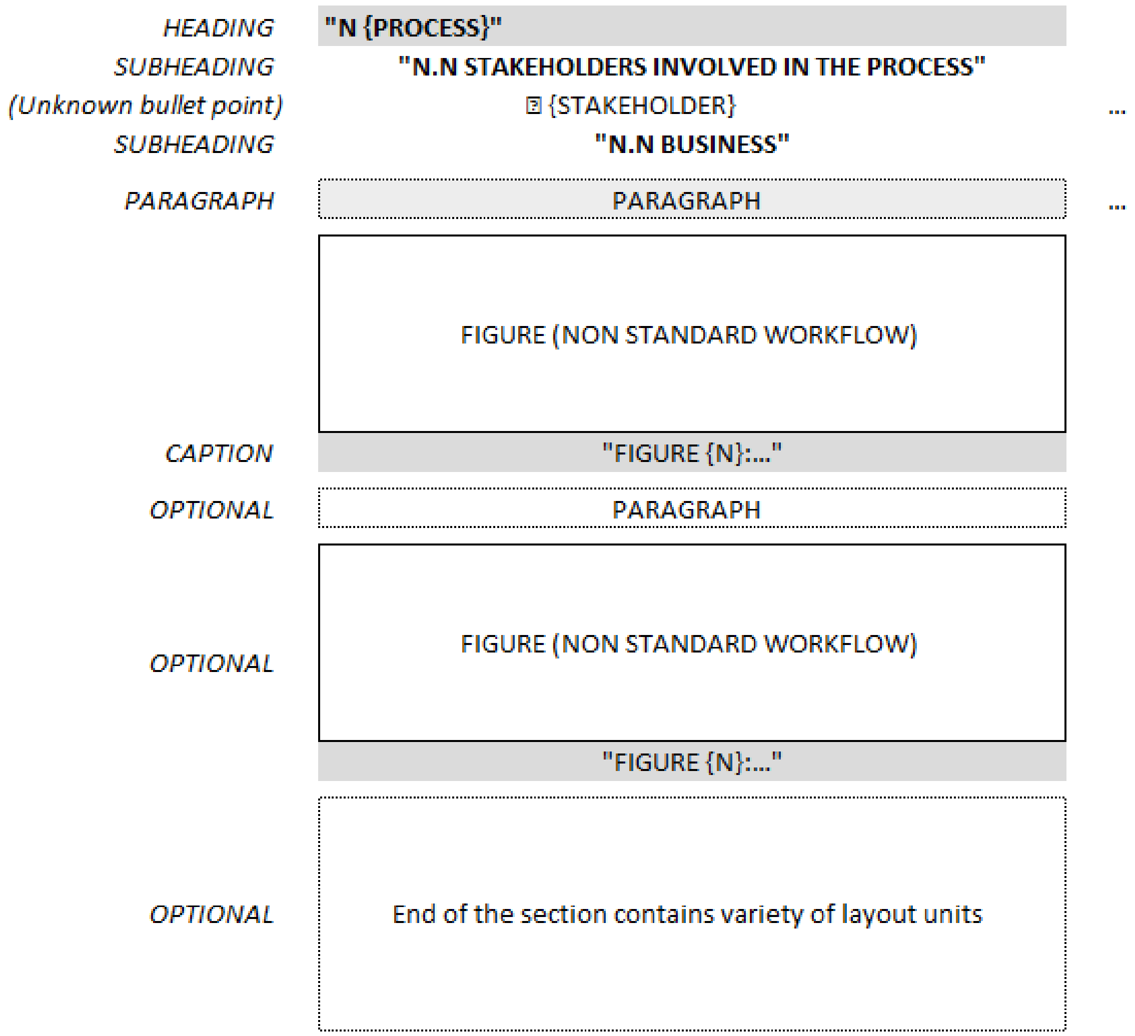


FIGURE 9 Customs Decisions Business User Guide's business process description structure

Before the first figure, there are paragraphs that with few sentences describe what happens in the process and positions the process in the high-level process. These paragraphs may mention the process name explicitly but not always. Also, the process reference may differ between the heading and the paragraph. For example, ”Take Decision” process becomes the decision-taking process. This ambiguity may cause difficulties in information extraction.

4.1.2 NSW Prototype System Design Document

NSW Prototype System Design Document defines the National Single Window (NSW) system. NSW is part of the Safeseanet system developed by the European Maritime Safety Agency. So basically this document’s context is in the maritime vessel operations. The purpose of the document is ”to present a comprehensive architectural overview / the technical details of the NSW system components”. The target audiences of the document are ”system designers and system builders", and it is somewhat technically oriented. SHIP document's process descriptions are inside a table layout. The table heading mentions the process explicitly by preceding with word "process". The process tables only include process name, a business process figure and a description of the steps as can be seen from the next figure (FIGURE 10).

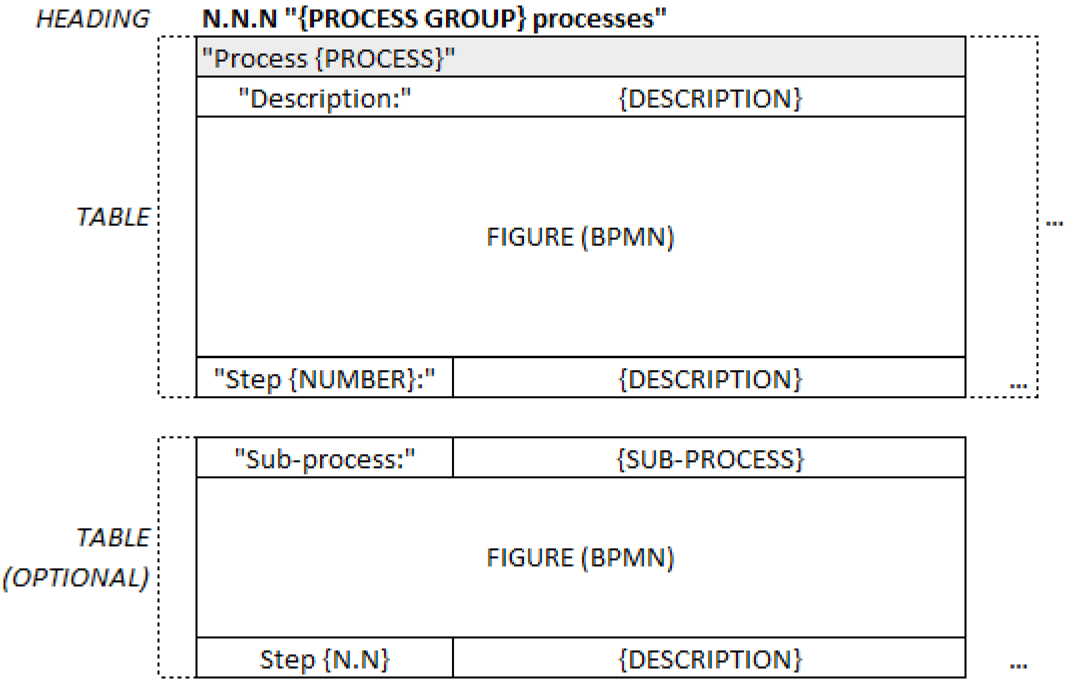
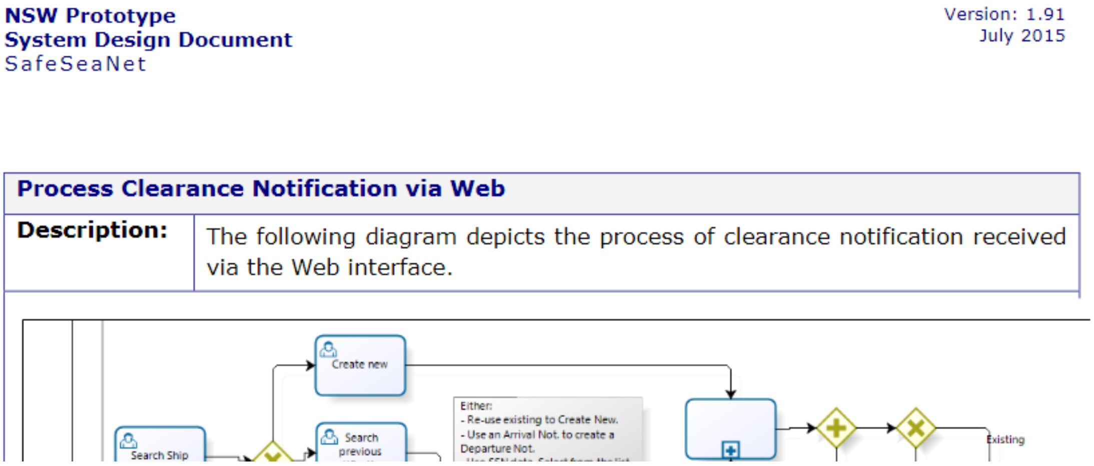


FIGURE 10 NSW Prototype System Design Document's Process Description Structure

However, the document includes use cases that may contain a reference to the business process. In the literature research, Sawant et al., used NLP methods for use cases (2014) their solution might be useful for similar documents.

The descriptions for process and steps are very brief. The length of the description is usually only one sentence. For example, the description for Clearance Data Routing process is "The following diagram depicts the process of routing clearance notification data to the Authorities" and description for Manage Ship process is "Defines the steps executed by the National Administrator to manage Ships". Example of the process description can be seen in the following figure (FIGURE 11).



NSW Prototype
System Design Document
SafeSeaNet

Version: 1.91
July 2015

Process Clearance Notification via Web	
Description:	The following diagram depicts the process of clearance notification received via the Web interface.

Create new
Search Ship
Search previous
Either:
- Re-use existing to Create New.
- Use an Arrival Not. to create a Departure Not.
Existing

FIGURE 11 Process name and the description of the process in the document

4.1.3 TARGET Instant Payment Settlement User Requirements

The primary purpose of the TARGET Instant Payment Settlement User Requirements document is to inform the reader about the user requirements in the instant payment system. TARGET Instant Payment Settlement (TIPS) is a service aimed for instant payments in the euro system and regulated by the European Central Bank. The document does not explicitly define what the audience of the user requirement document is. However, the content is relatively technical (it defines data formats and ISO standards that are used). Although the process descriptions are relatively coherent and explicit, only a minor part of the chapters include business processes described in detail. Most of the document pages are dedicated to the user requirements as the name of the document implies.

The paragraph contents that are below heading (see FIGURE 12) differ between processes. The payment process' paragraph describes the start and end events of the process. The recall process' paragraph explains briefly what are the two recall processes. Liquidity management has a similar description. However, liquidity management includes an overview that describes the liquidity management processes almost step by step.

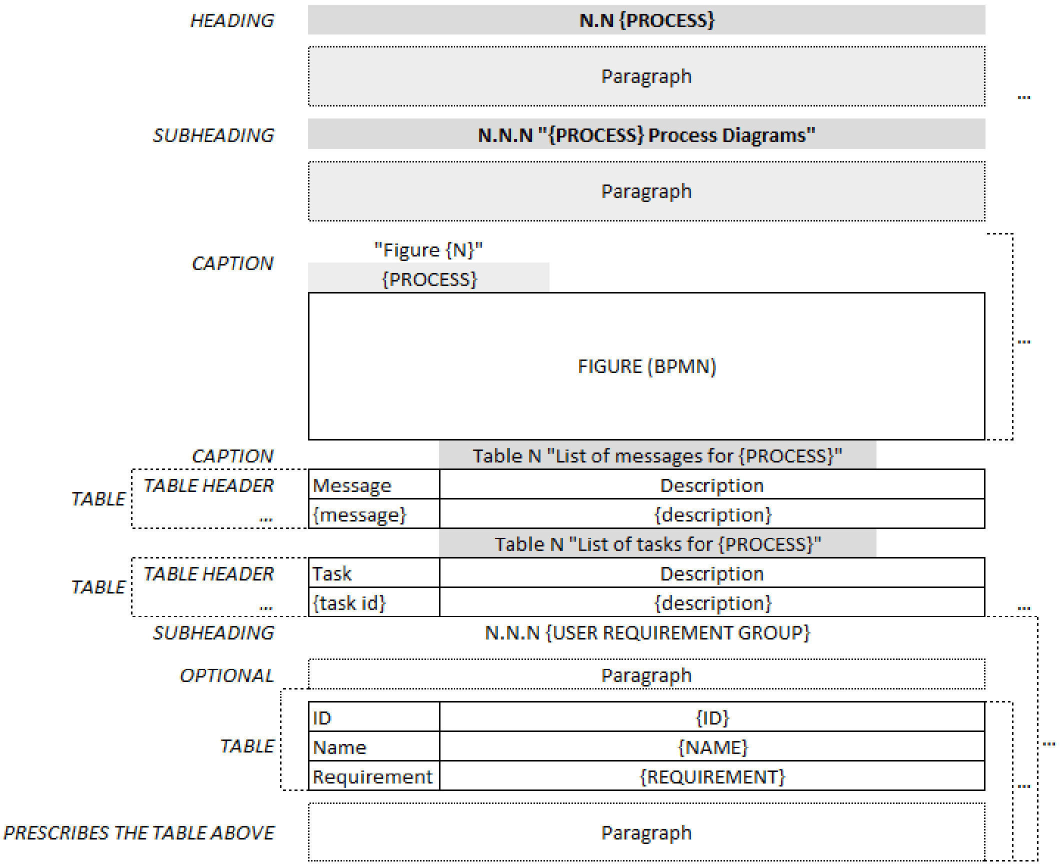


FIGURE 12 Process description structure of the TARGET Instant Payment Settlement User Requirements document

Parsing the unstructured text from the TIPS document breaks the heading number and the heading into two rows. Also, the header and footer break the unstructured text flow. The table header and the text content from the table cell are now in an inconsistent order. This makes the document more challenging for natural language processing.

The tasks in the TIPS documentation have an identifier. The ID is used as a reference between the BPMN diagram and the table. The BPMN diagram includes the name of the task and the ID (FIGURE 13).

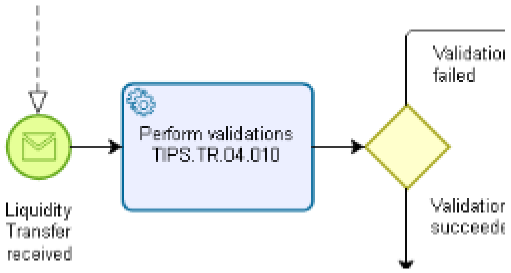
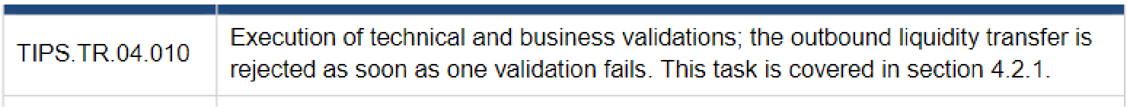


FIGURE 13 Presentation of the task in the process diagram

In the list of tasks, the tasks are referred only by their IDs (FIGURE 14). If using an ID as a reference would be a common practice, a machine could easily recognize it. However, this was the only document that used an ID this way, and BPMN diagrams do not always contain extractable text.



TIPS.TR.04.010	Execution of technical and business validations; the outbound liquidity transfer is rejected as soon as one validation fails. This task is covered in section 4.2.1.

FIGURE 14 Presentation of a task in the list of tasks

4.1.4 DLV02.01 – Business Processes

The DLV02.01 – Business Processes document describes the requirements and the business processes in the Single Digital Gateway. Single Digital Gateway provides access to information, procedures and assistance services for the citizens of the single market[8]. The purpose of the document is to analyse "the business processes necessary to put in place and run the different IT tools of the Single Digital Gateway". The document does not explicitly describe its target audience; however, the purpose indicates that the audience is process experts and the technical personnel.

The DLV document is dedicated almost exclusively to business processes as the document name indicates. However, the SIPOC model presented uses steps and processes interchangeably, and most of the steps are marked in the BPMN figure as subprocesses. This ambiguity might prevent the machine from extracting subprocesses and texts correctly, primarily if the elements in the process column are extracted as columns.

When targeting the figures and the headings, they include the process name. The processes are explicitly called as processes in process chapter paragraphs and the description & activities columns in the process fiche. Occasionally the process is also bolded. Before the process sections, the processes are listed in the Process group preceding a paragraph "This section contains the following processes". The structure of the document is presented in the next figure (FIGURE 15).

[8] https://ec.europa.eu/growth/single-market/single-digital-gateway_en

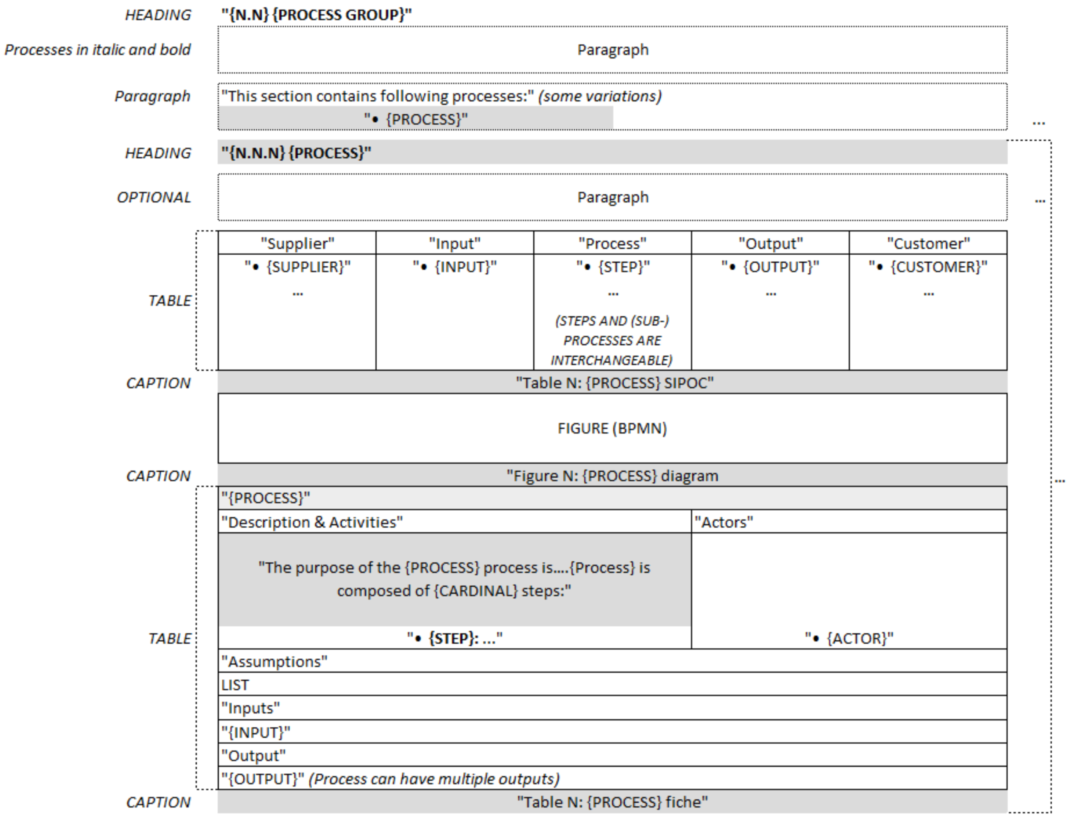

HEADING
"{N.N} {PROCESS GROUP}"
Processes in italic and bold
Paragraph
Paragraph
"This section contains following processes:" (some variations)
"• {PROCESS}"
...
HEADING
"{N.N.N} {PROCESS}"
OPTIONAL
Paragraph
...
TABLE
"Supplier"
"Input"
"Process"
"Output"
"Customer"
"• {SUPPLIER}"
"• {INPUT}"
"• {STEP}"
"• {OUTPUT}"
"• {CUSTOMER}"
(STEPS AND (SUB-) PROCESSES ARE INTERCHANGEABLE)
CAPTION
"Table N: {PROCESS} SIPOC"
FIGURE (BPMN)
CAPTION
"Figure N: {PROCESS} diagram
TABLE
"{PROCESS}"
"Description & Activities"
"Actors"
"The purpose of the {PROCESS} process is....{Process} is composed of {CARDINAL} steps:"
"• {STEP}: ..."
"• {ACTOR}"
"Assumptions"
LIST
"Inputs"
"{INPUT}"
"Output"
"{OUTPUT}" (Process can have multiple outputs)
CAPTION
"Table N: {PROCESS} fiche"

FIGURE 15 Process description structure of the DLV02.01 – Business Processes document

This document has the most consistent structure in terms of Business Process Management. This document is very explicit about the processes and their components. The document would probably be an ideal example of a Business Process documentation for this research.

4.1.5 UNESCAP Business Process Analyses

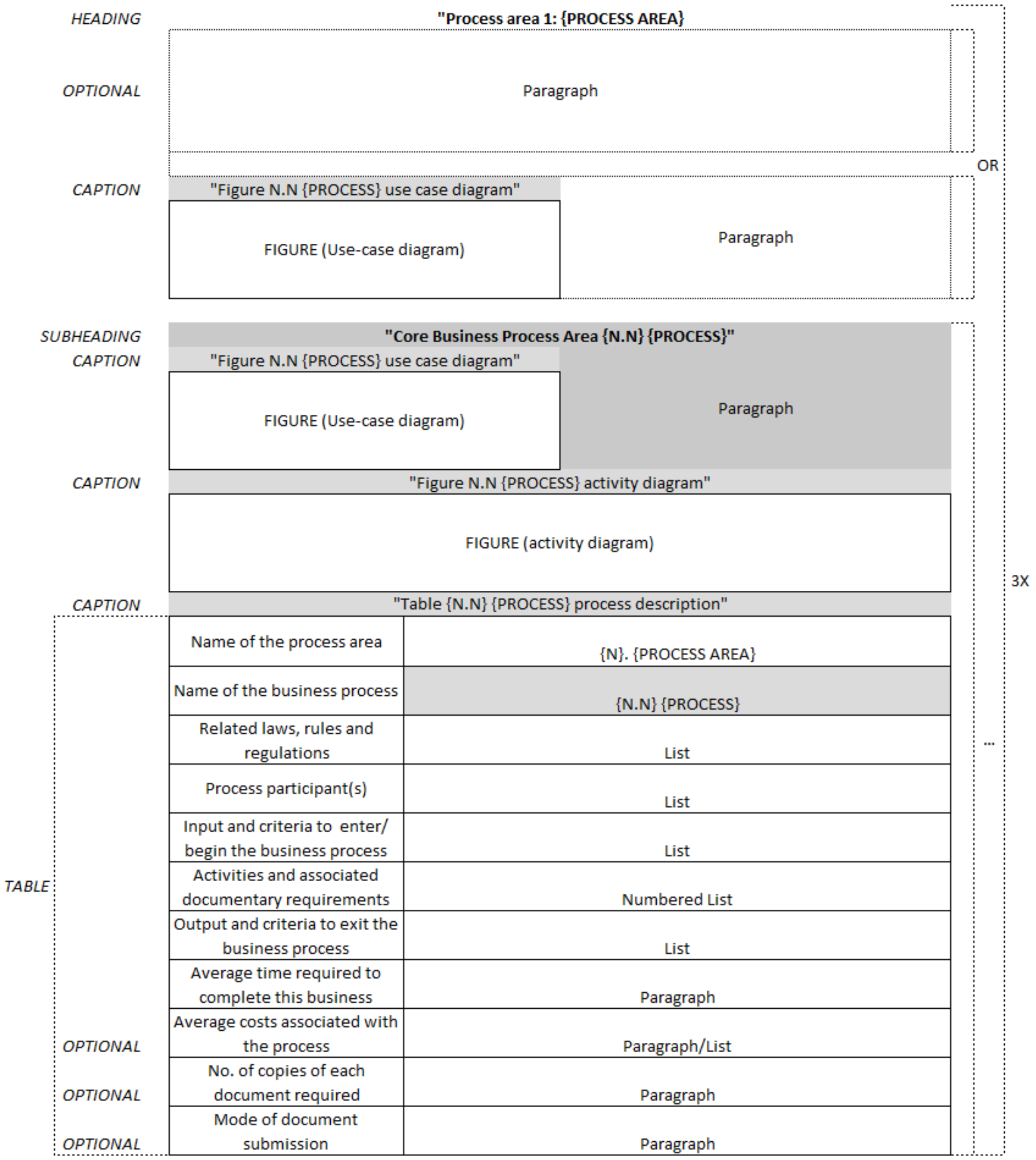
This examination includes Five Business Process Analysis (BPA) documents that were not included in the corpus of the named entity recognition design. The purpose of the BPA documents is to describe the current business processes in a specific international trade context, recognize bottlenecks, and thus give recommendations for the relevant authorities and inform the shareholders. The UNESCAP Business Process Analyses differ from documents provided by the organizations of the European Union in that UNESCAP BPA documents have predefined guidelines that the documents follow (United Nations ESCAP, 2012). Since the documents are mainly centred around business context, UNESCAP Business Process Analyses lack any technical architecture.

The titles of the documents answer to what and where questions in quite a specific style, e.g., “Business Process Analysis of Export of Plastic Kitchenware and Tableware from Bangladesh to Bhutan.” The Business Process Analysis Guide instructs the processes to be named with a “descriptive verb-noun phrase”

(United Nations ESCAP, 2012). In the documents, the process names are quite detailed, for example, “Conduct the registration of goods which arrived at the Bishkek customs clearance place.”

Documents contained more statistical information in the process description like average time. The statistics were further analysed to evaluate the bottlenecks of the trade processes at the end of the documents.

All the BPA documents contain process description sections that are grouped into Buy, Ship, and Pay process areas. Each process area introduces its processes in a paragraph or a use-case diagram. The document structure is presented in the following figure (FIGURE 16).


HEADING
"Process area 1: {PROCESS AREA}
OPTIONAL
Paragraph
OR
CAPTION
"Figure N.N {PROCESS} use case diagram"
FIGURE (Use-case diagram)
Paragraph
SUBHEADING
"Core Business Process Area {N.N} {PROCESS}"
CAPTION
"Figure N.N {PROCESS} use case diagram"
FIGURE (Use-case diagram)
Paragraph
CAPTION
"Figure N.N {PROCESS} activity diagram"
FIGURE (activity diagram)
3X
CAPTION
"Table {N.N} {PROCESS} process description"
...
TABLE
OPTIONAL
OPTIONAL
OPTIONAL

Name of the process area	{N}. {PROCESS AREA}
Name of the business process	{N.N} {PROCESS}
Related laws, rules and regulations	List
Process participant(s)	List
Input and criteria to enter/ begin the business process	List
Activities and associated documentary requirements	Numbered List
Output and criteria to exit the business process	List
Average time required to complete this business	Paragraph
Average costs associated with the process	Paragraph/List
No. of copies of each document required	Paragraph
Mode of document submission	Paragraph

FIGURE 16 UNESCAP Business Process description structure

The process descriptions contain one table, figures (use-case or activity diagram or both), and paragraphs. Table condenses the information about the business process providing the most relevant aspects of the process such as process name, process area, participants, activities, time consumption and optionally statistical data on costs and document requirements.

4.1.6 Results

The following table (TABLE 19) shows the layout components that include process names in each document. All the documents except the Ship document had process names in the headings. Also, the captions included process names as often as headings.

TABLE 19 Layout elements that contain Business Process names

Layout Component	BPA	DLV	SHIP	TIPS	CUSTOMS
CAPTION	X	X		X	X
HEADING	X	X		X	X
PARAGRAPH	X	X		X	X
TABLE	X	X	X		

The DLV and the BPA documents had the most fine-grained business process description. These documents had more business process components in the layout structure than the other documents. The table in the appendix (Appendix 4) shows the different business process elements that appear in the process descriptions.

The BPA documents differ from other documents in that they are describing as-is processes, and thus they have a more specific process and activity names. We can also argue that since the purpose is to inform about the bottlenecks and to give recommendations. Thus the underlying goal is to get the relevant authorities and the shareholders who operate in these trade contexts to take action to change their processes. In the ship, tips and DLV documents the purpose is to inform about the system in somewhat technical detail. These three documents seem to aim to inform the technical readers on how to integrate or implement the systems in their organizations. The processes are ”to-be” processes that are about to be implemented. The customs document’s purpose was also to inform about the system. However, the document’s audience is not as technically oriented. Thus customs document avoids technical details and describes processes on the end-user level. Purposes and audiences of the documents can be seen from the following table (TABLE 20).

TABLE 20 The purpose and the audience of the documents

Document	Purpose	Audience
Customs	inform about the system	end-users with variable backgrounds and roles
SHIP	inform about the system	″system designers and system builders″
TIPS	inform about the system	technical audience
DLV	analyse the business processes and IT tools required by the system	process experts and technical personnel
BPA Documents	inform about the bottlenecks and give recommendations.	authorities and shareholders

The process naming is somewhat more ambiguous in these documents, probably because the audience may include multiple organizations in different countries. Customs document differs the most from the other documents as it is a user guide for end-user and for this reason it includes more semiotics to highlight essential points and understanding of the processes does not require competence in BPMN or UML diagrams. However, the process naming is still quite similar to ship, tips and DLV documents.

4.2 Discussion

The genre theory and discourse analyses offer multiple methods to analyse the documents. One of these is the move-step analysis that divides the text into sections that fill the purpose inside the genre. A move may include several steps to fulfil its purpose. The other is a multimodal content analysis that looks into the layout and structure of a document.

In this chapter, the structure and the content of the business process descriptions in nine documents were compared. Also, this chapter looked into the purposes and the audience of the documents. The goal of this analysis was to inspect what constrictions and possible solutions for the information extraction do these document pose.

The business process descriptions have been referred to as unstructured text, and indeed, the text that the natural language processing tools consume are unstructured. In practice, the PDF documents are parsed into plain text for Natural Language Processing. Nevertheless, when parsing documents, much of the information is lost. Especially in the business process domain where multiple visual tools like graphs and diagram are being used to propagate the business process knowledge. Following components of the PDF documents seem to cause problems when parsing to plain text format:

- Tables
- Headers and Footers
- Text in diagrams
- Different text styles.

Especially parsing of the tables is problematic since tables are rather ubiquitous in the process descriptions. Tables also often included the process names in the compared documents.

The analysis also shows that the audience and the purpose have some correlation with the language. The UNESCAP's Business Process Analysis has a specific ″as-is″ processes that the authors want the audience to improve, and these processes are named unambiguously. The other documents contain more ambiguously named processes like ″Payment″ and ″Login″ as an extreme example. Unambiguously named processes might be easier to extract and recognize.

The multimodal content analysis would indicate that process names appear in the headings in 8 out of 9 cases. In the SHIP document, the process names were in the tables but not as easily machine recognizable heading. Still, the classification of headings could be one approach to discover business processes.

The layout analysis was qualitative, and the number of analysed documents was very small. Indeed, these nine documents do not present all the document types that the information worker might seek. It should also be noted that the current traditional office documents and PDF documents are not the only existing means to present the business process. Additional documentation may be found, for example, from Business Process Management and Modelling applications. Also, genres may change (Orlikowski & Yates, 1994). One of these changes can be that the documentation happens in other formats than PDF documents. This change should be taken into account or at least monitored if the Natural Language Processing methods become viable tools in Business Process Management. Although this analysis may not be the most profound contribution to the genre theory, it has given some information to plan for feasible solutions in terms of design science research methods objective definition.

The conclusion of the genre analysis is that process names are most likely to be found in the headings of the process documents. Even if the process names are not always in the headings, it could be a low hanging fruit, for example, for the extraction of the process names and finding the chapter where the relevant process information is located. Discovering process names from headings should not be the only method to extract the process. The extraction from heading could be the primary method to discover processes that could be validated with other methods. Alternatively, process discovery could be validated by the processing of the headings, or heading classification could be one of the parallel methods for information extraction. The next chapter will describe discovering process names from the headings through classification.

5 CLASSIFICATION OF HEADINGS

Next step is to develop and design a solution that can extract process names from a document's headings. This solution would be useful for discovering processes from a large set of documents. For example, a knowledge worker could get a list of documented processes in the organization without exploring manually every single document and without any previous knowledge of the processes. This solution would also be useful for further automatic discovery of process knowledge when knowledge worker does not know which documents and sections of the documents contain references to process documentation. It is more likely that these sections and documents contain information about the activities, events and actors of the processes.

The content analysis revealed that most of the process description documents had process names in the headings. A set of headings of which some are process names, and some are not, can be seen as a classification problem. Classification is part of the machine learning methods where observation are assigned to a class, and the classifier uses these observations to train the algorithm to predict the class of observation of which class is not known.

This development step implemented scikit-learn machine learning library to build a classifier solution. Scikit-learn is a python library that supports multiple different Machine Learning tasks. It was selected because of its ease-of-use

and multiple available classification methods. The classification evaluation compared seven classifiers simultaneously. Further information about the classifiers can be found from the Classifiers section (5.2).

Scikit-learn's Python API made it possible to prototype the pipelines rapidly using Google's Colab Notebook. During the development, four different pipelines were created. Compositions and of each pipeline their results are reported in the pipeline section (5.4). Scikit-learn also included cross-validation tools that are described in the validation section (5.3).

The headings were manually picked from the table of contents or the document headings. The real-life implementation of this solution would require sophisticated PDF document parsing to separate the headings and the table of contents automatically. This research, however, concentrates on the classification after the extraction of the headings. This research skips the automatic structure extraction since the classification is more relevant to Business Process context, and the solution can extend to HTML, Office document and PDF document formats.

After the extraction, non-process headings were labelled with "O", and process headings with "PROCESS" tag. The following Data section describes the data used in the classification in more detail. This chapter attempts to answer the following questions: which pipeline has the best performance, and what are the best classifiers for this solution? Since the pipeline development here is an iterative development challenge, pipelines from two to four have their own hypotheses.

5.1 Data

The training data was made from the headings of the process description documents. Some of these headings were process names, and some were not. The dataset in the heading classification included documents in the content analysis and one additional document, the European Central Bank's T2S Business Process Description (TABLE 21). The headings were gathered manually from the documents' Table of Contents.

TABLE 21 Source documents of the headings

Document Name	Code	Organization	Document Version	Year
DLV02.01 – Business processes	DLV	EUROPEAN COMMISSION (EU)	5.0	2018
NSW Prototype System Design Document SafeSeaNet	Ship	European Maritime Safety Agency (EU)	1.91	2015
TARGET Instant Payment Settlement User Requirements	Tips	ECB (EU)	1.0	2017
Customs Decisions Business User Guide	Customs	EUROPEAN COMMISSION (EU)	2.00	2017
BUSINESS PROCESS DESCRIPTION	T2S	ECB (EU)	1.4	2016
Business Process Analysis of Import of Wool to Nepal	BPA1	UNESCAP (UN)		2017
Business Process Analysis of Import of Light Motor Vehicles from the third Countries to Bhutan via Kolkata Port	BPA2	UNESCAP (UN)		2017

(to be continued)

TABLE 21 (to be continued)

Business Process Analysis of Export of Plastic Kitchenware and Tableware from Bangladesh to Bhutan	BPA3	UNESCAP (UN)		2017
Business Process Analysis of Import of Lentil from Nepal to Bangladesh	BPA4	UNESCAP (UN)		2017
Business process analysis of trade procedures in selected Central Asian countries	BPA5	UNESCAP (UN)		2015

The dataset contains 674 headings. Majority of the headings were non-process headings. Nearly two-thirds (~ 65 %) of the headings are non-process headings as can be seen from the following table (TABLE 22).

TABLE 22 Percentages of the heading tags

Tag	Count	Percentage
O	438	64.99%
PROCESS	236	35.01%
Total	674	100.00%

The number of process headings and the ratio of process headings varied notably between the documents. As can be seen in the figure (FIGURE 17), the customs document has the most significant amount of non-process headings.

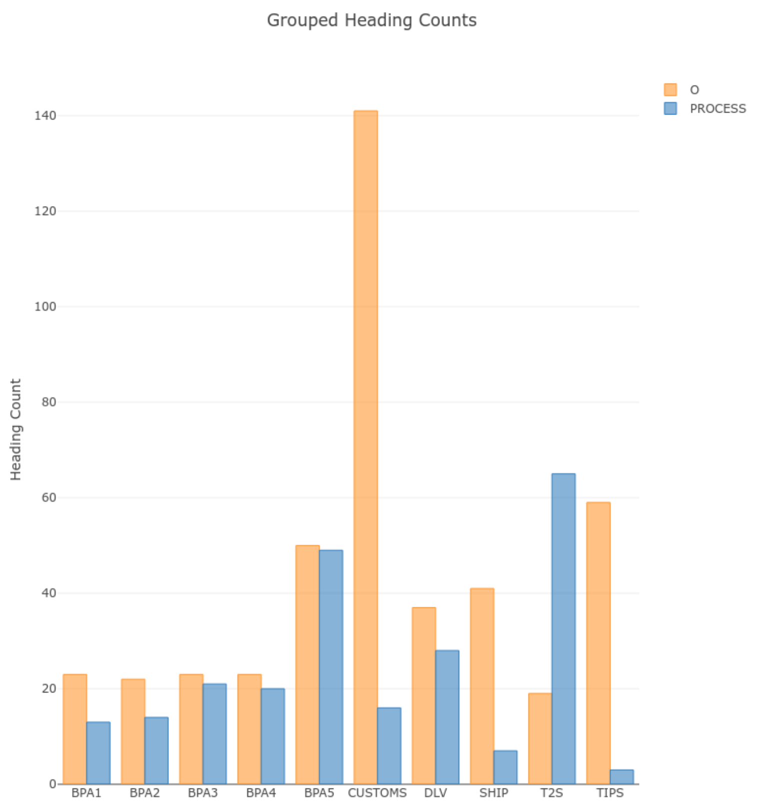


FIGURE 17 Grouped bar chart of the headings

Of all the non-process headings ~32% are headings from the Customs document, and ~48 % of the process headings are from t2s document and the BPA5 document. The following graphs visualize portions of the non-process headings (FIGURE 18) and process headings (FIGURE 19).

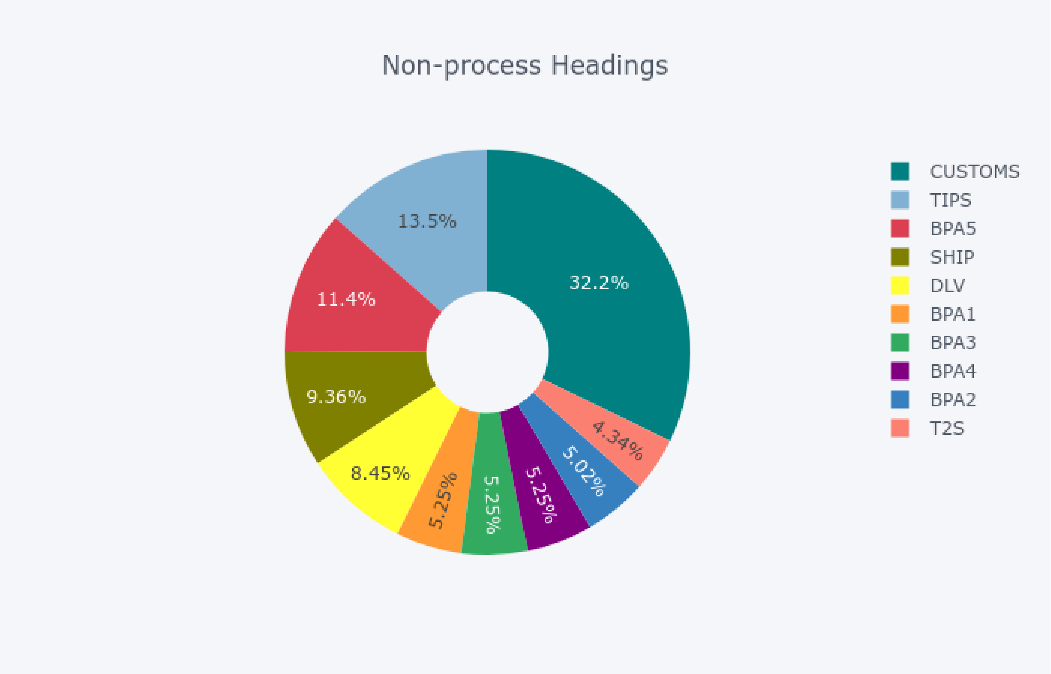


FIGURE 18 Portions of each document's non-process headings

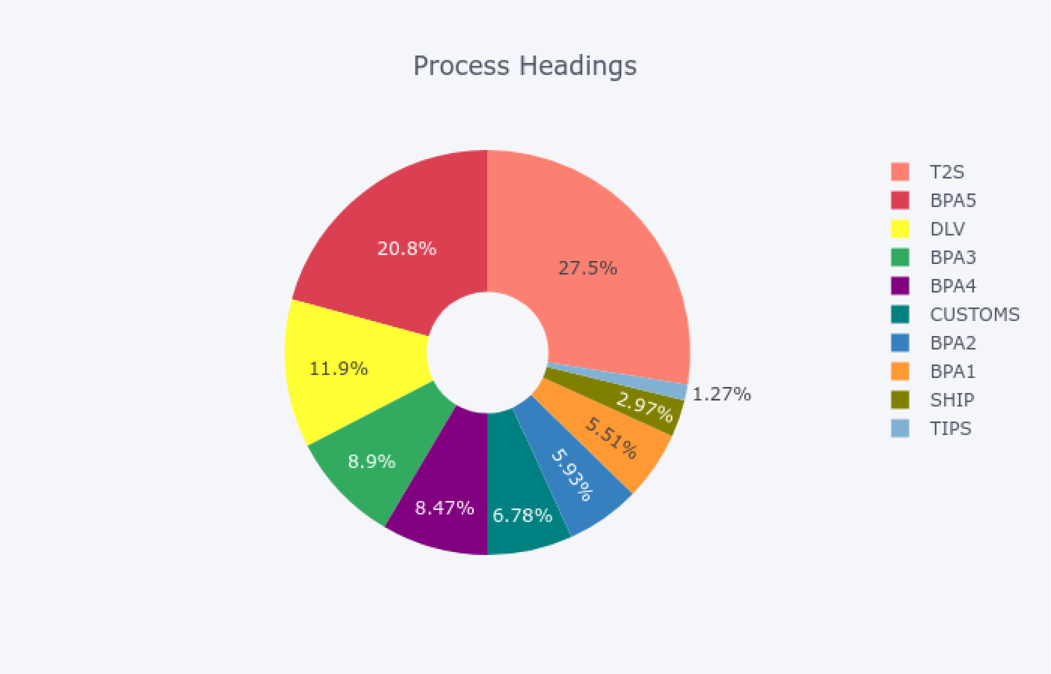


FIGURE 19 Portions of each document's process headings

Further inspection of the data might reveal possible weaknesses in the data that may skew the classifiers. The customs document has the most non-process headings, which might seem like a problem, especially since the customs document is not explicitly a business process document, as it was mentioned in the Genre Analysis chapter (chapter 4). However, it does provide headings that would be considered as business process related but not clearly as process names like

"Stakeholders involved in the Process". Since it would be important to separate these business process related headings from the process names, the inclusion of the customs document is arguable.

The most substantial proportion of the Process headings are from Target2Securities (T2S) document. A large portion of processes in a single document may cause challenges for the classifiers since T2S process names are not verb-noun phrases but noun-verb phrases (e.g. "CANCELLATION OF SETTLEMENT INSTRUCTION").

Another possible weakness in the classifier solution is that it treats the headings as independent elements. This might pose some ambiguity for the classification since there is some lack of context. As it was mentioned in the Genre Analysis, ambiguous processes like "Login" might be difficult to recognize as processes. When looking at the whole Table of Contents or the text context where they appear in the document, the processes become somewhat more apparent.

Drawing conclusions from the dataset and then applying these conclusions to classifiers and preparation of data may lead to overfitted model. Overfitted model would be only useful for predicting the data that it is trained with and trying to classify new data could lead to incorrect classifications. For this reason, the presumptions from the data should be drawn from Linguistics and Business Process Management Domain. For example, the naming conventions of the Business Processes are used in this research to argue the use of part-of-speech tagging. All the documents have been selected because they contain process names. Arguably despite some speculated problems, this provides a presentative data for the classifiers, and the data and the classifier configurations will remain unchanged between different pipelines to give us consistent metrics for evaluation.

5.2 Classifiers

Classifiers that were selected for the pipeline evaluation are presented in the next table (TABLE 23). Multiple classifiers were evaluated to see which are suitable classifiers for the design of the solution and also to see the differences between pipelines. These classifiers were included in all the pipelines, and their parameters stay the same between the pipelines. Naïve Bayesian, random forest and decision tree classifiers also appeared in the literature research.

TABLE 23 Classifiers used in the evaluation of the pipelines

Classifier	Abbreviation
Multinomial Naïve Bayes	MNB
Stochastic Gradient Descent	SGD
Bernoulli Naïve Bayes	BNB

(to be continued)

TABLE 23 (to be continued)

K-Nearest Neighbour	KNN
Random Forest	RF
Decision Trees	DT
Logistic Regression	LR

The abbreviation is used in this research to refer to these classifiers. Next sections will look briefly into the classifiers' basic mechanisms and the implementations of these classifiers in the literature.

5.2.1 Stochastic Gradient Descent

Gradient descent classification's purpose is to minimize the error between prediction and the actual observation. In order to achieve this error reduction, the gradient descent attempts to find the lowest point of the loss function. In the stochastic gradient descent (SGD), the weight that is estimated for the input to predict the output is updated after the training of each randomly picked example (Bottou, 2010). Unlike, for example, batch gradient descent, the stochastic gradient descent picks only one example. SGD has been used, for instance, in movie sentiment analysis of the IMDB reviews (Tripathy, Agrawal & Rath, 2016).

5.2.2 Logistic Regression

The outcome of the logistic regression is the probability between 0 and 1. Logistic regression is applicable in binary classification. The scikit-learn documentation recommends this for one-versus rest classification problems (1.1. Generalized Linear Models, n.d.).

Lee & Liu used logistic regression on a newsgroup text classification and deemed it as an effective method (2003). Lee et al. compared the logistic regression alongside with decision tree classifier, K-nearest neighbour, support vector machine in the topic classification of the twitter messages. In this comparison, in terms of accuracy, the logistic regression fared worse than the other methods. (2011.)

5.2.3 Decision Trees and Random Forest

Decision trees and random forest classifiers were found in the used methods in the literature review of this research. As the name suggests decision tree is a tree that consists of nodes that branch according to if-else decisions, and in the final leaf the presumed class is assigned (James et al., 2013, pp. 311-313). Depending on the decision boundary, the decision tree may provide a better fit to the training data (James et al., 2013, pp. 315).

Decision trees were part of the comparison in the text categorization of the Reuters articles along with a "variant of Rocchio's method for relevance feedback", naïve Bayes and linear support vector machine (LSVM). The decision trees and the LSVM produced a high classification accuracy compared to the other methods. (Dumais, Platt, Heckerman & Sahami, 1998.) Xhemali, Hinde & Stone compared decision tree, neural networks and naïve Bayes in classification for training purposes. In this research, naïve Bayes outperformed decision trees and neural networks. (2009.)

In random forest classification, decision trees are built from the training samples. The training data and the variables for creating the decision trees are selected randomly (Ali, Khan, Ahmad & Maqsood, 2012). Akinyelu & Adewumi used the random forest for spam classification, where it outperformed the previous Machine Learning method that was deemed as the best (2014).

5.2.4 K-Nearest Neighbours

K-Nearest Neighbour (KNN) classification uses the *k* closest training observations to determine the class of given observation (James et al., 2013). For example, if the *k* was 3, the label of the three nearest observations to the unlabelled observation would determine the label.

KNN has been researched for example in personality classification from twitter texts (Pratama & Sarno, 2015); categorization of online documents (Trstenjak, Mikac & Donko, 2014) and classification of news texts (Lan, Tan & Low, 2006).

5.2.5 Naïve Bayes

Naïve Bayes classifiers were also observed in the Literature Review. Bayesian classification uses Bayes' rule to assign a class to observation from given values probabilistically. Naïve Bayes assumes that these values that are the attributes of the observation are independent of each other. Multinomial naïve Bayesian classifier is able to classify text by using multinomially distributed features. In the text classification context, Bernoulli naïve Bayes classifier treats words in the texts as a binary vector (McCallum & Nigam, 1998).

Manning, Raghavan & Schütze recommend using Bernoulli naïve Bayes with shorter documents and with fever variables. Multinomial naïve Bayes can handle longer documents with multiple variables better (2008, p. 268). McCallum & Nigam compared multinomial naïve Bayes and Multi-variate Bernoulli naïve Bayes classifiers in webpage classification. Multinomial naïve Bayes outperformed the Multi-variate Bernoulli naïve Bayes. (1998.)

5.3 Validation

Cross-validation was used to evaluate the classifier pipelines. The metrics used in the evaluation were precision, recall, F1-score, received operating characteristics (ROC) and area under the curve (AUC). The ROC and AUC evaluation were conducted separately from the precision, recall and F1-score. Received operating characteristics in this research compared the relation of true positives and false positives. The area under the curve is the area under the ROC curve. In the best scenario, the ROC curve reaches the top left corner (AUC = 1.0), and in the worst scenario, the curve is a straight diagonal line (AUC = 0.5). The cross-validations with the ROC measure was done with built-in tools in Scikit-learn[9]. Cross-validation used stratified K-folds and group K-fold validations that were available in scikit-learn. In stratified K-folds cross-validation strategy, the data was split into *k* sub-samples and evaluated *k* times. The classes in the one sub-sample were predicted by using the remaining sub-samples as training data in each iteration (FIGURE 20). Stratified K-fold also retains the ratio of the classes in the testing set as in the original set.

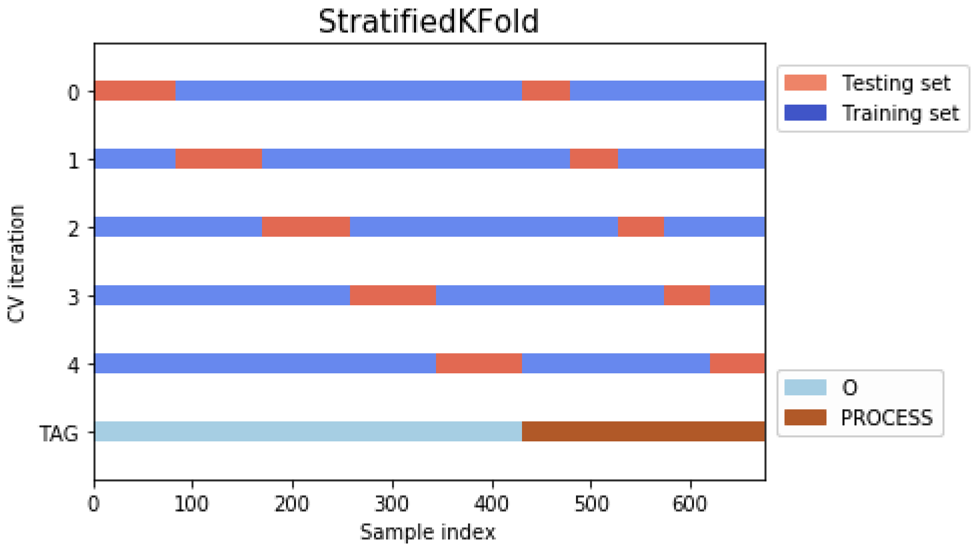


FIGURE 20 Rations of the training and testing data in stratified K-fold cross-validation

Unlike stratified K-fold validation, which splits the observations into equal-sized parts, the group K-fold validation splits the data according to predefined group labels. Groups here present the documents that provide the process names and the headings (FIGURE 21).

[9] https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc_crossval.html

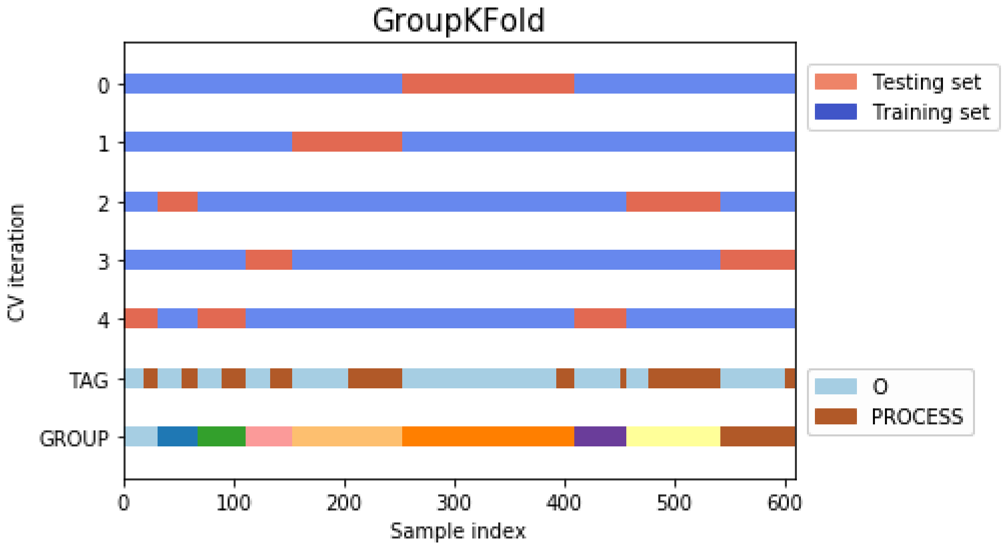


FIGURE 21 Rations of the training and testing data; tag portions and the groups in group K-fold cross-validation

The group K-fold was used in this research since group validation can prevent the headings from the same document to be in the testing set and in training set at the same time. Group K-fold is thus used to provide a more reliable validation together with stratified K-fold validation. Group K-fold will help us evaluate how well the model would perform when a new document was presented to the model as an input. Thus, it could be argued that the group K-fold presents an evaluation that would predict a real-world classification more accurately.

In this evaluation, there were five cross-validation iterations (*k=5*) in both group K-fold and in the stratified K-fold. Both of these cross-validation were run with all the classifiers, and the results were averages of the iterations.

5.4 Pipelines

In the final design and development step, four different pipelines were created. All the pipelines had the same classifiers with the same configuration. The way the data for classifiers was pre-processed was the main difference between pipelines. This was done to prevent the overfitting, which might have occurred if classifiers were configured continuously to get the best results from the evaluation. All the pipelines included scikit-learn's CountVectorizer, TfidfTransformer and a classifier. The pipelines were evaluated with seven different classifiers that were described in the classifier section. CountVectorizer and TfidfTransformer are described in more detail together with Pipeline 1. The solutions from the Pipeline 2 to Pipeline 4 utilize Spacy for part-of-speech tagging and the third pipeline utilized NLTK's Wordnet implementation and the word similarity algorithms. The first three experimental pipelines' evaluation results were compared and based on these results; the fourth pipeline was developed and evaluated. The fourth pipeline differed from others by using the phrases that contained the headings as observations. The last pipeline is discussed separately in the final resulting pipeline chapter (chapter 6). The following table (TABLE 24) summarizes the methods, tools and the hypotheses of the pipelines.

TABLE 24 Methods, tools and hypotheses of the pipelines

Pipeline	Methods and tools	Hypotheses
Pipeline 1	· vectorization · tf-idf	
Pipeline 2	· vectorization · tf-idf · part-of-speech	Adding part-of-speech tagging increases the overall results of the classifiers in terms of F1-score and the AUC compared to Pipeline 1 (H1).
Pipeline 3	· vectorization · tf-idf · part-of-speech · Business Process Taxonomy verb list · Wordnet · Wu-Palmer similarity	Implementing similarity measures of Business Process Taxonomy verbs increases the overall results of the classifiers in terms of F1-score and the AUC compared to Pipeline 2 (H2).
Pipeline 4	· vectorization · tf-idf · part-of-speech · context	When headings are presented with more context, the classifier will perform better in terms of F1-score in both stratified cross-validation and group cross-validation than the other pipelines (H3).

5.4.1 Experimental Pipeline 1

Scikit-learn's classification pipeline consists of feature extraction and classifier. In the primary scikit-learn text classification pipeline, CountVectorizer and TfidfTransformer extract the features.

CountVectorizer tokenizes the strings, creates a vocabulary from them and counts the occurrences of the tokens. In this case, the tokens were character sequences separated by whitespace. When n-gram range was set between 1 and 2, from the process name "create feedback report", following unigram and bigram sequences are created:

- "create"
- "feedback"
- "report"
- "create feedback"
- "feedback report."

CountVectorizer produces a matrix where rows present the index in the vocabulary, and the rows present a text. The elements of the matrix show how many times the word in the vocabulary appears in the text.

The output matrix of the CountVectorizer is an input for the TfidfTransformer, which counts the term frequency-inverse document frequency (tf-idf) for each value in the matrix. These frequency values are then passed to the classifier as parameters for training to predict the class they belong to (process or non-process). The next figure (FIGURE 22) presents the simplified visualization of Pipeline 1 with only the unigram matrix and vocabulary.

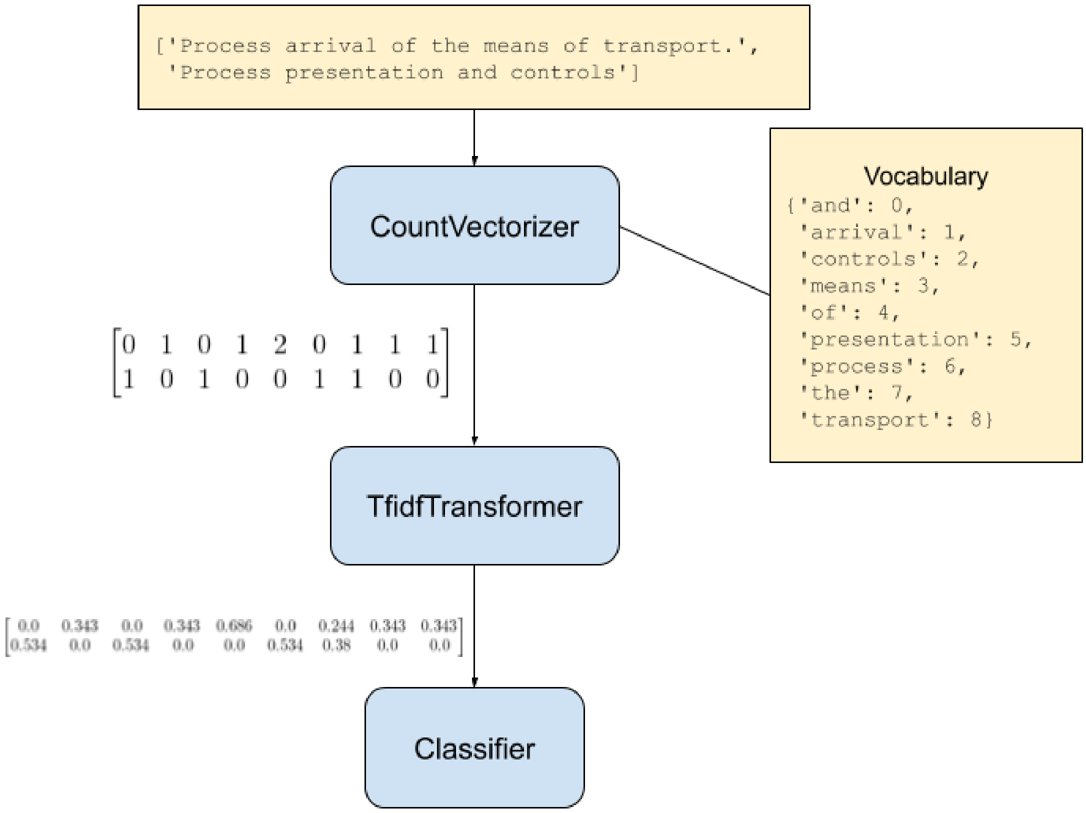


FIGURE 22 Components of the first pipeline

5.4.2 Results of Experimental Pipeline 1

These observations include the area under the curve (AUC), precision, recall and the F1-score. The results also include the fit time and score time. The fit time is the time taken by the fitting of the training data. The score time is the time taken by the scoring of the test data. These time observation can give some indication of the relative prediction speed in real use. Receiving Operating Characteristic curve shows the relation of true-positive rate to false positive rate, i.e. the recall.

When using the stratified K-fold for validation, all of the models except Bernoulli naïve Bayes had F1-score higher than 0.80, which seems promising. The BNB had a notably low recall. The stochastic gradient descent had the best F1-

score. The precision, recall and F1-measure for each classifier can be seen from the table below (TABLE 25).

TABLE 25 Stratified K-fold cross-validation results for Pipeline 1

Classifier	Fit Time	Score Time	Precision	Recall	F1
MNB	0.023	**0.022**	0.922	0.822	0.868
SGD	0.022	0.023	0.960	0.801	**0.871**
BNB	0.024	0.024	**0.991**	0.580	0.729
KNN	**0.021**	0.055	0.837	0.843	0.838
RF	0.345	0.086	0.937	0.712	0.808
DT	0.053	0.022	0.883	0.771	0.822
LR	0.025	0.022	0.888	**0.856**	0.870

However, Area Under the Curve in Receiving Operating Characteristics of the Bernoulli naïve Bayes is rather high as can be seen from the following figure (FIGURE 23), and it does have the highest precision 0.991. This might indicate that configurations of the Bernoulli naïve Bayes might not be optimal in this evaluation.

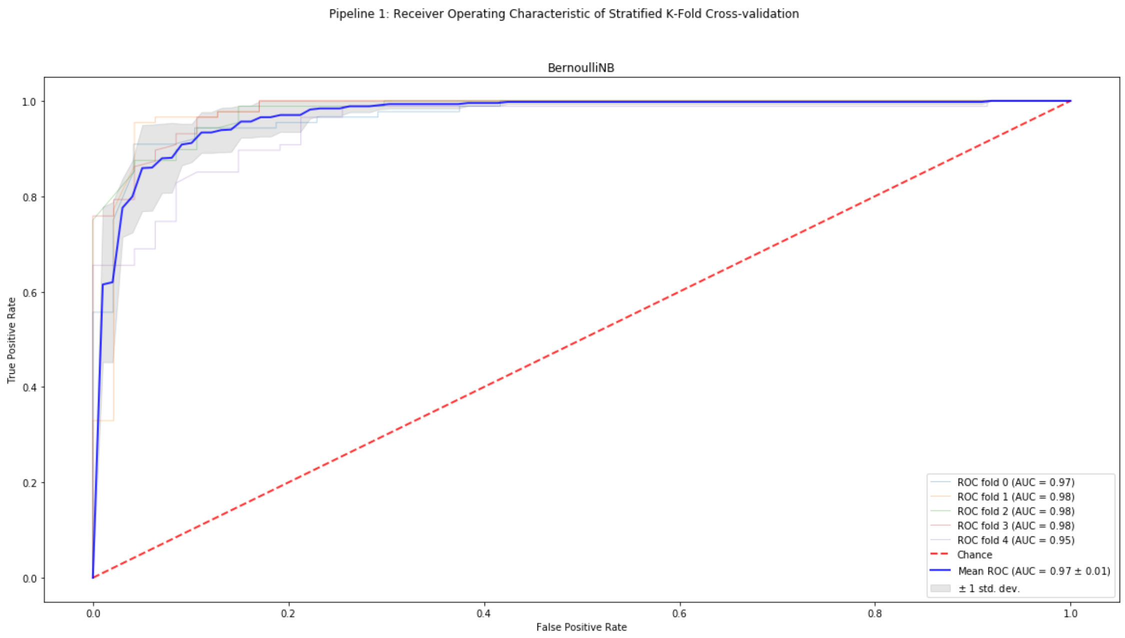


FIGURE 23 ROC of stratified K-fold cross-validation of Bernoulli naïve Bayes classifier

In the group K-Fold cross-validation, the K-nearest neighbours fared best in terms of the F1-measure. Again Bernoulli naïve Bayes had the weakest performance. The group K-fold results for classifiers are listed in the following table (TABLE 26).

TABLE 26 Group K-fold cross-validation results for Pipeline 1

Classifier	Fit Time	Score Time	Precision	Recall	F1
MNB	0.023	0.021	**0.675**	0.472	0.490
SGD	**0.021**	**0.021**	0.657	0.439	0.437
BNB	0.024	0.023	0.600	0.153	0.241
KNN	0.023	0.058	0.642	**0.617**	**0.556**
RF	0.335	0.085	0.634	0.323	0.408
DT	0.046	0.022	0.631	0.410	0.455
LR	0.028	0.022	0.650	0.527	0.514

Random forest and decision trees performed in one fold with AUC = 0.55, which is almost as good as a random guess. Decision tree had the lowest average AUC (0.66) as it is presented in the next figure (FIGURE 24).

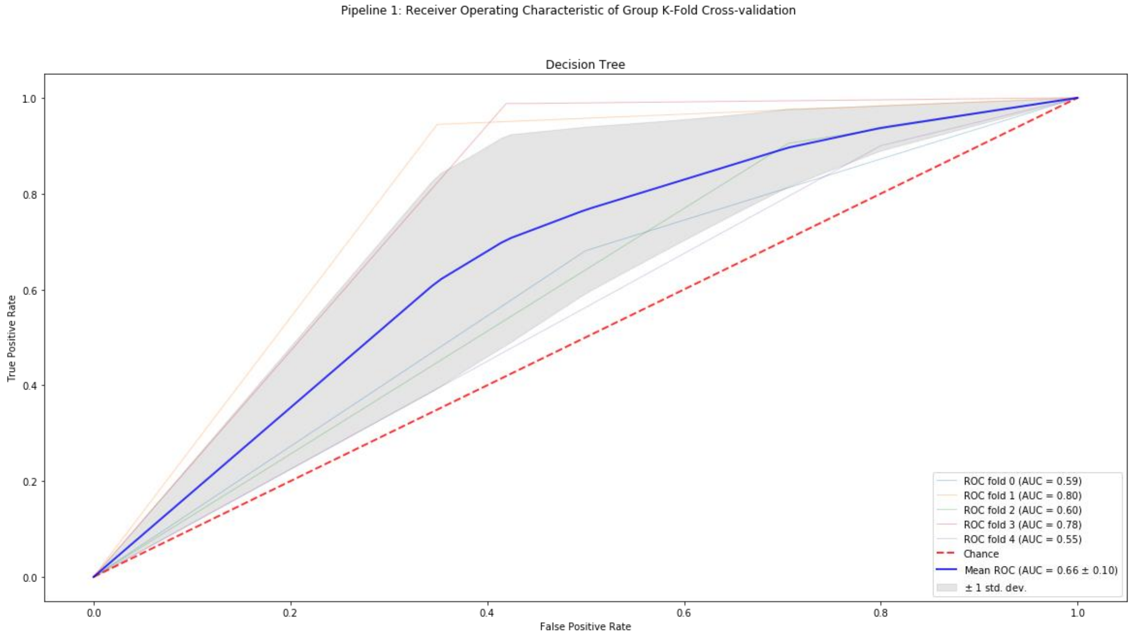


FIGURE 24 ROC of group K-fold cross-validation of the decision tree classifier

Overall the F1-measures were below 0.6 in the group K-fold cross-validation. These results would indicate poor results when presenting new documents with previously unknown words to the classifiers.

The stratified K-fold cross-validation's area under the curve is above 0.9 in all classifiers except for decision trees as can be seen from the following table (TABLE 27). The AUC of group K-fold cross-validations is lower than the stratified, and the standard deviation is notably higher. This could indicate that some of the documents provide better training data for predictions than others or some document are just too different from each other.

TABLE 27 Area under curve (AUC) results of the stratified and group K-fold cross-validations in Pipeline 1

Classifier	Stratified AUC	Standard Deviation	Group AUC	Standard Deviation
MNB	0.97	0.01	0.86	0.09
SGD	0.97	0.01	0.85	0.09
BNB	0.97	0.01	0.83	0.11
KNN	0.93	0.02	0.84	0.07
RF	0.95	0.02	0.77	0.17
DT	0.83	0.03	0.66	0.1
LR	0.97	0.01	0.84	0.11

In short, the results of the first pipeline's stratified cross-validation were rather promising. On the other hand, group cross-validation results were poor, which would indicate poor performance when the classifier is labelling headings with previously unknown words. There is still room for improvement.

5.4.3 Experimental Pipeline 2

Pipeline 2 uses part-of-speech tagging. The purpose of this solution is to address the classifiers inability to use words that are unknown in the vocabulary. The sparsity is alleviated by adding the grammatical properties of the headings to the feature matrix. Feature extraction consists of two pipelines — the regular feature extraction that was presented in the Pipeline 1 and the part-of-speech Feature Extraction. Part-of-speech Feature Extraction implements Spacy's part-of-speech tagging and produces a string that is a sequence of part-of-speech tags. For example, the phrase "apply for certificate of origin" is transformed to 'VB IN NN IN NN'. The following table (TABLE 28) shows the corresponding tags to tokens and their parts of speech according to Santorini's guidelines (1990).

TABLE 28 Part-of-speech tags of the phrase, "apply for certificate of origin"

Token	Tag	Part-of-Speech(Santorini, 1990)
apply	VB	Verb, base form
for	IN	Preposition or subordinating conjunction
certificate	NN	Noun, singular or mass
of	IN	Preposition or subordinating conjunction
origin	NN	Noun, singular or mass

After the part-of-speech transformation, the feature extraction method in the scikit-learn vectorizes the part-of-speech strings and calculates tf-idf. The union of the matrices from the regular feature extraction are provided for the classifier as features. The simplified visualization of the pipeline is presented in the next figure (FIGURE 25).

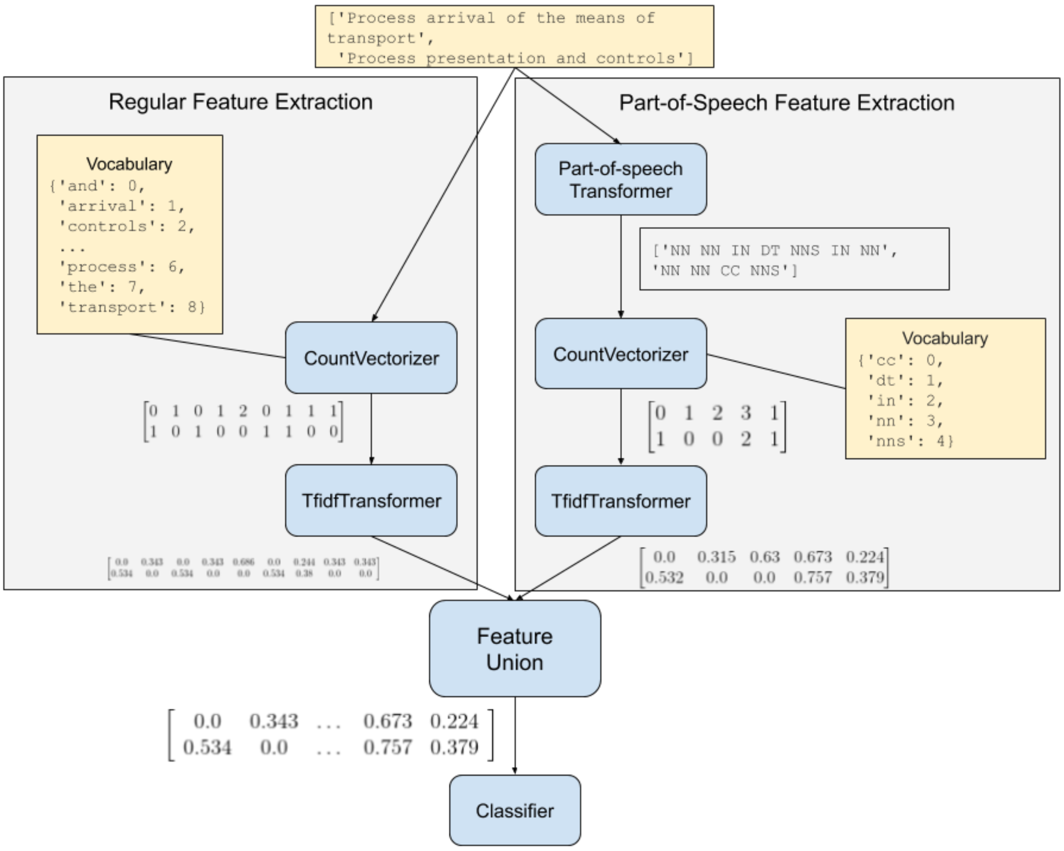


FIGURE 25 Components of the second pipeline

Possible caveats in this solution are the automatic Part-of-speech tagging which may produce erroneous tagging and the lack of sequential information in this solution as well. However, the vectorizer was configured to take also the trigrams into account along with the unigrams and bigrams. More than one n-grams should alleviate the lack of sequential information to some extent, and since the vocabulary with part-of-speech tags is limited, bigrams and unigrams may have more weight in the classification than unigrams. This pipeline is expected to perform better than the Pipeline 1, and the hypothesis (H1) is: adding part-of-speech tagging (Pipeline 2) increases the overall results of the classifiers in terms of F1-score and the AUC compared to Pipeline 1.

5.4.4 Results of Experimental Pipeline 2

As can be seen from the following table (TABLE 29), majority of the classifiers have mean F1-score between 0.8 and 0.9 in the stratified K-fold cross-validation.

Bernoulli naïve Bayes had once again the most inferior performance in terms of F1-Score.

TABLE 29 Stratified K-fold cross-validation results of the second pipeline

Classifier	Fit Time	Score Time	Precision	Recall	F1
MNB	6.984	5.260	0.920	**0.860**	**0.888**
SGD	7.025	5.204	0.925	0.818	0.867
BNB	6.929	5.266	**0.940**	0.593	0.723
KNN	6.961	5.232	0.834	0.746	0.785
RF	7.246	5.207	0.906	0.784	0.839
DT	**6.923**	**5.139**	0.859	0.801	0.829
LR	6.950	5.199	0.897	0.844	0.869

In the group K-fold cross-validation, Bernoulli naïve Bayes performance has increased, but it still performs poorly. All the other classifiers have an F1-score above 0.5. The results are still, however, below 0.6. Decision trees have the best F1-score (~ 0.577), and the stochastic gradient descent has the second-best performance (~ 0.570). The results for the precision, recall and F1-measure can be seen in the following table (TABLE 30).

TABLE 30 Group K-fold cross-validation results of the second pipeline

Classifier	Fit Time	Score Time	Precision	Recall	F1
MNB	6.935	5.181	0.670	0.613	0.542
SGD	6.914	5.196	0.678	0.631	0.570
BNB	**6.882**	**5.143**	**0.680**	0.310	0.414
KNN	6.910	5.227	0.664	0.619	0.547
RF	7.276	5.262	0.653	0.502	0.501
DT	6.976	5.202	0.660	**0.689**	**0.577**
LR	6.943	5.210	0.666	0.621	0.562

The stratified cross-validation's area under the curve results are still mostly above 0.9 in the evaluation of Pipeline 2. Only decision tree has AUC below 0.9 (AUC = 0.87). Also, in the group cross-validation, the decision tree has the only area under curve value below 0.8 (AUC = 0.77). The group K-fold cross-validation has still overall higher standard deviation than the stratified K-fold cross-validation. The AUC results can be seen from the following table (TABLE 31).

TABLE 31 Area under curve (AUC) results of the stratified and group K-fold cross-validations in Pipeline 2

Classifier	Stratified AUC	Standard Deviation	Group AUC	Standard Deviation
MNB	0.97	0.01	0.88	0.06
SGD	0.91	0.01	0.88	0.07
BNB	0.95	0.01	0.82	0.14
KNN	0.91	0.01	0.8	0.09
RF	0.95	0.02	0.84	0.11
DT	0.87	0.01	0.77	0.1
LR	0.96	0.02	0.88	0.08

Overall, the results of the second pipeline are mixed. The use of part-of-speech tagging seems to improve the group K-fold cross-validation results slightly. This slight increase could indicate that the part-of-speech tagging, along with regular vectorization improves the recognition of the process headings that contain previously unknown words for the trained model.

5.4.5 Experimental Pipeline 3

The use of the verb list from Process Lifecycle Verb Taxonomy (von Rosing, Foldager, Hove, von Scheel, & Bøgebjerg, 2015) was implemented in the third pipeline to try a method more in line with Business Process context. In essence, the solution is testing if the headings contain verbs that are included in the verb taxonomy. The assumption is that the business process names contain more likely the verbs that are similar to verb taxonomy.

Rather than making a Boolean comparison between words in the verb list and words in the heading, the WordNet similarities are calculated. Reason for the similarity calculation is to get a more flexible evaluation of the words. For example, the verb list contains the word "customize" which would return false in Boolean evaluation with the word "customise" even though they have the same meaning but different spelling.

Since the wordnet usually gives a list of synsets for a word, this solution selects the one that is a verb and has the verb in the synset name. If there are multiple synsets still after filtering the function returns the first one. If there is no synset of which name consists of the verb, the first synset is returned. The example synset definitions are presented in the following table (TABLE 32). The bolded row in the table presents the selected synset.

TABLE 32 Definitions of the "manage" word's WordNet synsets

Definition	Synset
be successful; achieve a goal	pull_off.v.03
be in charge of, act on, or dispose of	**manage.v.02**
come to terms with	cope.v.01
watch and direct	oversee.v.01
achieve something by means of trickery or devious methods	wangle.v.01
carry on or function	do.v.11
handle effectively	wield.v.02

Most of the documents name the processes with verb-noun phrases. European Central Bank's T2S Business Process Description document, however, uses almost exclusively noun-phrases without verbs (e.g. "maintenance of roles") in process names.

The use of nouns instead of verbs can create false negatives. For example, the transformer would ignore the nouns that have verb root, like 'management'. Thus, this solution also attempts to nominalize verbs to address these false negatives. So the noun tokens from the string are compared against the nominalized version of the verb list.

This solution implements the Natural Language Toolkit's (NLTK) WordNet implementation. WordNet was used to find the closest synsets from the verbs and the appropriate nominalization of the verb automatically when initializing the transformer. Spacy was used to select only the verbs and nouns for similarity calculations from the text array input.

The Natural Language Toolkit calculated the similarities by using the Wu-Palmer similarity (Wu & Palmer, 1994). Wu-Palmer similarity score gives a similarity between 0 and 1. The similarity scores were reduced to the best similarity of each phrase. A quick comparison of the similarity distribution shows that there are some differences with non-process headings and process headings (FIGURE 26).

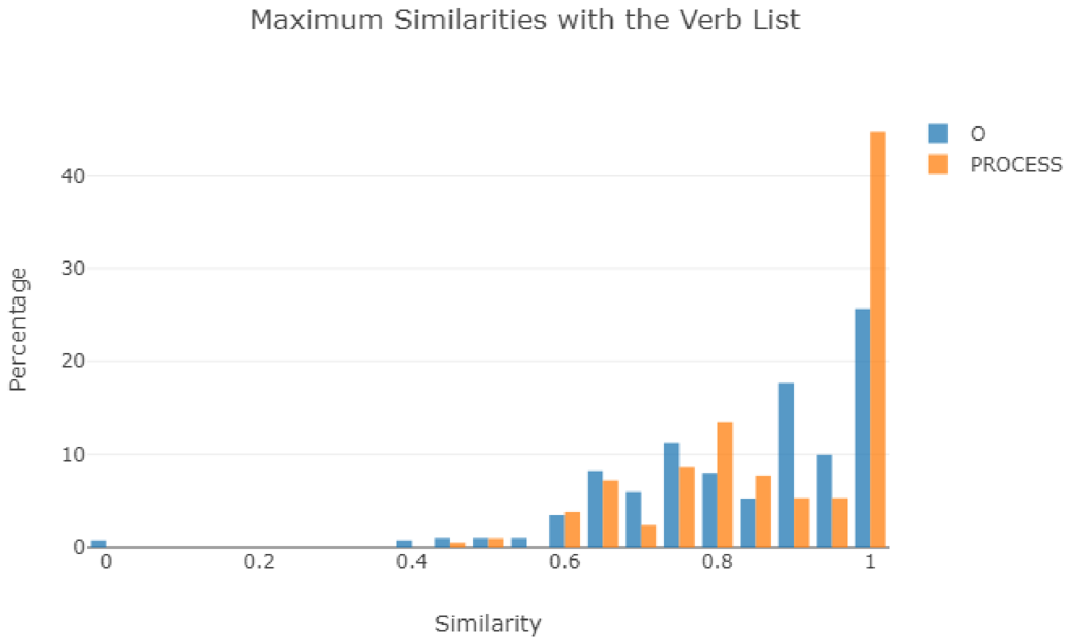
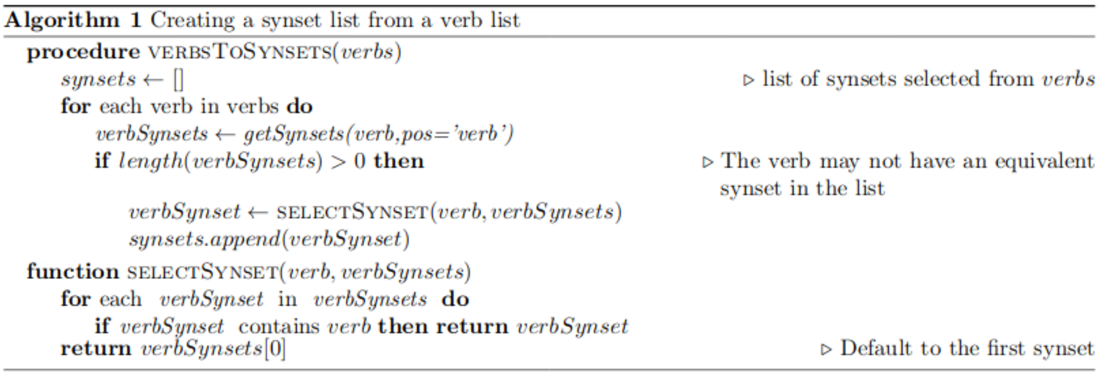


FIGURE 26 Maximum similarities of the headings when compared to the verb list

The final version of this feature extraction solution returned a matrix that included four values in each row, a maximum similarity of the verbs from the phrase, the position of the most similar verb, a maximum similarity of the nouns and the position of the most similar noun. The presumption here is that the most likely a process label would include a verb that is at the beginning of the phrase and is in the verb list. This algorithm is presented in the next figure (FIGURE 27).



```
Algorithm 1 Creating a synset list from a verb list
  procedure VERBSTOSYNSETS($verbs$)
      $synsets \leftarrow []$                                  ▷ list of synsets selected from $verbs$
      for each verb in verbs do
          $verbSynsets \leftarrow getSynsets(verb,pos='verb')$
          if $length(verbSynsets) > 0$ then                ▷ The verb may not have an equivalent
                                                             synset in the list
              $verbSynset \leftarrow$ SELECTSYNSET($verb, verbSynsets$)
              $synsets.append(verbSynset)$
  function SELECTSYNSET($verb, verbSynsets$)
      for each $verbSynset$ in $verbSynsets$ do
          if $verbSynset$ contains $verb$ then return $verbSynset$
      return $verbSynsets[0]$                                ▷ Default to the first synset
```

FIGURE 27 Algorithm for creating a synset list from the verb list

The matrix produced by the transformer is then included in the feature union with the features provided by the previously presented pipelines. The next figure (FIGURE 28) shows this pipeline.

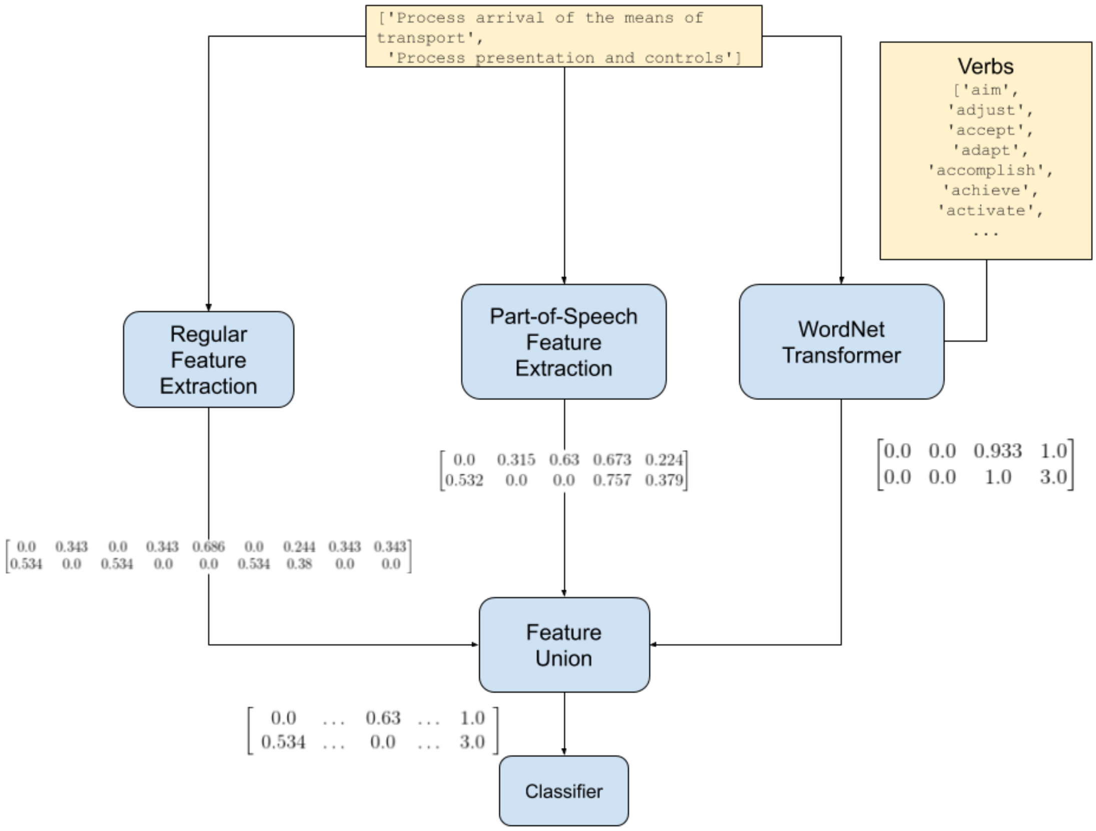


FIGURE 28 Components of the third pipeline

The hypothesis (H2) for this pipeline is: implementing similarity measures of Business Process Taxonomy verbs (Pipeline 3) increases the overall results of the classifiers in terms of F1-Score and the AUC compared to Pipeline 2.

5.4.6 Results of Experimental Pipeline 3

The best performing classifier in the stratified K-fold cross-validation was logistic regression (~0.872). K-nearest neighbours had the poorest performance (~0.763) which was a slightly lower result than the Bernoulli naïve Bayes classifier that had the poorest performance in the previous pipelines. The results of the stratified cross-validation are presented in the following table (TABLE 33).

TABLE 33 Stratified K-fold cross-validation results of the third pipeline

Classifier	Fit Time	Score Time	Precision	Recall	F1
MNB	46.330	34.490	0.912	0.788	0.842
SGD	**44.722**	35.360	0.883	0.767	0.809
BNB	45.541	**34.033**	**0.958**	0.653	0.776

(to be continued)

TABLE 33 (to be continued)

Classifier	Fit Time	Score Time	Precision	Recall	F1
KNN	59.386	47.975	0.786	0.750	0.763
RF	82.202	61.586	0.914	0.780	0.840
DT	58.723	40.290	0.856	0.780	0.815
LR	59.870	47.543	0.884	**0.860**	**0.872**

All the classifiers in the pipeline still have an F1-score below 0.6 in the group K-fold cross-validation, however the best performing classifier, logistic regression has F1-score ~0.599. Although the KNN had the worst performance in the stratified K-fold cross-validation, it did perform relatively well in the group K-fold cross-validation (~0.557). The Bernoulli naïve Bayes classifier had the poorest performance in group K-fold cross-validation again. The following table presents the results of the group K-fold cross-validation (TABLE 34).

TABLE 34 Group K-fold cross-validation results of the third pipeline

Classifier	Fit Time	Score Time	Precision	Recall	F1
MNB	59.112	46.248	**0.697**	0.619	0.584
SGD	43.906	32.804	0.647	0.605	0.510
BNB	44.450	**32.694**	0.696	0.342	0.442
KNN	**43.701**	33.198	0.610	0.624	0.557
RF	44.148	33.240	0.675	0.566	0.548
DT	44.024	32.888	0.650	0.544	0.532
LR	44.884	33.823	0.678	**0.654**	**0.599**

Random forest had the fourth-best group K-fold result in terms of F1-measure, although it had the second-lowest AUC (0.77) and the highest AUC standard deviation (0.17). As can be seen from the following graph (FIGURE 29), one of the folds has 0.50 AUC. This indicates the inconsistency of the random forest classifier in this pipeline.

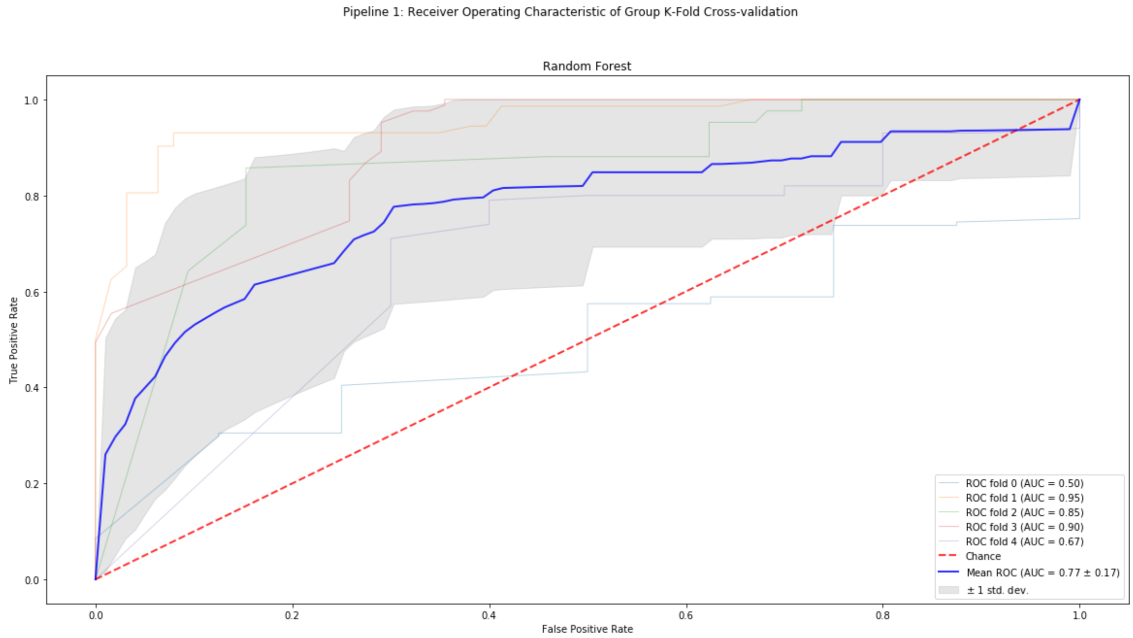


FIGURE 29 ROC of group K-fold cross-validation of random forest classifier

Bernoulli naïve Bayes and logistic regression had the largest area under the curve. The AUC results can be seen from the following table (TABLE 35).

TABLE 35 Area under curve (AUC) results of the stratified and group K-fold cross-validations in Pipeline 3

Classifier	Stratified AUC	Standard Deviation	Group AUC	Standard Deviation
MNB	0.95	0.02	0.87	0.07
SGD	0.92	0.02	0.86	0.06
BNB	0.96	0.02	0.85	0.11
KNN	0.87	0.02	0.78	0.08
RF	0.95	0.03	0.85	0.09
DT	0.88	0.04	0.75	0.1
LR	0.96	0.02	0.89	0.08

After implementing the Wordnet similarity to the pipeline, the average group K-fold cross-validation's F1-scores have slightly increased, and the AUC and stratified results have slightly decreased. The hypothesis of the use of verb taxonomy might itself not be invalid. There could be problems in the implementation. Since synsets are created automatically from the verb strings and corresponding nouns are created automatically from the verb's synset, there might be errors in this automation.

Closer look of the automatically selected synsets does indeed show that some synsets do not refer to correct definition. For example, the first definition of 'execute' verb is "kill as a means of socially sanctioned punishment" which is hopefully undesirable in the business context.

Since the synsets are created from the verb strings and nouns are derived automatically from the verb strings, there may arise problems. Also, the Wu-Palmer similarity might not be the correct solution here. When looking at the Wu-Palmer similarities between the synset of “perform” (perform.v.01) and “execute” verb's synsets, the most morbid definitions still gives the highest similarity besides the actual synset of the verb “perform” itself. Next table (TABLE 36) lists all the synsets retrieved with the word “execute” and their Wu-Palmer similarity with the “perform” synset (perform.v.01).

TABLE 36 Wordnet synsets for the word "execute" and their Wu-Palmer similarity with the "perform" synset

Definition	Similarity	Synset
kill as a means of socially sanctioned punishment	0.4	execute.v.01
murder in a planned fashion	0.3333333333	execute.v.02
put in effect	0.2857142857	carry_through.v.01
carry out the legalities of	0.2857142857	execute.v.04
carry out a process or program, as on a computer or a machine	0.2857142857	run.v.19
carry out or perform an action	**1**	**perform.v.01**
sign in the presence of witnesses	0.2	execute.v.07

Overall, the results of the third pipeline are mixed. Although the group K-fold cross-validation results are notably better than in Pipeline 1, it should be noted that the training and prediction have become now much more time-consuming. In the stratified K-fold cross-validation the results are lower than in the first pipeline. When comparing the results to Pipeline 2, the slight improvement might not be worth the loss of efficiency.

5.5 Comparison of the Experimental Pipelines

This section compares the F1-measures of the three experimental pipelines. Between the classifiers and average results. All the Pipeline comparisons were measured using precision, recall, F1-Score and the AUC. Following graph presents the performance of each classifier and the mean F1-score of the stratified K-fold cross-validation (FIGURE 30).

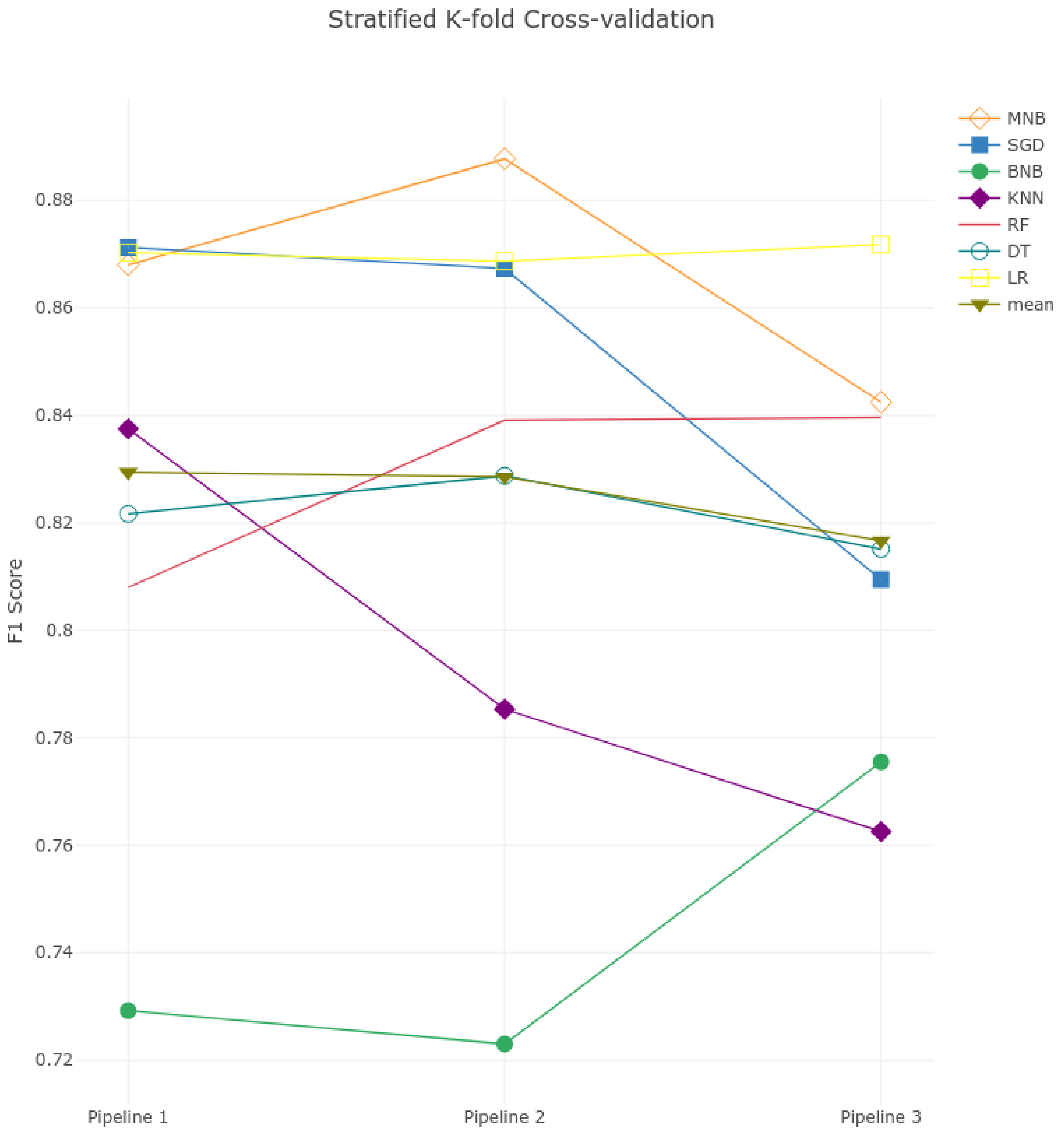


FIGURE 30 Stratified K-fold cross-validation's F1-score for the first three pipelines

Most prominent observation from the graph is the drastic decline of the K-Nearest Neighbours classifier. The K-Nearest Neighbours' score is decreasing from ~0.84 to ~0.76. Only results of random forest increase along the pipelines. The results of the stratified cross-validation show that the mean of the results is slightly lower in Pipeline 2 and 3 than in Pipeline 1. Without the KNN, the second pipeline's F1-score would have been higher than the first pipeline's.

As can be seen from the following figure (FIGURE 31), the group K-fold validation's results are better in the Pipeline 2 than in Pipeline 1 and slightly higher in the Pipeline 3 than in the Pipeline 2. In the group cross-validation, the KNN's results remain rather stable across the pipelines. The logistic regression, random forest and multinomial naïve Bayes improve their F1-score between the pipelines.

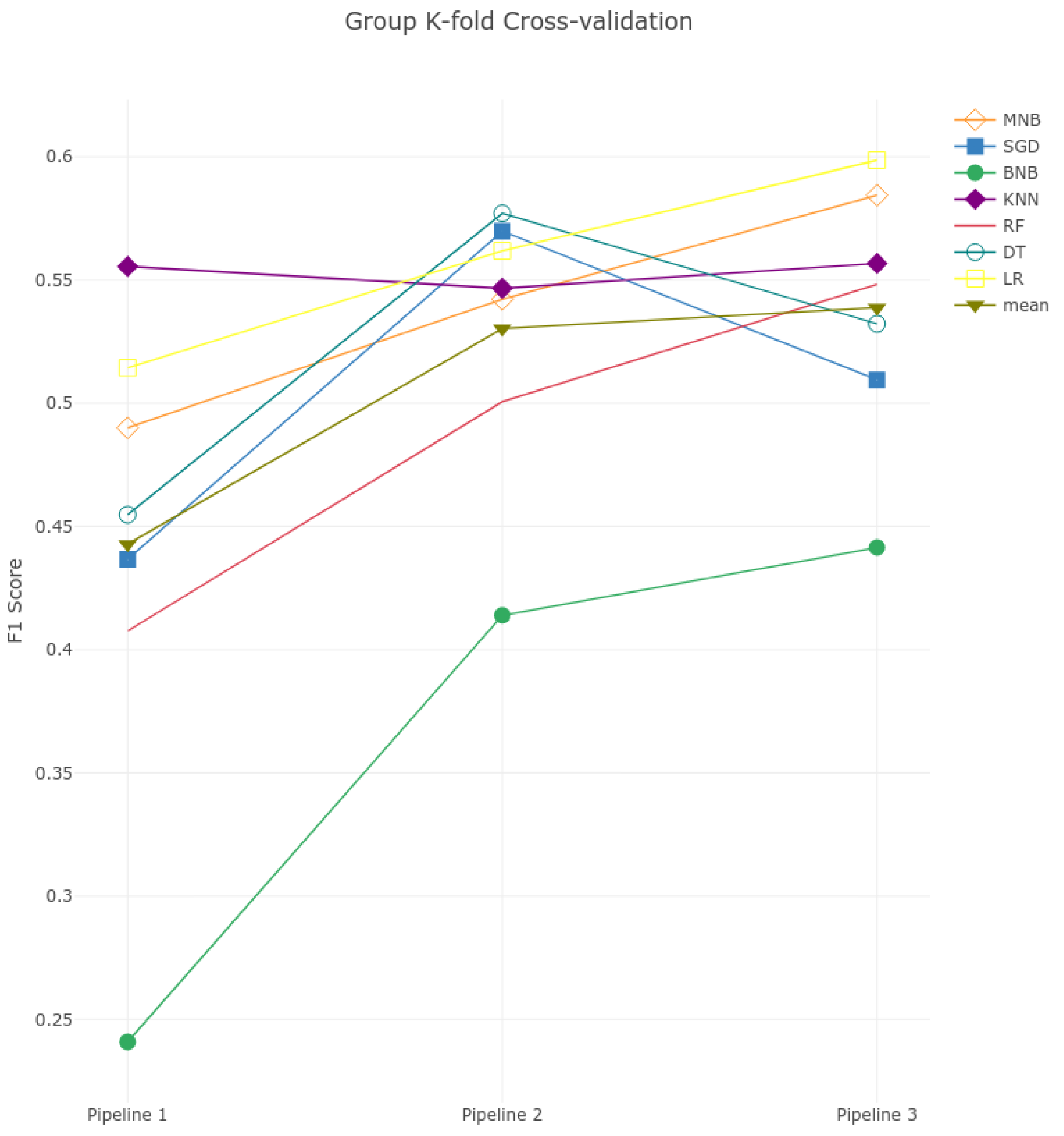


FIGURE 31 Group K-fold cross-validation's F1-score for the first three pipelines

The results of the area under the curve are similarly descending as the results of F-measures in the stratified K-fold cross-validation. The stratified area under curve decreases between pipelines and is highest at Pipeline 1. The group K-fold cross-validation shows some improvement in the Pipeline 2. Pipeline 2 and Pipeline 3 did have a higher mean AUC than Pipeline 1. The mean area under the curve results can be seen in the following table (TABLE 37).

TABLE 37 Mean AUC of the Pipelines

K-fold Cross-validation	Pipeline 1	Pipeline 2	Pipeline 3
Stratified	0.941	0.931	0.927
Group	0.807	0.839	0.836

As can be seen from these comparisons and the previous result sections, the results are inconclusive. There is some improvement in the group K-fold cross-validation, but the stratified K-fold cross-validation did decrease. Some of this decrease is because of the notable decline of the KNN classifier's results.

However, the hypotheses for these evaluations were:

> H1: Adding part-of-speech tagging (Pipeline 2) increases the overall results of the classifiers in terms of F1-Score and the AUC compared to Pipeline 1.
>
> H2: Implementing similarity measures of Business Process Taxonomy verbs (Pipeline 3) increases the overall results of the classifiers in terms of F1-score and the AUC compared to Pipeline 2.

The next table (TABLE 38) shows each validation and if these hypotheses were confirmed or not.

TABLE 38 Outcomes of the hypotheses

K-fold cross-validation	H1	H2
Group F1 Mean	Yes	Yes
Group AUC Mean	Yes	No
Stratified F1 Mean	No	No
Stratified AUC Mean	No	No

Although there were some improvements, the Pipeline 2 and 3 did not confirm the hypotheses. If the aim is, however, to classify headings that have different words than in the training data, the Pipeline 2 seems like a better solution than Pipeline 1. Pipeline 3 did perform slightly better in group K-fold cross-validation, but with much larger fit time. From the classifiers, the logistic regression or multinomial naïve Bayes might be the best choices since they made consistently reliable predictions compared to other classifiers.

6 RESULTING PIPELINE 4

The last pipeline used sentences where the heading appears in the document. This solution supposes that classification requires more context to produce reliable results. For example, the payment process, that has been used as an example of an ambiguous process name in this thesis, can be challenging to recognize as a process if the only input is the string ″payment″. When more context is added, for example, if the input contains a sentence ″Payment process begins from payment request event″ it is easier to classify ″payment″ as a process. Of course, this pipeline's solution could have some caveats. For example, if the process name as a word is so ubiquitous that the input is filled with multiple irrelevant sentences that do not indicate if the heading is a process or not. Also, in the genre analysis, it was observed that process name might not appear in the text paragraph as in the headings. For example, ″Take Decision″ process becomes the decision-taking process and thus, the ″Take Decision″ may not appear in the paragraphs at all.

Still, the prototype of this solution can be more easily created than the named entity recognition, which might need annotation of large corpora before it becomes viable. Also, the general labels that have multiple different meanings in the document may confuse the named entity recognizer as well as the classification solution.

6.1 Structure of the Pipeline 4

The fourth pipeline retrieves the headings from the text and takes the sentences that contain the possible process names as a context. The sentences in this solution are retrieved from PDF documents with regular expression. Sentences in this context were phrases that end with full stops, exclamation or question marks. Some queries of the phrases returned an empty string. Thus the heading was included at the beginning of the context observation to mitigate the lack of observations. The simplified pipeline is presented in the following figure (FIGURE 32).

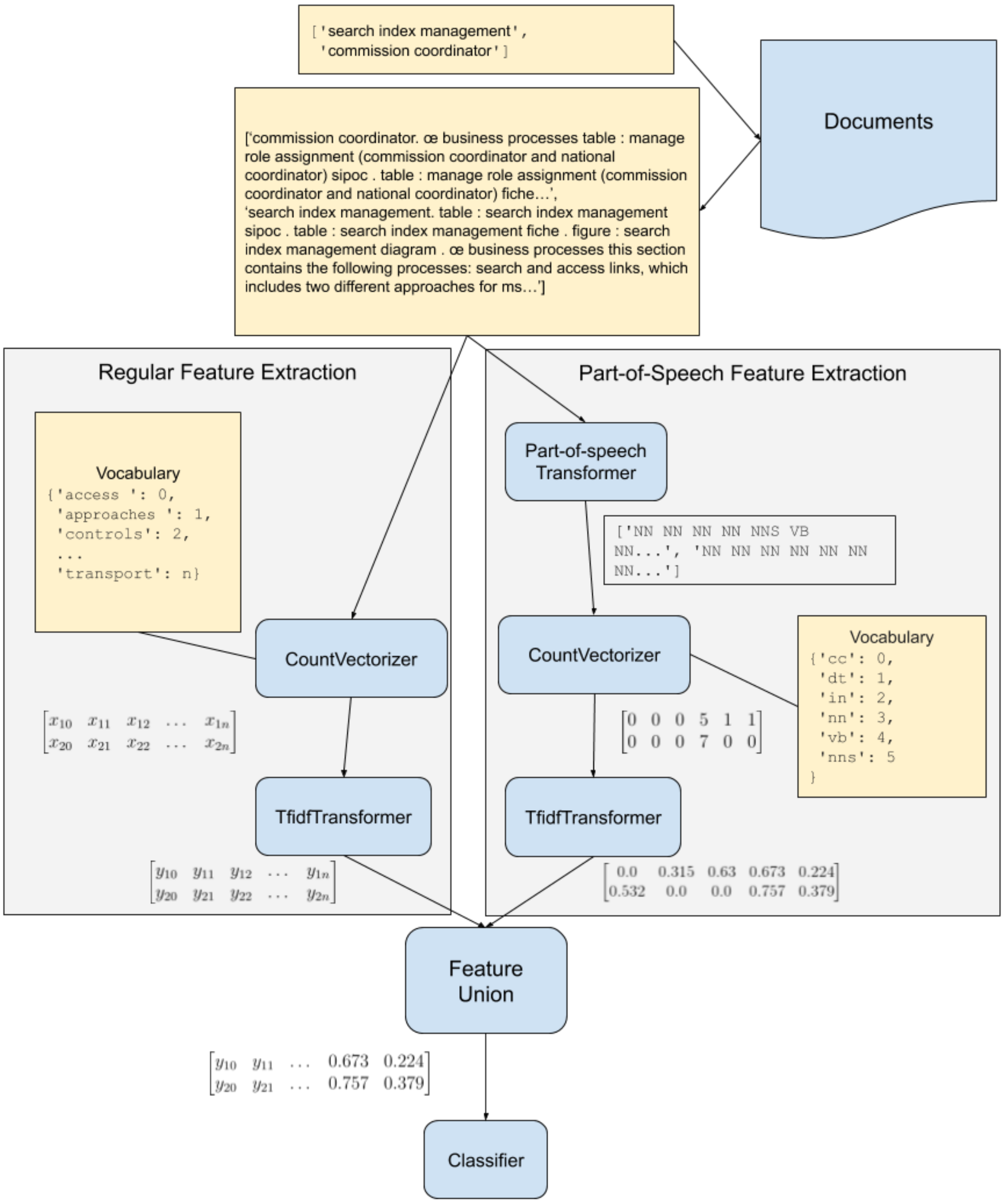

['search index management' ,
'commission coordinator']
Documents
['commission coordinator. œ business processes table : manage role assignment (commission coordinator and national coordinator) sipoc . table : manage role assignment (commission coordinator and national coordinator) fiche…',
'search index management. table : search index management sipoc . table : search index management fiche . figure : search index management diagram . œ business processes this section contains the following processes: search and access links, which includes two different approaches for ms…']
Regular Feature Extraction
Vocabulary
{'access ': 0,
'approaches ': 1,
'controls': 2,
...
'transport': n}
CountVectorizer
$\begin{bmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1n} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2n} \end{bmatrix}$
TfidfTransformer
$\begin{bmatrix} y_{10} & y_{11} & y_{12} & \cdots & y_{1n} \\ y_{20} & y_{21} & y_{22} & \cdots & y_{2n} \end{bmatrix}$
Part-of-Speech Feature Extraction
Part-of-speech Transformer
['NN NN NN NN NNS VB NN...', 'NN NN NN NN NN NN NN...']
CountVectorizer
Vocabulary
{'cc': 0,
'dt': 1,
'in': 2,
'nn': 3,
'vb': 4,
'nns': 5
}
$\begin{bmatrix} 0 & 0 & 0 & 5 & 1 & 1 \\ 0 & 0 & 0 & 7 & 0 & 0 \end{bmatrix}$
TfidfTransformer
$\begin{bmatrix} 0.0 & 0.315 & 0.63 & 0.673 & 0.224 \\ 0.532 & 0.0 & 0.0 & 0.757 & 0.379 \end{bmatrix}$
Feature Union
$\begin{bmatrix} y_{10} & y_{11} & \cdots & 0.673 & 0.224 \\ y_{20} & y_{21} & \cdots & 0.757 & 0.379 \end{bmatrix}$
Classifier

FIGURE 32 Components of the final pipeline

This pipeline was chosen to use the structure of the second pipelines over the structure of the pipeline three since the third pipeline's performance was inconclusive when compared to Pipeline 2. The average F1-measure in group cross-validation was 0.009 lower in the second pipeline than in the third pipeline, and the stratified cross validation's F-measure was 0.012 higher in the pipeline two than in the pipeline three. Also, the presumption was that the verb list used in Pipeline 3 was not as relevant when using longer phrases as parameters. All in all, a six-fold increase of the Pipeline 3's score time made the second pipeline a more justified starting point for the last pipeline.

This solution has the following hypothesis (H3): when the possible process names are presented with more context, the classifier will perform better than the other pipelines in terms of F1-score and AUC in both stratified cross-validation and group cross-validation.

6.2 Results of the Pipeline 4 Evaluation

Random forest performed the best in stratified K-fold cross-validation in terms of F1-score (0.883) and the AUC (0.97). Bernoulli naïve Bayes had the most inferior F1-measure (0.635). In the AUC of the stratified K-fold cross-validation, Decision trees had the worst performance (0.88). Results of the stratified K-fold cross-validation can be seen in the following table (TABLE 39).

TABLE 39 Stratified K-fold cross-validation results of the final pipeline

Classifier	Fit Time	Score Time	Precision	Recall	F1
MNB	25.082	6.299	0.957	0.699	0.806
SGD	25.263	6.402	0.920	0.839	0.876
BNB	25.115	6.273	0.481	**0.936**	0.635
KNN	25.142	6.320	0.893	0.800	0.843
RF	25.787	6.363	**0.960**	0.822	**0.883**
DT	25.371	**6.244**	0.880	0.868	0.873
LR	**24.975**	6.327	0.906	0.826	0.862
Mean	25.248	6.318	0.857	0.827	0.825

In group cross-validation, the logistic regression had the best F1-score. The best performing classifier in the stratified cross-validation, the random forest was the penultimate classifier as can be seen from following table (TABLE 40). Bernoulli naïve Bayes had yet again the lowest F1-score. Overall the group K-fold F1-score is lower in the Pipeline 4 (~0.431) than in the Pipeline 1 (~0.443).

TABLE 40 Group K-fold cross-validation of the final pipeline

Classifier	Fit Time	Score Time	Precision	Recall	F1
MNB	25.692	6.353	0.817	0.281	0.412
SGD	25.454	6.396	0.647	0.393	0.461
BNB	25.327	6.308	0.380	0.517	0.359
KNN	**24.871**	6.321	0.612	0.411	0.464
RF	25.595	6.287	**0.805**	0.299	0.391
DT	25.179	6.266	0.571	0.517	0.451
LR	24.962	**6.262**	0.640	0.464	**0.480**
Mean	25.297	6.313	0.639	0.412	0.431

Results of the area under the curve have the largest standard deviations of all the pipelines. The following table presents the area under the curve results of the final pipeline (TABLE 41).

TABLE 41 Area under curve (AUC) results of the stratified and group K-fold cross-validations in Pipeline 4

Classifier	Stratified AUC	Standard Deviation	Group AUC	Standard Deviation
MNB	0.95	0.02	0.75	0.20
SGD	0.96	0.01	0.79	0.15
BNB	0.91	0.01	0.69	0.20
KNN	0.94	0.02	0.70	0.17
RF	0.97	0.01	0.81	0.13
DT	0.88	0.03	0.62	0.11
LR	0.96	0.01	0.80	0.17

Every group K-fold cross-validation had one fold that performed poorly compared to the other folds. As can be seen from the ROC curve of the decision trees (FIGURE 33), the mean AUC is near 0.50.

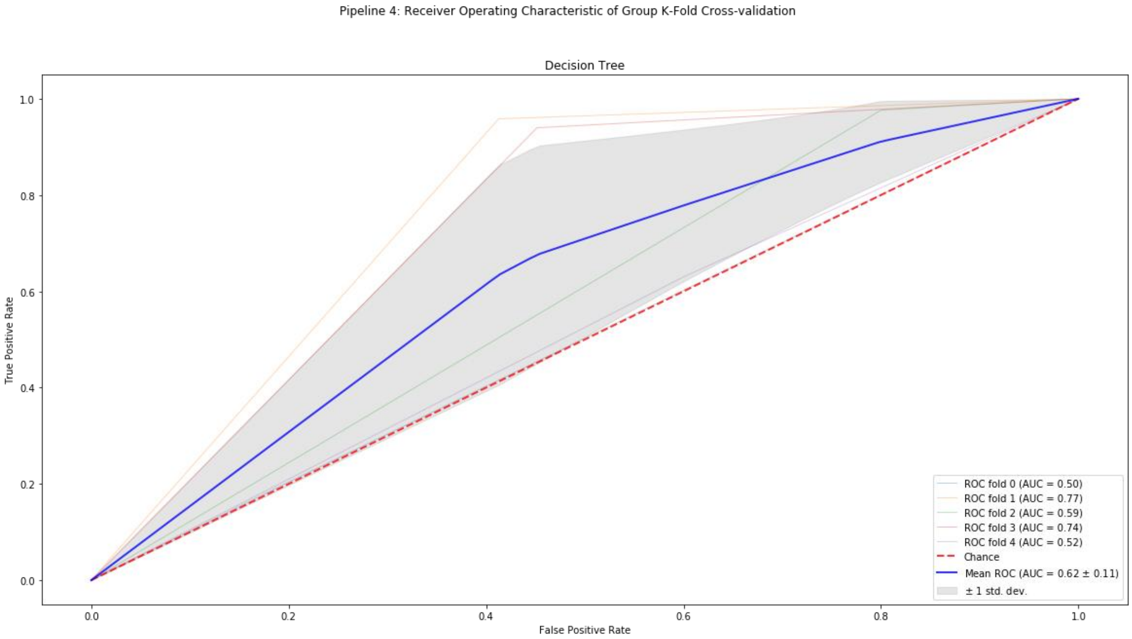


FIGURE 33 The receiver operating characteristic curve of group K-fold cross-validation of a decision tree classifier

The part-of-speech part of the pipeline might not produce relevant data for the classifier. Since it basically offers only the count of each part-of-speech in the text.

These values can vary more between each process sentence than between process and non-process tags.

Instead of adding useful information to the pipeline, the context seems to add more noise for the training data of classifiers. The weakness in this solution is also the increase in fit time, although it is not as large as in the third pipeline. The increase in score time is slightly longer than in Pipeline 2. Score time is still notably shorter than in Pipeline 3. However, it should be noted that the extraction of the phrases related to process name candidates was also rather time-consuming, although this was not measured in this evaluation. Pipeline 4 did not perform better than other pipelines. This can be seen from the table below that shows the mean F-measures of all the pipelines (TABLE 42).

TABLE 42 Mean F-measures of all the pipelines

Cross-validation	**Pipeline 1**	**Pipeline 2**	**Pipeline 3**	**Pipeline 4**
Mean F1 Group	0.4428	0.5303	0.5388	0.4311
Mean F1 Stratified	0.8294	0.8286	0.8166	0.8255

When the possible process names are presented with more context, the classifier does not perform better than the other pipelines in this case, and the third hypothesis (H3) has not been confirmed.

7 CONCLUSION

In this Design Science Research, the goal was to find a solution for discovering business processes from the unstructured text by natural language processing. The purpose was to find a way to alleviate the workload of the knowledge worker with process definition assignments for e.g. process modeling.

The literature research chapter delved into the current state of the art of natural language processing. This literature review's purpose was to provide knowledge for defining the objectives step in the design science research method. The results provided a set of tools, measures, and methods that are available in natural language processing. Recall, precision, and F-measure were deemed as relevant measures since they were most popular and would enable comparison with other results if needed.

The first solution for discovering processes was developed by using named entity recognition. The named entity recognition solution used the named entity recognition implementation of the Spacy Natural Language Processing library to extract the processes from documents containing Business Process descriptions. The named entity recognition was chosen as the most applicable method for this problem. It is readily available in multiple Natural Language Processing tools and libraries. Spacy's implementation of named entity recognition was selected because of its ease of use that enabled quick prototyping of the process extraction. The preliminary results of the solution were, however, weak. Instead of continuing to develop it, the research took a step back to look at the Business Process documentation through the lens of genre analysis.

The method that was used in this genre analysis was a content analysis that attempted to discover the semiotic and structural means of communicating the process information. From these results, many observations were made. First, the process phrases differ notably between the documents. Inside the document, they are somewhat consistent. Also, although the documents' structure differs, they show similarities through the content analysis. Since most of the process names were in the headings, the analysis would indicate that a text classification could be a viable option in pre-processing the document to recognize the chapter where the process labels are more likely located. Rather than treating the chapters equally, more weight should be put on the low hanging fruits. It could be a waste of time to recognize entities from irrelevant chapters. In the best case, the table of contents may provide all the process labels as subheadings of the process chapter, preferably even labelled as Business Processes, and the process chapter provides all the other salient process components. At worst case, there is not even a proper table of contents, and the process names are distributed evenly in the document.

Because of observations from the genre analysis, the classification of document headings into process names and non-process headings was selected as a solution for the last design phase. Four different pipelines were built to demonstrate the classification. These pipelines contained the pre-processing strategies of the observations to provide useful data for the classifiers. The first pipeline produced promising results in the evaluation. However, the group K-fold cross-validation, that was supposed to evaluate how well does the pipeline work when presenting process names from different documents performed poorly.

The following two pipelines attempted to improve these results. Although the group K-fold cross-validation results improved in the second and third pipeline, the other cross-validation results were worse. These two pipelines did not fully confirm the proposed hypotheses.

The final pipeline strived to enhance the classification performance by taking the context into account. The context, in this case, was the sentences that included the process name. This solution, however, produced more noise for the classifiers than useful data and did not improve the performance.

Even though the preliminary hypotheses were not proved, the pipelines two and three might be more useful in practice. The stratified K-fold cross-validation results were not abysmally low, and the first pipeline might require a large corpus to perform adequately in real-world settings.

7.1 Previous Research

The literature review presented the F-measure that was used in several papers. This enables the examination of how this research aligns with results from other papers. Ferreira et al. attempted to discover the business process elements from short sentences with a semi-automatic approach. They had F-measure between ~0.81 and ~0.93. (2017.)

Leopold et al. attempted to classify textual processes into manual, user, and automatic labels. The classification was done to recognizes candidate tasks for robotic process automation. They performed 10-fold cross-validation, and the F1-scores for their classification were between 0.56 and 0.85. The classification of manual tasks had the highest F1-measure, and the automated label had the lowest F-measure. Their solution also had the area under the curve between 0.75 and 0.78. (2018.)

Jlailaty et al. attempted to discover the Business Process activities from email logs by clustering. They compared the LSI and Word2vec together with clustering. Their F-measures were between 0.42 and 0.56. In this research, the Word2vec fared better than LSI. (2017a.)

These results are not far from the results from the cross-validation of the classification. The research by Jlailaty et al. had quite low F-measures like the results from the group k-fold cross-validation. The stratified k-fold cross-validation results were similar to the result of Leopold et al. The solutions presented in this thesis would require still more improvement to achieve similar results as the solution of Ferreira et al.

7.2 Further Research

This research subject offers multiple possibilities for related research. Observations from this thesis can be extended to discovering other process components, using other methods and tools for the same problem, conducting user experiments on the field or examining Business Process documentation more thoroughly.

There are more challenges in finding other process components, such as activities. As the genre analysis revealed that the activities may not be mentioned at all in the text and only show up in the process diagram, or the activities have ids that are used to refer to them. Also, the components of the business process do not appear in the headings. However, this information can be found from the tables, so a more sophisticated way to parse tables from documents could be useful. Extraction of all the components would also require a much larger corpus.

Although the Spacy and straightforward named entity recognition was discarded at its early stage, it is still a viable option for discovering processes. It does, however, require extensive attention for the annotation. Some commercial solutions aim to alleviate the manual work that the creation of the corpus requires, like Prodigy[10]. Other tools like BERT have recently gained popularity in the Natural Language Processing (Devlin, Chang, Lee & Toutanova, 2019). Tools like BERT might make the named entity recognition from the text a more viable solution. Also, a dependency graph might provide a better result than just an unordered bag-of-words that the pipeline solutions used.

Another way to approach the problem would be to get more insights about the needs and the behaviour of the knowledge worker. This research did not include a real-world experiment with users, which would have been the most decisive evaluation. As it was mentioned, a user experiment in the field with a more developed system would provide valuable insights about the solution. It could also be relevant to know if the knowledge worker is more in need of improved retrieval of process components and relevant process descriptions or get the selected text turned into a business process model.

Although Natural Language Processing methods offer multiple possibilities for discovering the processes, the problem of processing the PDF documents may remain. It would be interesting to know what kind of documentation would serve both human and machine processing needs. Business Process Modelling should have an important role in the early stages of the documentations. The documents used in this research would have been difficult to recognise as Business Process descriptions without process diagrams.

We should pay more attention to the usability of our documentation. The amount of the information is not going to decline, and the work hours for processing of the information are not going to increase. We can only make the time we spend processing the information more efficiently.

[10] https://prodi.gy

REFERENCES

1.1. Generalized Linear Models — scikit-learn 0.21.2 documentation. (n.d.). Retrieved June 23, 2019, from https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

Abiteboul, S. (1997, January). Querying semi-structured data. In International Conference on Database Theory (pp. 1-18). Springer, Berlin, Heidelberg.

Akinyelu, A. A., & Adewumi, A. O. (2014). Classification of phishing email using random forest machine learning technique. Journal of Applied Mathematics, 2014.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. International Journal of Computer Science Issues (IJCSI), 9(5), 272.

Annervaz, K. M., George, J., & Sengupta, S. (2015). A Generic Platform to Automate Legal Knowledge Work Process Using Machine Learning. In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA) (pp. 396-401). IEEE.

Bateman, J. (2008). Multimodality and genre: A foundation for the systematic analysis of multimodal documents. Springer.

Bateman, J., Delin, J., & Henschel, R. (2007). Mapping the multimodal genres of traditional and electronic newspapers. *New directions in the analysis of multimodal discourse*, 147-172.

Bawden, D., & Robinson, L. (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. Journal of information science, 35(2), 180-191.

Berardi, G., Esuli, A., Fagni, T., & Sebastiani, F. (2015). Classifying websites by industry sector: a study in feature design. In Proceedings of the 30th Annual ACM Symposium on Applied Computing (pp. 1053-1059). ACM.

Biber, D., Connor, U., & Upton, T. A. (2007). Discourse on the move : Using corpus analysis to describe discourse structure. Retrieved from https://ebookcentral.proquest.com

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 (pp. 177-186). Physica-Verlag HD.

Connor, U., & Mauranen, A. (1999). Linguistic analysis of grant proposals: European Union research grants. *English for specific purposes, 18*(1), 47-62.

Delin, J., Bateman, J. A., & Allen, P. (2002). A model of genre in document layout. *Information Design Journal, 11*(1), 54-66.

De Medio, C., Gasparetti, F., Limongelli, C., & Sciarrone, F. (2017). Automatic Extraction and Sequencing of Wikipedia Pages for Smart Course Building. In 2017 21st International Conference Information Visualisation (IV) (pp. 378-383). IEEE

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-73.

Dumas, M., La Rosa, M., Mendling, J. & Reijers, H. A. (2018). *Fundamentals of Business Process Management* (2nd ed. 2018.). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization.

Ferreira, R., Thom, L., & Fantinato, M. (2017). A Semi-automatic Approach to Identify Business Process Elements in Natural Language Texts. In ICEIS (3) (pp. 250-261).

Friedrich, F., Mendling, J., & Puhlmann, F. (2011). Process model generation from natural language text. In International Conference on Advanced Information Systems Engineering (pp. 482-496). Springer, Berlin, Heidelberg.

Fortier, S. C., & Dokas, I. M. (2008, May). Setting the specification framework of an early warning system using IDEF0 and information modeling. In Proceedings of the 5th International ISCRAM Conference (pp. 441-450).

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144.

Gao, F., & Bhiri, S. (2014). Capability annotation of actions based on their textual descriptions. In 2014 IEEE 23rd International WETICE Conference (pp. 257-262). IEEE.

Harmon, P. (2010). The scope and evolution of business process management. In Handbook on business process management 1 (pp. 37-81). Springer, Berlin, Heidelberg.

Hammer, M. (2015). What is business process management?. In *Handbook on business process management 1* (pp. 3-16). Springer, Berlin, Heidelberg.

Heikkinen, V., & Voutilainen, E. (2012). Genre – monititeteinen näkökulma In V. Heikkinen, E. Voutilainen, P. Lauerma, U. Tiililä & M. Lounela (Eds.), *Genreanalyysi – tekstilajitutkimuksen käsikirja* (pp. 17-47). Helsinki, Finland: Gaudeamus.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

Hirschberg, J., & Manning, C. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. In Ldv Forum (Vol. 20, No. 1, pp. 19-62).

Iren, D., & Reijers, H. A. (2017). Leveraging business process improvement with natural language processing and organizational semantic knowledge. In Proceedings of the 2017 International Conference on Software and System Process (pp. 100-108). ACM.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jlailaty, D., Grigori, D., & Belhajjame, K. (2017a). Mining business process activities from email logs. In 2017 IEEE International Conference on Cognitive Computing (ICCC) (pp. 112-119). IEEE.

Jlailaty, D., Grigori, D., & Belhajjame, K. (2017b). Business Process Instances Discovery from Email Logs. In 2017 IEEE International Conference on Services Computing (SCC) (pp. 19-26). IEEE.

Jurafsky, D., & Martin, J. H. (2009). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition (2nd ed). Upper Saddle River, N.J: Pearson Prentice Hall.

Kreuzthaler, M., Schulz, S., & Berghold, A. (2015). Secondary use of electronic health records for building cohort studies through top-down information extraction. Journal of biomedical informatics, 53, 188-195.

Lan, M., Tan, C. L., & Low, H. B. (2006, July). Proposing a new term weighting scheme for text categorization. In *AAAI* (Vol. 6, pp. 763-768).

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011, December). Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 251-258). IEEE.

Lee, W., & Liu, B. (2003, August). Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*(Vol. 3, pp. 448-455).

Leopold, H., Pittke, F., & Mendling, J. (2015). Automatic service derivation from business process model repositories via semantic technology. Journal of Systems and Software, 108, 134-147.

Leopold, H., van der Aa, H., & Reijers, H. A. (2018). Identifying candidate tasks for robotic process automation in textual process descriptions. In Enterprise, Business-Process and Information Systems Modeling (pp. 67-81). Springer, Cham.

Lindsay, A., Read, J., Ferreira, J. F., Hayton, T., Porteous, J., & Gregory, P. (2017). Framer: Planning models from natural language action descriptions. In Twenty-Seventh International Conference on Automated Planning and Scheduling.

Liu, Q., Javed, F., & Mcnair, M. (2016). Companydepot: Employer name normalization in the online recruitment industry. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 521-530). ACM.

Manning, C., Raghavan, P. & Schütze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41-48).

Mehmood, M. K., Iftikhar, A., & Iftikhar, E. (2016). Automated goal detection from natural language constraints. In 2016 Sixth International Conference on Innovative Computing Technology (INTECH) (pp. 491-496). IEEE.

Mikkonen, K. (2012). Multimodaalisuus ja laji In V. Heikkinen, E. Voutilainen, P. Lauerma, U. Tiililä & M. Lounela (Eds.), *Genreanalyysi – tekstilajitutkimuksen käsikirja* (pp. 296-308). Helsinki, Finland: Gaudeamus.

Müller, O., Junglas, I., Debortoli, S., & vom Brocke, J. (2016). Using text analytics to derive customer service management benefits from unstructured data. MIS Quarterly Executive, 15(4), 243-258

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

Niboonkit, S., Krathu, W., & Padungweang, P. (2017). Automatic discovering success factor relationship entities in articles using named entity recognition. In 2017 9th International Conference on Knowledge and Smart Technology (KST) (pp. 238-241). IEEE.

Okoli, C., & Schabram, K. (2010), A Guide to Conducting a Systematic Literature Review of Information Systems Research, Sprouts: Working Papers on Information Systems, 10(26).

Orlikowski, W. J., & Yates, J. (1994). Genre repertoire: The structuring of communicative practices in organizations. Administrative science quarterly, 541-574.

Paré, G., Trudel, M. C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. Information & Management, 52(2), 183-199.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of management information systems, 24(3), 45-77.

Pinggera, J., Zugal, S., Weber, B., Fahland, D., Weidlich, M., Mendling, J., & Reijers, H. A. (2010). How the structuring of domain knowledge helps casual process modelers. In International Conference on Conceptual Modeling (pp. 445-451). Springer, Berlin, Heidelberg.

Piskorski, J., & Yangarber, R. (2013). Information extraction: Past, present and future. In Multi-source, multilingual information extraction and summarization (pp. 23-49). Springer, Berlin, Heidelberg.

Pittke, F., Leopold, H., & Mendling, J. (2015). Automatic detection and resolution of lexical ambiguity in process models. IEEE Transactions on Software Engineering, 41(6), 526-544.

Pratama, B. Y., & Sarno, R. (2015, November). Personality classification based on Twitter text using Naive Bayes, KNN and SVM. In 2015 International Conference on Data and Software Engineering (ICoDSE) (pp. 170-174). IEEE.

Process. (2019a). In Lexico.com. Retrieved September 30, 2019, from https://www.lexico.com/en/definition/process

Process. (2019b). In Merriam-Webster.com. Retrieved September 30, 2019, from https://www.merriam-webster.com/dictionary/process

Pustulka-Hunt, E., Telesko, R., & Hanne, T. (2018, August). Gig Work Business Process Improvement. In *2018 6th International Symposium on Computational and Business Intelligence (ISCBI)* (pp. 10-15). IEEE.

Ramakrishnan, C., Patnia, A., Hovy, E., & Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. Source code for biology and medicine, 7(1), 7.

Reichert, M., Hallerbach, A., & Bauer, T. (2015). Lifecycle management of business process variants. In *Handbook on Business Process Management 1* (pp. 251-278). Springer, Berlin, Heidelberg.

Revindasari, F., Sarno, R., & Solichah, A. (2016). Traceability between business process and software component using Probabilistic Latent Semantic Analysis. In 2016 International Conference on Informatics and Computing (ICIC) (pp. 245-250). IEEE.

Rosemann, M. and vom Brocke, J., 2015. The six core elements of business process management. In Handbook on business process management 1 (pp. 105-122). Springer, Berlin, Heidelberg.

Rummler, G. A., & Ramias, A. J. (2010). A framework for defining and designing the structure of work. In Handbook on Business Process Management 1 (pp. 83-106). Springer, Berlin, Heidelberg.

Rummler, G. A., Ramias, A. J., & Rummler, R. A. (2009). Potential pitfalls on the road to a process-managed organization (PMO), Part 1: The organization as system lens. Performance Improvement, 48(4), 5-16.

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. In Intelligent natural language processing: Trends and Applications (pp. 373-397). Springer, Cham.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project*. University of Pennsylvania, School of Engineering and Applied Science, Department of Computer and Information Science.

Sawant, K. P., Roy, S., Parachuri, D., Plesse, F., & Bhattacharya, P. (2014). Enforcing structure on textual use cases via annotation models. In Proceedings of the 7th India Software Engineering Conference (p. 18). ACM.

Strnadl, C. (2006). Aligning business and it: The process-driven architecture model. Information Systems Management, 23(4), 67-77.

Swenson, K., & von Rosing, M. (2015). Phase 4: What Is Business Process Management? In *The Complete Business Process Handbook* (pp. 79–88). Elsevier. https://doi.org/10.1016/B978-0-12-799959-3.00004-5

Tripathy, A., Agrawal, A., & Rath, S. K. (2016). Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, *57*, 117-126.

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. Procedia Engineering, 69, 1356-1364.

United Nations ESCAP. Business Process Analysis Guide to Simplify Trade Procedures. (2012). Retrieved July 15, 2019, from https://www.unescap.org/resources/business-process-analysis-guide-simplify-trade-procedures

Van Auken, K., Jaffery, J., Chan, J., Müller, H. M., & Sternberg, P. W. (2009). Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation. BMC bioinformatics, 10(1), 228.

Weske, M. (2012). Business Process Management. Springer, Berlin, Heidelberg.

von Rosing, M., Laurier, W., & M. Polovina, S. (2015). The BPM Ontology. In *The Complete Business Process Handbook* (pp. 101–121). Elsevier. https://doi.org/10.1016/B978-0-12-799959-3.00007-0

von Rosing, M., Foldager, U., Hove, M., von Scheel, J., & Bøgebjerg, A. F. (2015). Working with the Business Process Management (BPM) Life Cycle. In The Complete Business Process Handbook (pp. 269–345). Elsevier. https://doi.org/10.1016/b978-0-12-799959-3.00014-8

Wu, Z., & Palmer, M. (1994, June). Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 133-138). Association for Computational Linguistics.

Xhemali, D., Hinde, C. J., & Stone, R. G. (2009). Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages. *International Journal of Computer Science Issues (IJCSI)*, *4*(1), 16.

Yates, J., & Orlikowski, W. J. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of management review*, *17*(2), 299-326.

APPENDIX 1 THE BUSINESS PROCESS COMPONENTS

TABLE 43 Business Process components according to Dumas et al. and von Rosing et al.

Component	Definitions by Dumas et al. (2018)	Definitions by von Rosing et al. (2015)	Component
Business process	A collection of inter-related events, activities, and decision points that involve several actors and objects, which lead to an outcome that is of value to at least one customer.	"A set of structured activities or tasks with logical behaviors that produce a specific service or product."	Business Process
Event	Thing that happen "atomically, which means that they have no duration"	"A state change that recognizes the triggering or termination of processing."	Event
Activity	"fine-grained or coarse grained units of work"	"A part of the actual physical work system that specifies how to complete the change in the form or state of an input, oversee, or even achieve the completion of an interaction with others actors and which results in the making of a complex decision based on knowledge, judgment, experience, and instinct."	Process Activity
Decision Point	"points in time when a decision is made that affects the way the process is executed"	"Determines the forking and merging of paths, depending on the conditions expressed."	Gateway
Actors	"human actors, organizations, or software systems acting on behalf of human actors or organizations"		

(to be continued)

Table 43 (to be continued)

Objects	“Physical objects, such as equipment, materials, products, paper documents” and “Informational objects, such as electronic documents and electronic records”	“A real-world thing of use by or which exists within the enterprise and information objects reveal only their interface, which consists of a set of clearly defined relations. In the context of the business competency, the relevant objects are only those which relate to the enterprise’s means to act.”	Object (business and information)
		“A logical cluster of all sets of related data representing an object view of a business object.”	Data object
Outcome	Is either negative or positive	“a result and output generated by the enterprise. It has a combination of tangible and intangible attributes (features, functions, usage)”	Product

APPENDIX 2 THE SEARCH PHRASES AND THE NUMERICAL RESULTS OF THE QUERIES

TABLE 44 Search Phrase Combinations Used and the Numerical Results. All the queries had “since 2014” constraint

Phrases	EBSCO-Host	Google Scholar	IEEE	ACM
"Business Process Management" AND "Natural Language Processing"	8	1100	3	5
"Business Process" AND "Natural Language Processing"	11	4060	16	10
"Business Process" AND "Information Extraction"	2	2240	10	8
"Business Process" AND "Information Retrieval"	91	13 300	23	38
"Business Process Management" AND "Information Extraction"	2	601	2	4
"Business Process Management" AND "Text Mining"	1	1050	0	0
"Business Process Management" AND "Information Retrieval"	59	2660	6	25
"Business Process" AND "Text Mining"	3	3090	5	1
"Business Process" AND "Unstructured text"	0	913	0	2
"Business Process Management" AND "Unstructured text"	0	241	0	1
"Process Description" AND "Natural Language Processing"	0	374	2	0
"Process Description" AND "Information Extraction"	0	279	0	0
"Process Description" AND "Information Retrieval"	0	909	0	1
"Process Description" AND "Text Mining"	0	193	0	0

APPENDIX 3 THE PAPERS SELECTED FOR FURTHER INSPECTION

TABLE 45 Papers selected for further inspection.

EbscoHost (1)
Leopold, Pittke, & Mendling, 2015
IEEE (10)
Annervaz, George, & Sengupta, 2015 Pittke, Leopold, & Mendling, 2015 Niboonkit, Krathu & Padungweang, 2017 Jlailaty, Grigori & Belhajjame, 2017 Mehmood, Iftikhar & Iftikhar, 2016 Revindasari, Sarno & Solichah, 2016 Gao & Bhiri, 2014 Jlailaty, Grigori, & Belhajjame, 2017b De Medio, Gasparetti, Limongelli & Sciarrone, 2017 Pustulka-Hunt, Telesko & Hanne, 2018
ACM (4)
Iren & Reijers, 2017 Sawant, Roy, Parachuri, Plesse & Bhattacharya, 2014 Liu, Javed & Mcnair, 2016 Berardi, Esuli, Fagni, & Sebastiani, 2015
Google Scholar (3)
Lindsay, Read, Ferreira, Hayton, Porteous, & Gregory, 2017 Leopold, van der Aa & Reijers, 2018 Ferreira, Thom, & Fantinato, 2017

APPENDIX 4 PROCESS RELATED INFORMATION INCLUDED IN THE DESCRIPTION STRUCTURE

TABLE 46 Process related information included in the description structure

DLV	BPA	TIPS	SHIP	Customs
Process	Process	Process	Process	Process
Actors	Process participant(s)			Stakeholders
Steps	Activities	Tasks	Steps	
Inputs	Input and criteria to enter/ begin the business process			
Outputs	Output and criteria to exit the business process			
Assumptions				
		Messages		
	Associated documentary requirements criteria to exit the business process	Requirements		
	Rules			Reason
BPMN	Use case & Activity Diagrams	BPMN	BPMN	Other Diagrams