TEXT MINING IN FINANCIAL INDUSTRY: IMPLEMENTING TEXT MINING TECHNIQUES ON BANK POLICIES

MASTER'S THESIS

Masters of Business Informatics

Faculty of Science, Department of Information and Computing Science

First Supervisor:	Dr. Marco Spruit	Student (ID):	Drilon Ferati (5676932)
Second Supervisor:	Dr. Matthieu Brinkhuis		

Utrecht University

Utrecht, Netherlands

13/07/2017

ABSTRACT

With the increase in data, that organisations collect and create, the necessity to leverage from these resources has become apparent. This pool of data distinguishes two primary data structures, namely structured data and unstructured data. Both of these formats come with their own bag of techniques for scrutinising the data and extracting information and knowledge subsequently. Besides not having a predefined structure or representation, unstructured data also comprise roughly 80% of all the data that organizations possess. Policy documents are a good illustration of this kind of data, with their text-heavy format and domain specific language. As written guidelines of acceptable actions to which organisations must adhere, policy documents are present across industries and in a large number. This is especially true for organisations in the financial industry, such as banks, who continuously introduce policies in order to be fully compliant with regulations that governing bodies impose. In an attempt to bring order and some understanding to policies, this research investigates the applicability and benefits of TM on processing such documents. Relying on the DS principles, initially, the literature was consulted, to determine the extent to which such techniques have been exploited on policies. This investigation revealed that the use of TM on policy documents fell short in both qualitative and quantitative aspect. Next, to the limited amount of publications that treated these concepts, the variety of techniques that were examined was narrow. Hence, through a CS in one of the biggest banks in Netherlands, a set of unprecedented techniques were applied to policy documents. The use of IE to extract references between policies, together with the use of automatic summarization and keyword extraction, to retrieve a concise representation of the documents and a set of descriptive labels (tags) respectively, were evaluated both statistically and by experts. The results showed that to a large extent, these techniques are capable of analysing internal policies and extracting reliable information from them. Furthermore, this led to the introduction of a new TM framework for processing policies. The framework is a MAM of the approach followed in this study and it represents the harmonic use of three different techniques and the results that derive from their utilisation. Thus, next to unveiling the current state of literature, this research also introduces a novel approach for processing policies with the use of TM techniques.

ACKNOWLEDGMENT

Moments like this, make one take a step back and reflect on the path they have passed. It is at this point that, next to hard and joyful times, one sees the people that have stood by him along the way, making him understand how important they are. Thus, this part is not about me or the thesis, it is about everyone of *You* that have played a role during the two years of my Masters.

I would like to start with the person who made this thesis possible and who has guided me through every step of the way, my first supervisor Dr. Marco Spruit. His network together with his expertise and constructive feedback, made it possible to conduct a solid study. At the same time, I would like to thank my second supervisor, Dr. Matthieu Brinkhuis, for his valuable feedback along the process.

Next, I would like to thank Alexander Ackema for agreeing to take me in as an intern and the whole Provision Office at ABN AMRO for immediately accepting me as a member of the team. The lessons I learned from all of you, I will carry with me wherever I go. Next to this, I would like to thank the Policy Writers within ABN AMRO, for accepting to participate in this study and evaluate the results.

At the same time, I would like to thank my close family and relatives for accompanying me in every step of the way. Afraid to not forget anyone of significance, I would like to thank all of you for making this journey look easy and for giving the best advices to overcome every obstacle. I owe this thesis to all of you.

Last but not least, a big thank you goes to my friends for bringing the *"fun"* factor in this journey, with their cheerful personalities and excellent sense of humour.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATION

LIST OF ABBREVIATIONS

Abbreviation	Meaning
AI	Artificial Intelligence
BS	Behavior Science
CS	Case Study
DM	Data Mining
DNB	De Nederlandse Bank
DS	Design Science
EBA	European Banking Association
F	F-measure
HMM	Hidden Markov Model
IE	Information Extraction
IFRS9	International Financial Reporting Standard
IR	Information Retrieval
IS	Information Systems
IT	Information Technology
KDD	Knowledge Discovery in Database
KW	Keywords
LD	Levenshtein Distance
LSA	Latent Semantic Analysis
MAM	Meta-Algorithmic Model
MUC	Message Understanding Conference
NE	Named Entity
NER	Named Entity Recognition
NLIP	Natural Language Input Processing
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
P	Precision
PDD	Process Deliverable Diagram
POS	Part of Speech
QA	Question Answering
R	Recall
RQ	Research Questions
SQ	Sub-Questions
TF-IDF	Term Frequency - Inverse Document Frequency
TM	Text Mining
TOC	Table of Content
XML	Extensible Markup Language

1. INTRODUCTION

"*Software is eating the world*" - this is how Andreessen (2011) referred to the phenomena of organisations becoming more digital. Within itself, digitalization entails the daily creation and collection of large stream of data, which Spence (2010) referred to as "*the pollution of the information age since they are created and they are here to stay*". The exponential increase in data possession necessitated the need for organisations to leverage from these resources and extract information and subsequent knowledge from them. A frontrunner in enabling this process is the Knowledge Discovery in Databases (KDD) framework of Fayyad, Piatetsky-Shapiro and Smyth (1996). Furthermore, this large stream of data also revealed two distinct data formats, namely structured and unstructured data. Besides being different in structure, each of these data formats also required different treatment methodologies. Structured data are mostly numerical and they are found and collected from their designated fields in a proper format (Baars & Kemper, 2008). Data Mining (DM), processes structured data and attempts to extract patterns and the hidden relationship among them. Although this process is faster, and some even argue that is easier, only 20% of the overall data that an organisation generates and collects, is found in a structured format. The rest 80% of data come in an unstructured format (Grimes, 2008). These data have no consistency in appearance, and are usually text-heavy, making it a challenge to extract patterns and relationships from and among them. The existence of such data format, together with the large quantity, has made it highly important to develop methodologies that will bring meaning to their disorganised structure, thus separating the signal from the noise. This has led to the inception of the Text Mining (TM) field, which Witten (2004) defines loosely as *"…the process of analysing text to extract information that is useful for particular purpose".* Nevertheless, it is more eminent that it may sound, with its prominence being shown from the wide range of techniques that unveil hidden thematic in text, techniques that automatically extract relevant information, answer questions and so on, to its applicability in the biomedical, financial, security, social, and other similar domain (Bholat, Hansen, Santos, & Schonhardt-Bailey, 2015; Fan, Wallace, Rich, & Zhang, 2006b; Friedman, Johnson, Forman, & Starren, 1995; Haug, Ranum, & Frederick, 1990; Zhao, 2013). Nevertheless, next to all of these, there are still industries and sectors that have yet to utilise the benefits that TM and its techniques bring. An example of this is the use of TM in the financial industry, more specifically its use on internal bank documents. Evidence shows that banks have not been able to fully appreciate the benefits of TM on their internal documents, such as using them to acquire a better understanding of the internal policies. Hence, this study presents an initiative to contribute in this aspect. The study relies on the design science paradigm (Hevner, March, Park, Ram, & Ram, 2004), and its novelty is two folded: *1) identifying the current application of TM on internal policy document through a systematic literature study,* and *2) design an artefact for TM of bank policy documents*. By reviewing the scientific body of literature, TM techniques and frameworks that are applicable on bank policies will be identified and their utilization on bank policies will be evaluated. Ideally, the use of these techniques should employ an easy navigation and search through documents that are internal and integral for the organisation. Subsequently, through a case study in one of the biggest banks in Netherlands, the artefacts will be developed and evaluated. Additionally, the validity of this artefact will be estimated by means of user acceptance testing, where a stakeholder from the bank will evaluate the relevance of the retrieved results from the artefacts.

1.1 PROBLEM DESCRIPTION

The widespread applicability of TM, in various fields, has resulted in multiple TM frameworks being developed with some of them even resulting into systems. An example of this is the biomedical domain, which has been able to greatly benefit from the development of TM framework and systems that treat the complexity of the biomedical text. Nonetheless, not all fields have been able to taste the same riches. Policies are textual documents that are present in all businesses. The Oxford Dictionary (Dictionary, 2007) defines policies as "*a course of the principle of action adopted or proposed by an organisation or individual*". As such, policies represent written guidelines of acceptable actions to which organisations must adhere. Furthermore, although these type of documents are industry wide, they still lack standardisation. The lack of standardisation ramifies policies based on the purpose of compilation and the industry they are applied to (A. I. Anton & Earp, 2003). Additionally, since policies are distinct per industry, they make use of a domain-specific dictionary. This dictionary has often proven to make policies incomprehensible, which has subsequently resulted in difficulties of extracting meaning and knowledge from them (A. Anton & Earp, 2004). The financial sector is a broad term and in itself encapsulates various industries such as banks, insurance companies, accounting companies, stock markets, investment funds and so forth. Organisations in these sectors have witnessed the continuous introduction of policies that impact them on a daily basis. Banks, as industries in the financial sector, have their business activities regulated by internal and external governing bodies, which bodies introduce regulations to which banks should fully comply. In order to do so, banks continuously introduce new policies or update existing ones. Additionally, the manual effort that is put on performing these changes is extensive. Hence, banks are a testimonial to the difficulties that come with the policies. These difficulties together with the number of policies that banks have in place culminate to the following problem statements.

"It is not yet certain whether the appliance of Text Mining on (bank) policy documents has previously been studied and whether it is an appropriate technique for automatically extracting valuable information from them"

This problem statement is two folded. The significance of TM has been made apparent in the latest years. Its large scale of applicability has made it an appealing discipline to leverage from. Moreover, from a scientific perspective, there are numerous TM frameworks that have been developed for various purposes. Furthermore, it is evident that it can successfully provide a solution to various problems regarding textual documents. Nevertheless, it is not yet certain to what extent this technique has been applied to internal policies. Additionally, it is advocated that TM is an appropriate method for extracting relevant information from textual documents. Yet, it is not certain whether parallels can be drawn between this statement and bank policies. The problem statement also introduces the scope of the research that will follow. Although it was mentioned that policies are found across industries, this research will be limited only to bank policies. Furthermore, it will only concern internal bank policies, thus excluding policies that the bank shares with the general public. Thus, this indicates that the TM techniques will be evaluated based on the benefits they introduce when used for internal purposes.

1.2 RESEARCH QUESTION

Drawing from the defined problem statement, the following research questions are compiled.

RQ 1 – To what extent has TM been applied on policy documents?
As a multidisciplinary technique, TM has found applicability in a variety of situations. By conducting a thorough literature investigation, the first research question attempts to identify practices of applying TM on policy documents. Furthermore, driven by the fact that the focus point of this research is bank policies, the following sub-question derives.

SQ 1.1 – To what extent has TM been applied on bank policy documents?
Since the first research question focuses on policies in general, this sub-question attempts to be more specific. Building upon the results of the first research question, SQ 1.1 focuses in determining whether the identified literature on policy documents actually treats bank policies.

RQ 2 – Which TM techniques or frameworks have been applied on policy documents?
The following research question aims to explicitly identify techniques and frameworks that have been used in analysing policy documents. The existence of multiple TM frameworks is evident, with each of them treating a distinct problematic. With this in mind, the second research questions focus on TM frameworks that have been specifically developed for policy documents. Nevertheless, the fact that policies tend to differ from one another, the following sub-question derives.

SQ 2.1 – Which TM techniques or frameworks have been applied on bank policy documents?
Policies are domain-specific, meaning that they fully depend on the business domain for which they have been compiled. Thus, this sub-question aims to identify the existing techniques and frameworks that have been explicitly applied on bank policies.

SQ 2.2- Which linguistically oriented techniques have been applied on bank policy documents?
The text-heavy format and at the same time the lexicon that policies use, have an impact in the comprehensibility of such documents. These factors deem it necessary to identify linguistically oriented TM techniques that will enable to extract information with the highest linguistic values, which is what this research question aims to answer.

SQ 2.3 - What is the level of similarity between the bank and non-bank policies, in terms of TM techniques and frameworks that have been applied on them?
Depending on the outcome of the aforementioned question and sub-question, the third sub-question aims to identify whether there is a similarity in the techniques or frameworks that have been used in analysing bank and non-bank policies.

RQ3 – Which TM techniques can be used to obtain information that would enable an easier navigation through the policies?
The large number of policies that bank have in place calls for a solution that will enable an easy navigation through the same. This solution may be in a form of a repository that stores all bank policies, together with some relevant information for each. Thus, it is of a great importance to use techniques, that once used will enable the compilation of a policy repository that will employ an easy search of the same. Moreover, the answer to this research question will strongly rely on the identified TM techniques and frameworks from the previous questions.

SQ 3.1 – *To what extent can the followed case study approach, be generalised in a new TM framework?*
Taking into consideration that thus far there are multiple TM frameworks in existence, the following question aims to determine, whether the followed approach in the case study introduces a novelty and can be generalised to a TM framework for processing business policies. Hence, also enabling to contribute to the scientific body of literature.

1.3 OUTLINE

The structure of the thesis is as follows: Chapter 1 provides a brief introduction to the study at hand and the problematic it treats together with the research questions that derive from this problematic. Chapter 2 provides a description regarding the chosen research method. At the same time, it introduces the organisation that will facilitate the evaluation of the artefact by means of a Case Study (CS). Chapter 3 provides a thorough literature review of TM. Besides defining and introducing the difference between TM and DM, in this chapter some literature about the major TM techniques is provided. Additionally, here we dwell into the literature regarding the use of TM on policy documents. Furthermore, some literature regarding Natural Language Processing (NLP) is introduced, given its importance in the TM process. Chapter 4 gives a detailed description of the techniques that have been utilised in the CS and how they have been used. Chapter 5 reports on the results that derived from the CS and what these results mean for the scientific body of literature. Chapter 6 provides an elaborate answer to the research questions, whereas Chapter 7 discusses the generalizability of the followed approach, its limitations and future work.

2. RESEARCH METHOD

The defined problem statement, together with the research question, call for an extensive review of the existing work regarding the use of TM on policy documents. This was done by means of a literature review and Design Science (DS) paradigm. The literature review followed a snowball approach, as defined by (Wohlin, 2014), in order to create a corpus of studies that have investigated the use of TM. Furthermore, search terms such as "Text Mining in Finance", "Text Mining in Banks", "Text Mining on Policy", "Natural Language Processing" and multiple variations of the same terms, were employed to pinpoint studies that have treated a similar problematic. This review, besides providing empirical evidence on the current state of research regarding TM, also unveiled TM frameworks and models that can be used in analysing policy documents. With these inititatives the research questions were answered. Whereas, the designing of the artefact followed the DS paradigm as introduced by (Hevner et al., 2004).

2.1 DESIGN SCIENCE

In order to improve and make the most out of their Information Technology (IT), organisations rely on the use of Information Systems (IS), which is considered as a social science that has IT embedded in it. Since the early inceptions of this field, scientists have invested a vast amount of attention on investigating various phenomena regarding these systems. Peffers, Tuunanen, Rothenberger and Chatterjee (2008) defined it as an applied research discipline, meaning that theories from other disciplines are applied to solve problems at the intersection of IT and organisation. The research in IS can be summed up in two main fields of investigation: 1) research, whose subject of the investigation is the development of theories that explain human and organisational behaviour regarding the use of IS, and 2) means of creating and evaluating IT artefacts that are intended to solve organisational problems (Hevner et al., 2004). The former is known as Behavior Science (BS) paradigm, whereas the latter is known as Design Science (DS) paradigm. The DS paradigm has its roots in the "*science of the artificial*"(Simon, 1969), where the purpose is to create artefacts that will accomplish desired goals and resolve identified problems.

The DS research framework of Hevner et al. (2004) utilises the execution of a reliable and high-quality research. This also serving as a roadmap for the study at hand (Figure 1). In this framework, the "*environment*" section facilitates the problematic that is studied. Additionally, the business needs are defined by positioning the current technological infrastructure and architecture, against the organisational strategies and culture. This articulation of business needs is followed by two complementary IS research phases (BS and DS). Furthermore, the knowledge base provides materials (i.e. frameworks, theories, models, methods etc.) for these two types of research to be accomplished. In addition to this framework, Hevner also introduced guidelines to be followed in a DS research. These guidelines give a better understanding of the requirements that lead to an effective execution of the research. When adapted to the study at hand, these guidelines are as follows:

1. *Problem Identification*

   At all time, banks should be fully compliant with regulations that the governing bodies impose. In order to do so, banks continuously create new policies or adapt existing ones, resulting in an enormous amount of policies that they have in place. This, in conjunction with their text-heavy and domain specific dictionary imposes a challenge on the readability and manual extraction of information from them.

2. *Definition of the objectives of a solution*

   Given the challenges that come from the large cluster of policies, stakeholders require an easy and fast navigation through them. Thus the solution should employee an automatic approach of analyzing policies and retrieving relevant information in a timely fashion. If we try to classify the objectives according to the objective classes of (Peffers et al., 2008), it will fall under qualitative objectives since it introduces the benefits that an artefact will provide towards solving the problem.

3. *Design and Development*

   The data format of these type of documents, calls for TM technique to be used in the solution design. Thus, the new method relied on the combined use of both TM and NLP techniques, which combination relied on the literature findings.

4. *Evaluation*

   Evaluation assists in determining whether the developed artefact indeed has an impact on the problem. It puts the defined objectives of a solution against the actual solution, to determine to what extent the problem has been diminished. Hevner et al. (2004) present a number of artefact evaluation methods such as black box testing, white box testing, case study, field study, static analysis, dynamic analysis, simulation and so on. The difference between these methods is in the process of examining, monitoring and studying the performance of the artefacts. The evaluation of the developed artefact followed the case study approach, supported by test cases and statistical performance measurements.

5. *Communication*

   The subsequent coverage of the previous phases, enabled to communicate the generalizability and validity of the designed framework. This was based on the outcome from the evaluation.

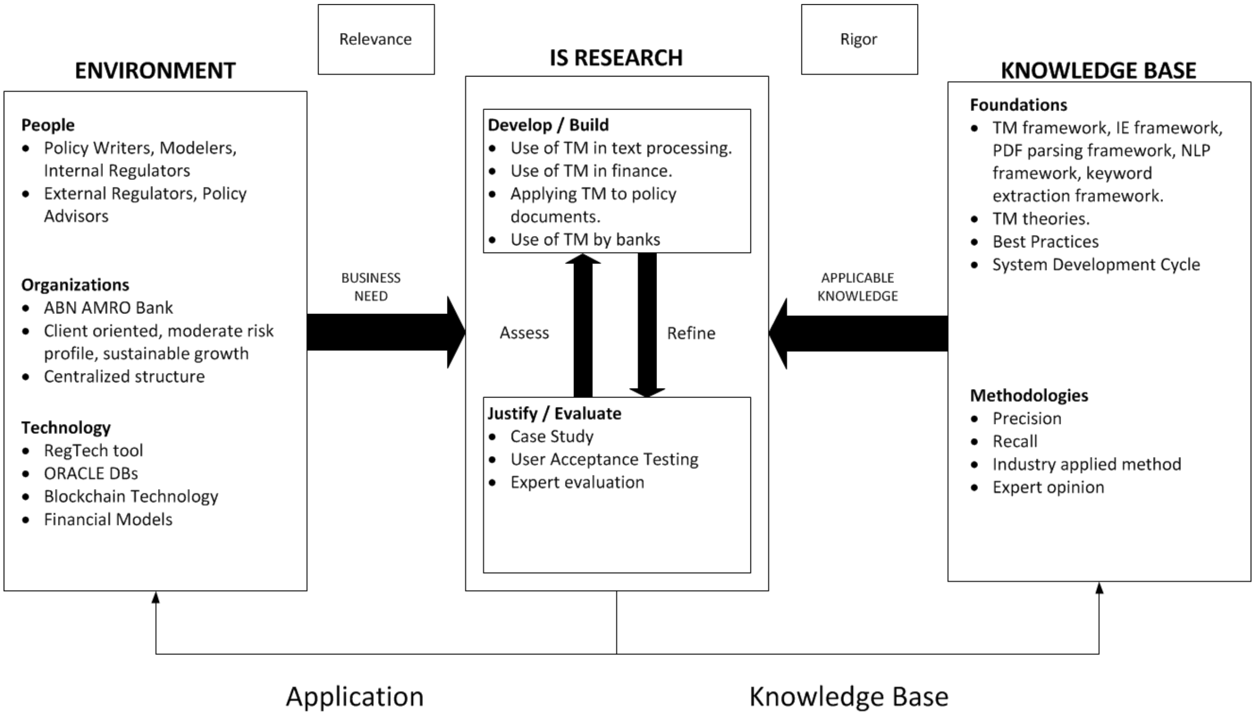
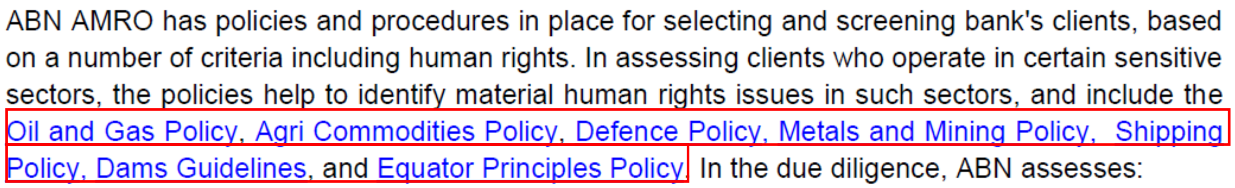


FIGURE 1 - DESIGN SCIENCE FRAMEWORK (ADAPTED FROM HEVNER ET AL., 2004)

2.2 CASE STUDY

When defining Case Study (CS), scientists have either focused on the research process (Yin, 2009) or the unit of study (the case) (R. E. Stake, 1995). Nevertheless, a good understanding of the goals of CS can be obtained by the work of (Zonabend, 1992). Zonabend states that in order to conduct a CS, a large amount of attention should be paid to the completeness of observation, analysis and reconstruction, of the case being studied. The general focus of the CS is in answering "*how*" and "*why*" questions, providing more attention towards what can be learned from the study (Stake, 1994). As such, CS can be single case or multiple case study. As the name implies, in a single CS the focus is entirely on one case being studied, whereas in multiple cases the focus is in two or more cases within the same study. The exponential use of CS led (Yin, 2009) to define three general approaches to designing a CS. Namely, these approaches are *a) exploratory, b) explanatory and c) descriptive.* In an exploratory approach, data are collected and are explored to identify the problem and subsequent solution or hypothesis. The explanatory is suitable when causality between entities is being studied (i.e. the effect of class A on class B). Whereas, descriptive CS requires the investigation to begin with a descriptive theory and face the possibility that problems will be faced along the study. Thus, drawing from these definitions, the study at hand represents an exploratory single case study.

With regards to this research, the CS was conducted at ABN AMRO Bank. ABN AMRO is one of the biggest banks in Netherlands with headquarters in Amsterdam. It provides financial services in 15 countries worldwide with roughly 22,048 employees. Alike other financial institutions, ABN AMRO has recognised the defined problem within their own settings. Currently, ABN AMRO together with other banks, operate under a lot of policies. The process of compiling a policy starts with a regulation that the governing bodies such as European Banking Association (EBA), Bank for International Settlements (BIS), De Nederlandsche Bank (DNB) and many others, introduce. These regulations may have been previously translated into policies, thus they are checked against the current repository of policies. Furthermore, policies are often related to each other, thus within the content of a policy, there are references to other related policies (Figure 2). The domain specific language together with the complicated network of interrelated policies introduces a bottleneck for fast and easy retrieval of information. A solution to this problem is thought to be through the composition of policy repository, which repository will consist of most representing information for each document. This approach is still in a design phase, where all the necessary attributes are being identified. Nevertheless, in the early phases of the design, one of the requirements was to have a list of all referencing policies that a given document refers to. Given the extensive amount of policies in place and at the same time their length, it is important to have this process automated. An additional requirement was to have a list of tags for each policy. The tags should be phrases and keywords that represent the main terms in the policy. The reason behind the automatic extraction of these entities was that policies are textual documents which leave space for multiple interpretations. Thus if they are given to competent entities for manual tagging, the risk of having inconsistent tags is high. Because of this, the use of TM techniques to generate such tags was thought to be a more reliable and unbiased solution. And last but not least, next to the references and tags, the stakeholders advocate that a concise summary of each policy should be present. This summary should draw on the main aspects of the documents, thus providing an understanding of the policy context in a timely manner.



ABN AMRO has policies and procedures in place for selecting and screening bank's clients, based on a number of criteria including human rights. In assessing clients who operate in certain sensitive sectors, the policies help to identify material human rights issues in such sectors, and include the Oil and Gas Policy, Agri Commodities Policy, Defence Policy, Metals and Mining Policy, Shipping Policy, Dams Guidelines, and Equator Principles Policy. In the due diligence, ABN assesses:

FIGURE 2 - REFERENCE TO OTHER POLICIES

2.2.1 VALIDITY OF THE CASE STUDY

In order to test the quality of the CS, Yin (2014) has introduced four logical tests that establish the quality of the research. These tests are as follows:

a) *Construct validity* **–** in general, this test deals with the selection of the correct operational measures for the concept that is being studied. In this case, it implies that from the wide range of TM techniques, only the ones that are applicable to the given problem statement should be selected. Critics have pointed out that in most of the case, scientists select measures based on “subjective” judgements. In order to avoid the threat to construct validity, the selection was based on the literature review and the techniques that it deemed as relevant for solving similar situations.

b) *Internal Validity* **-** this test is concerned with cases when casual effect between two events is being investigated. As such, this test is only applicable for explanatory studies. Given the fact that this study was earlier defined as exploratory in nature, this test was not applicable.

c) *External Validity* **–** is concerned with the generalizability of the findings beyond the study. In order to achieve this, we built upon the existing literature frameworks and architectures, and adapted them to the situation at hand.

d) *Reliability* – this test is concerned with the replicability of the study. This in case someone decides to repeat the study, they should be able to reach to the same findings. To assure that reliability is satisfied, a detailed description of the artefact was provided together with the technical solution behind it. Thus enabling the replication of the same study at any time.

2.3 TEST DESIGN

The developed TM method should provide a solution to the bottlenecks regarding the policies. This method should consist of TM techniques that introduce a potential for fulfilling the requirements. Nevertheless, when it came to evaluating the approach, there was a disruption in the knowledge base between the engineers of the artefact and the Policy Advisors or Policy Writers (the stakeholders). Given the domain-specific dictionary, the policies require some background knowledge in order to be understood. Individuals such as the stakeholders have the appropriate knowledge to extract meaning from these documents. On the other hand, engineering the algorithms requires a set of technical knowledge, an attribute that is not often found within the stakeholders. At the same time, it is not common to have engineers that are able to have an explicit understanding of what the relevant information within a policy are. This knowledge disruption had an impact on assessing whether the algorithm truly retrieves the relevant information. In order to bridge this gap, specific testing methods were used on evaluating the algorithmic results.

Generally speaking, testing compares the "*what is*" situation, with "*what ought to be*". A more formal definition of testing introduces it as "*The process of operating a system or component under specified conditions, observing or recording the results, and making an evaluation of some aspect of the system or component*"(Copeland, 2003). In the System Development Cycle, testing holds an important role, making it a crucial phase of the process. Nilsson (2013) introduces testing in terms of two practices, "*verification*" and "*validation*". Verification evaluates whether the system is built in the right way, whereas validation investigates whether the right system is built. This also introducing two main classes of testing, namely Black-Box and White-Box testing (Copeland, 2003). White-box testing focuses on the internal path, design and execution of the system that is being tested. Whereas, black-box testing solely focuses on the requirements and specification, thus requiring no knowledge on the design, implementation and internal path of the system. Furthermore, based on what the test will evaluate, there are several types of testing belonging to the two classes, such as unit testing, integration testing, functional and system testing, regression testing, beta testing and acceptance testing. As part of the black-box testing, acceptance testing determines whether the developed system satisfies the needs of the stakeholders. Thus to overcome the knowledge gap that was introduced earlier, and to have a proper result evaluation mechanism, the acceptance testing method was employed. Since it falls under the Black-Box testing class, acceptance testing does not dive into the technical specifications of the artefact. It only looks at the results and it evaluates their relevance. To follow this approach, test cases as in Table 1, were made available for the stakeholders, where they express their evaluation.

Test No.	**Test Type**	**Test Name**	**Purpose of Test**	**Test Date**	**Expected Results**	**Outcome**
1.	Algorithm	Extracting tags.	To determine whether the developed algorithm returns relevant information	12/02/2017	The meaningful tag phrases for this document are:	The algorithm returns tags that match the manually extracted tags.

TABLE 1 - EXAMPLE OF A TEST PLAN

3. LITERATURE REVIEW

By means of the Literature Review, the gap in the scientific body of literature is identified. Moreover, this review introduces various scenarios on how TM and its techniques have been implemented in different industries, with a greater emphasis on TM in the banking industries. Furthermore, these examples contribute in determining techniques that can be potentially used for the study at hand. To the best of our knowledge, there has not yet been a systematic literature review on TM, which could have been used as a starting point for the literature review in this study. Thus by using a collection of search phrases the initial cluster of scientific papers was generated, to which collection the snowballing approach was applied to further expand it for a more extensive review. The literature initially discusses the definitions of TM and its main techniques. It is then succeeded by reviewing the state-of-the-art research TM in Banks and Policy documents. Apart from providing an overview of the expansion of TM in different fields, the literature will also reveal methods TM and techniques that can be adapted to the study at hand.

3.1 DEFINING TEXT MINING

The existence of multiple TM definitions mirrors its wides-spread applicability.Currently, the literature provides a variety of definitions regarding TM, such as the one from Hearst (1999) who defines it as ”*the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources*. Nevertheless, this and many other definitions regarding TM have their roots in the definition of Feldman and Dagan (1995) who defined it as a *“ process of knowledge discovery from textual database”* (V. Gupta & Lehal, 2009; Tan, 1999). This definition in a way introduces the roots of TM since it originates from Frawley, Piatetsky-shapiro and Matheus (1992) definition of Knowledge Discovery (KD), which is defined as “*nontrivial extraction of implicit, previously unknown, and potentially useful information from given data*”. Given this definition, it is argued that TM has its roots in the KD field. Furthermore, the same is also used as a benchmark definition for defining Data Mining (DM). Nonetheless, although these two fields have similar definitions, multiple studies try to emphasise the difference between them. Thus introducing the early attempts of the scientist to differentiate these two fields. Frawley et al. (1992) and Rajman (1997) implied that the difference between these two fields is the type of the data that they use for KD. Thus, while DM employs information extraction techniques on structured data, TM does the same on unstructured or semi-structured data, which are often referred as textual data (V. Gupta & Lehal, 2009). Besides the data type, (Dijrre, Gerstl, & Seiffert, 1999) differentiated the fields on the complexity of steps they follow. The general steps that DM follows are *1) identification of collection, 2) preparation and feature selection* and *3) distribution analysis*. Although TM does not deviate from these steps, the feature selection is different since it is not practical to have human examine the features and to decide which ones to use. The other differing point is in the distribution analysis where highly dimensional vectors should be handled. This implies that there should be special versions and implementations of the DM algorithms. However, these differences do not prevent (Hearst, 1999) to state that TM is and extension of DM. It is not clear to what extent this statement may be true since there are no studies that agree or disagree with it. Some basic differences between TM and DM are also introduced in the work of (Berkouwer, 2009), which are also depicted in Table 2. Nonetheless, (Checklist & Management, 2008) introduces TM as an interdisciplinary field that relies on other disciplines such as data mining, information retrieval, computational statistics, computer science and linguistics. Aside from the distinction between TM and DM, (V. Gupta & Lehal, 2009) tries to draw the lines between TM and web-search. The difference is that in web search users look for something that is already found and it has been written by a person, whereas in TM the goal is to discover (previously) unknown information (Berkeley, 2003).

Text Mining	Data Mining
Relies on unstructured or semi-structured data	Relies on fielded (structured) data
Term extraction takes place based on semantic algorithm	Involves numerically based statistical analysis
Documents containing overlapping concepts can be organised together	Allows for temporal analysis
Documents containing overlapping concepts can be placed together partially	Clustering based on coding
	Involves co-occurrence matrices and histograms

TABLE 2 - DIFFERENCE BETWEEN TM AND DM (ADAPTED FROM BERKOUWER, 2009)

3.2 TEXT MINING TECHNIQUES

The rapid increase in the collection of unstructured data has urged scientist to investigate and develop techniques that will enable them to leverage from these type of data. According to (Xu, Liu, & Gong, 2003), these techniques are fundamental in enabling the efficient organisation, navigation, retrieval and summarization of large document corpus. The benefits of TM were always known, but only in the last couple of decades users have been able to leverage from it. Fan et al. (2006) state that the intention of TM is to create technologies that will combine the speed and accuracy of the computers with the human's linguistic capabilities. Nevertheless, in the early days of TM, (Fan et al., 2006b) enlisted technology as big limitation, since computers were far away from understanding the natural language. Feldman and Dagan (1995) made this obvious in their work where the technological limitations imposed only the execution of simple information extraction from the text in order to keep it at a reasonable cost. With the latest technological developments and the introduction of open source platforms for conducting TM operations, this limitation is being diminished. Nevertheless, as technology continued to improve, TM followed the same trend. Thus today the literature provides a large body of scientific literature regarding TM techniques such as Information Extraction (IE), Summarization, Clustering, Categorization, Topic Modelling and Question Answering (QA) and their applicability.

Wilson, Irwin and Lightbody (1997) define IE as "*a technique that locates and extracts specific information from textual documents, thus extracting structured information from text*". Furthermore, they introduce a framework that combines IE and data mining, in an attempt to develop a system that will extract a structured database from a textual corpus and will subsequently mine this database to acquire knowledge. Chang, Kayed, Girgis and Shaalan (2006) introduces an extensive survey regarding IE from web sources. Much of the progress in this technique, comes from the "Message Understanding Conference" (MUC) that was held between 1987 and 1998. The first two conferences focused on information regarding naval operations. MUC-3 and MUC – 4 treated the concept of analysing news articles about terrorist activities. Whereas, MUC- 5 and MUC- 6 treated news articles regarding joint ventures and change management (Mantrach, 2008). On another note, (Humphreys, Demetriou, & Gaizauskas, 2000) presents two scenarios of using IE in the biomedical domain for extracting information regarding enzyme reaction from journals and articles. Much of the research that has been conducted regarding TM in biomedicine, (Meystre, Savova, Kipper-Schuler, & Hurdle, 2008) breaks it down into two categories: focusing on biomedical text and focusing on the clinical text. The first implies the text that is found in literature in books, in articles, whereas the latter refers to the textual documents that are written by clinicians in clinics.

Badr (2011) defines summarization as a process of compiling a shorter version of a textual document that perceives the main information, content and meaning. The author also introduces extractive and abstractive as two summarization methods. Through abstractive approach, a main understanding of the main concepts is created, which concepts later are represented in a clear natural language. Whereas extractive approach extracts key textual elements based on statistical analysis regarding their frequency, location and so forth. The literature is overly populated with cases where automatic summarization has been applied. Teufel (2001) argues that with regards to scientific papers, automatic summarization is as useful as human-generated summarization and significantly more useful than the authors' abstract. Nenkova and Bagga (2003) investigated the usefulness of summarising threads in help forums in order to determine whether the thread is relevant or not, whereas (Tucker & Whittaker, 2008) summarised meetings for a better and easier navigation through meetings content.

Text categorization is defined as the assignment of textual documents to a set of predefined categories, based on their content (Niharika, Latha, & Lavanya, 2012). The authors (survey) also introduces a text categorization process together with some methods for categorization such as Decision Tree, k-Nearest Neighbour, Bayesian Approach, Neural Networks and so forth. Larkey (1999) uses text categorization to organise patent documents in order to make their search easier. Furthermore, (Drucker, Wu, & Vapnik, 1999) and (Weiss, Apte, & Damerau, 1999) have used text categorization for filtering spam e-mails and it has proven to be beneficial.

Iiritano and Ruffolo (2001) defines clustering as a technique which does not have predefined classes and as such it groups objects based on their similarities. There are numerous clustering algorithms, which the author also introduces in his work together with some similarity coefficients. Dijrre et al. (1999)

provide a deeper explanation of the differences between clustering and categorization as TM techniques.

Topic Modelling is another TM technique that has been widely used. Blei (2012) defines it is an algorithm that uncovers the hidden theme from a large collection of textual documents, which is not possible to be done manually by a human. At the same time, Blei states that Latent Dirichlet Allocation is one of the simplest topic modelling approaches. Denicia-Carral et al. (2006) define QA as a system that enables users, through a natural language query, to acquire not a collection of documents that contain the answer, rather acquire the answer itself. At the same time, he argues that currently, QA systems are capable of treating and answering questions of the factoid or definition type, whereas (Juárez-González et al., 2007) introduces frameworks for answering the two types of questions respectively. Liu (2016) introduces a web-based system that answers biomedical questions and at the same time provides encyclopedia-type commentary to the questions.

3.3 TEXT MINING APPLICATION AND FRAMEWORK

The enormous focus on its techniques and its applicability to various domains have resulted in multiple TM frameworks being developed. These frameworks differ from one another based on the purpose they are developed for, the abstraction level and so on. One of the most used and adaptable TM frameworks is the one presented in the work of (Fan et al., 2006b) and which is also visualised in Figure 3. Its generalizability is what it makes it so applicable. It initiates with a collection of textual documents, which documents are retrieved and preprocessed by the TM tool. Further on, the pre-processed documents follow an analysis phase, where the techniques in this phase are repeated until knowledge is discovered. In the visualisation of the framework, we witness only three techniques. These techniques are not imposed by the author in any way, thus in this phase, there can be a combination of other techniques depending on the goals that one intends to achieve with this process. With the help of the techniques, the extracted information is placed in a knowledge management system from where a considerable amount of value can be acquired by the user of the system.

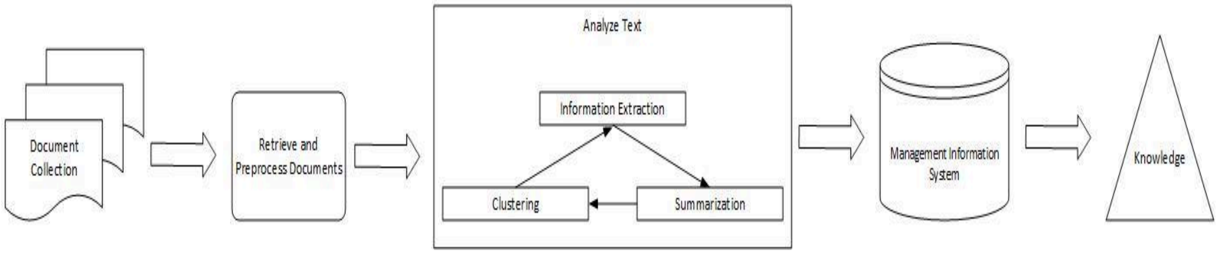


FIGURE 3- GENERIC TEXT MINING FRAMEWORK (ADAPTED FROM FAN ET AL., 2006A)

A generic TM process is also introduced by Kongthon (2004b). This process goes through five major steps, namely: 1) document retrieval, 2) data extraction, 3) data preprocessing (cleansing), 4) data analysis and 5) data visualisation. Nonetheless, there are frameworks that are more specific. An example is the Term Extraction Module developed by (R. Feldman et al., 1998). This framework goes through four main stages. It starts by reading the collection of documents and it resumes with the Linguistic Processing phase where various NLP tasks are executed on the textual data. Moreover, the Term Generation phase is self-explanatory, where terms are created based on the association coefficient between words, which leads to the last phase where terms are filtered based on a statistical computation of the terms and their appearance in the collection of documents. The authors present this as an advanced approach to generating tags for a collection of documents.

Another framework, which is more generic than the previous one, is that of (Tan, 1999). It consists of two main phases, text refinement and knowledge distillation. Text refining converts the collection of textual documents into an intermediate form, which can be either document-based or concept-based. The knowledge distillation phase extract patterns across documents from document based intermediate form (i.e. clustering, categorization, visualisation), and it extracts patterns across objects or concepts from the concept-based intermediate form (i.e. predictive modelling, associative discovery). The framework that (Kongthon, 2004a) introduces, is to a large extent, developed for knowledge discovery from articles related to technical fields. The framework is composed of four main phases, data extraction and cleansing, text summarization, text clustering and visualisation. Aside from the previously mentioned frameworks, the literature evidenced the existence of a collection of different TM frameworks, all being developed for a specific purpose.

3.4 TEXT MINING IN FINANCIAL INDUSTRY

Costantino, Collingham and Richard (1996) argues that the financial sector is suffering from a large stream of financial news and articles from various on-line news providers, financial newspapers and so on. They advocate that if these data are properly processed, they can aid in the process of decision making regarding various investments. Although in their paper they do not present an actual system, they introduce the benefits that the use of NLP and IE can have on these investments. In addition to this (Das, 2014), (Hagenau, Liebmann, & Neumann, 2013), (Junqu De Fortuny et al., 2014) and (Schumaker & Chen, 2009) have all built systems that attempt to predict stock movements by processing the textual news. Some of these systems have proven to be quite successful, even outperforming trading experts in some cases.

Kloptchenko et al. (2004) applied a combination of data and text mining techniques to financial reports of firms, in order to find hidden indication about the firms' future financial performance. They argue that although data mining has been used quite often for analysing financial reports, it is the textual part of the report that contains the most important information. The study indicated that the results from the qualitative and quantitative data analysis did not correspond. They argue that this is due to the fact that the numerical part of the report indicates the past performance of the company, whereas the textual part holds some details about the future performance of the company.

Gupta and Gill (2012) conducted a similar study on financial statements by examining the qualitative disclosure in the footnotes of the financial statements. In most of the times, textual information such as auditor's comment together with the financial ratio can be found in the footnotes of financial statements. They introduce a 5 steps TM approach for detecting financial statement fraud.

Nevertheless, the authors present a conceptual approach which to the best of our knowledge has not yet been tested and as such its validity is questioned. The work of (N. Street, 2000) is another example of using TM for analysing financial statements. He confirmed the feasibility of predicting the upcoming short-term financial performance of organisations, through text classification. As for (Kloptchenko, Magnusson, Back, Visa, & Vanharanta, 2004) and (Kamaruddin, Hamdan, & Bakar, 2007), both used TM for detecting deviations in financial statements, which is believed to be quite successful. Another study regarding TM in the financial sector is the one of (Karanikas, Tjortjis, & Theodoulidis, 2000). The author employs IE techniques to extract information regarding news where specific actors and the relationship among the actors is mentioned (i.e. Company A takeover Company B). The outcome of this research was a TM tool, regarding IE and clustering of documents, called TextMiner. These are just a small portion of the studies that have used TM for analysing financial statements. Nevertheless, with the exception of Karanikas et al. (2000), most of the other studies only introduce a conceptual framework which is yet to be tested in a later stage of the work.

3.4.1 TEXT MINING IN BANKING INDUSTRY

The usefulness of TM has not gone unnoticed by the financial sector. One of the early applications of TM in finance is the development of ATRANS (C. Street, 1986). This system was developed to extract information in a form of template from messages regarding money transfer between banks (often referred to as Telex message). Given the predictability of the content of the messages, this system resulted in being quite useful. This system was developed in a time when TM as a term was not yet incepted, at that time it was referred to as Text Analysis, also representing the start of using TM in banks.

Bholat et al. (2015) state that although the true potential of TM is ignored by the majority of banks, they do find the use of text mining on a daily basis. As an example of TM technique that operates in the background, the author enlists the use of spell checker before publishing documents, the spam detection firewalls or the use of queries for retrieving existing literature on a given topic. Due to the fact that the intentional use of TM by the banks is currently limited, the author presents some examples where TM can be used. Those examples concern the consistency in the messages that the bank communicates to the outside world and determining whether there is a conflict or complimentary between the policies that they enact.

Hendry and Madeley (2010) used TM to investigate the statements issued by Bank of Canada and to determine what type of information affected returns and volatility in short-term and long-term interest rate market. They used Latent Semantic Analysis (LSA) which was able to highlight the press releases of Bank of Canada which had more impact on market returns and volatility. The impact was more evident in the short term market.

Bruno (2016) used TM to analyse the Governor's Concluding Remarks of the Bank of Italy. The aim of this research was to 1) show the word frequency distribution features of documents, 2) extract the evolution of sentiment and polarity from the texts, 3) evaluate the index of readability of the texts and 4) measure the popularity that the document gained in the web. Besides providing a text analysis diagram, the author also provided the used algorithms for analysis, such as the formula for polarity evaluation which resulted that the text followed a neutral flow over the extent of the speech and the formula for estimating the Automated Readability Index (ARI) that showed that the documents had a 12-15 score which implies college degree readability level. Other studies have also treated the communications that central banks have conveyed and their impact. Such is the work of (Boukus & Rosenberg, 2006) who analysed the Federal Open Market Committee (FOMC) minutes through the LSA methodology and determined that these statements were correlated with current and future economic outlook, treasury yield changes and level of monetary policy uncertainty. On a similar note, (Moniz & De Jong, 2014) investigates the communications released by the central banks and their impact on financial investor's interest rate.

3.5 TEXT MINING OF POLICIES

In the large corpus of publication, only a small portion of them treated the subject of using TM on policy documents. Furthermore, most of the policies used in these studies were privacy policies, or more commonly known as "Terms and Conditions" or "Terms of Use". Surprisingly, these publications did not see it as relevant to apply the TM techniques on internal policies as well.

The ambiguity and text-heavy format of privacy policies has driven researchers to examine ways of bringing sense into them. Such is the case with Liu, Ramanath, Sadeh and Smith (2014) who contribute towards resolving this problematic. The majority of privacy policies address common issues such as the collection of user location, user contact, financial information, and so on. Furthermore, given the fact that the gathering of these data should be disclosed in the policy, the authors present a classification method that aligns sections of policies from different sources, based on their content similarity. Policies were collected from the most widely used website, resulting in a corpus of 1,010 policies. These policies then were manually split into sections and paragraphs in order to align segments of policies based on the privacy issue they address. In their quest to do so, the authors chose to employ the Hidden Markov Model (HMM). This choice was based on the motivation that HMM is able to capture the transition between topics (i.e. Topic B often follows after the discussion of Topic A). Additionally, the authors compared the outcome from HMM to the clustering algorithm implemented in the work of Zhong and Ghosh (2005), which algorithm performs bisections until it reaches a desired amount of clusters. The experiment showed that at the section level, the clustering algorithm performed better than HMM. Whereas, in the paragraph level HMM managed to outperform the clustering algorithm. This research is considered to be a follow-up investigation of a previous research from the same authors (Ramanath, Liu, Sadeh, & Smith, 2014). The motivation behind this research was two folded: a) initially to provide understanding to users on the implications of agreeing to the policies, and b) assisting legal entities in creating a better understanding of the content of the policies enabling them to provide improvement recommendations.

A similar investigation on website privacy policies was conducted by Massey et al. (2013), on a corpus of 2,061 privacy policies from Google Top 1000 visited websites and Forbes 500 companies. Their research focuses on three characteristics of privacy policies, namely: a) assessing the readability of these documents for requirement engineering; b) examine if automated TM can indicate whether a policy contains requirements outlined as either privacy protection or vulnerability; and c) whether the identification of privacy protection and vulnerability can be generalized to other policies. For each of the aforementioned characteristics, the authors relied on methods that have been proven to address the specified features. Thus, in an attempt to address the readability of the documents, the authors used five metrics, such as: Flesch Reading Ease (FRE) , Flesch Grade Level (FGA) (Flesch, 1948), Automated Readability Index (ARI) (Count, Reading, & Personnel, 1975), SMOG (McLaughlin, 1969) and FOG (Gunning, 1952). All these metrics measure different readability aspects from one another, where FRE assesses how challenging a document is to read (on a scale from 0-100) and FGA assesses how many years of education is needed in order to be able to understand the document. Whereas, for indicating whether the policy represents a privacy protection or vulnerability and whether it can be generalised to other policies, the authors used Topic Modeling. Topic Modelling enables to uncover the hidden thematic of a large collection of documents, which is nearly impossible to be done through human annotation (Blei, 2012). It does this by making three assumptions for each document: 1) The documents are made of topics and the topics are made of words; 2) Topic identification is done automatically rather than manually; and 3) The topics are shared across the collection of documents. The results show that all the policies in the study but at the same time policies, in general, are quite challenging to read and understand. Furthermore, the use of Topic Modeling could indeed determine whether a policy holds requirements that are expressed either as privacy protection or vulnerability. Additionally, the results introduce a preliminary support to the generalisation of the approach, nevertheless the authors argue that further research is needed to confirm the findings.

In parallel to the previous papers, Costante, Sun, Petković and den Hartog (2012) use TM and machine learning techniques to associate privacy categories with policies that address these categories. It is commonly known that the majority of websites today collect data from their visitors. As such, websites are required by law to disclose this fact on their platform together with an explanation of what data is being collected and how those data are being shared, used and stored. This information is provided either through a link on the website or by asking the visitors to agree to the privacy policies of the website. Nevertheless, a common assumption is that most of the visitors tend to neglect these notifications thus agreeing to the unconditional acceptance of sharing their personal data. Because of this, authors propose a system that uses TM and machine learning to assess the completeness of the privacy policy. It is anticipated that such a system informs users whether the privacy categories of their interest are being covered by the policy. The system requires from the user to select which privacy categories do they want to view (i.e. Data collection, Data sharing, Retention time, Security, Cookies etc.). Once deciding on this, the system detects the paragraphs in the policy that addresses these issues and assesses its completeness in terms of degree of coverage for the specific categories. The completeness is calculated with the following Equation 1:

$$G_c(p) = N \cdot \sum_{i=1}^{n} w_i \cdot c_i$$

EQUATION 1: COMPLETENESS EVALUATION

Here, $N = 10$ and $\sum_{i=1}^{n} w_i$ are a normalization factor which scales the scores in between 0 and 10. Furthermore, n represents the quantity of privacy policies, whereas w_i is the weight that may be in the range of 0 and 1. Finally, c_i is a binary variable that takes the value of 0 (if not covered) or 1(if covered). The machine learning and text categorization techniques are used to determine the value of c_i . Text categorization is used to label the paragraph to the thematic categories it belongs, whereas, machine learning is used to build an automatic classifier by learning from a pre-classified set of documents. Since text classification has been widely treated in the literature, it has resulted in multiple learning algorithms being developed for this purpose. In an attempt to choose the best performing algorithms, the authors compared the results from multiple algorithms, such as: Naïve Bayes, Linear Support Vector Machine (LSVM), Ridge Regression, k-nearest neighbor (k-NN), Decision Tree (DT), and Support Vector Machine (SVM). Their fitness was measured by means of precision and recall (Makhoul & Kubala, 1999). With respect to a class, Precision measures the rate of items being assigned to the class which actually are from that class. Whereas, Recall measures the rate of items from the respective class which actually are labeled as belonging to that class. This two values are combined in a single score through the F-measure formula. The results showed that the best results came from LSVM and Ridge Regression, which reported an F-measure of approximately 92%. This result was tested against a manual annotation from an expert and it showed that the algorithms performed reasonably well.

Given the fact that privacy policies are intended to protect the vendor rather than the user, they are intentionally written in a complex and ambiguous way. Often, there are ambiguous words in the policies, which are used to hide concerning issues and often result in giving vendors access or rights to actions that may not be in the user's best interest. Because of this, Stamey and Rossi (2009) introduce Hermes. Hermes is a system that attempts to provide a better understanding of privacy policies. It considers policies as a collection of topics expressed in a single document. Since topics are made up of words, it identifies the most important words for a specific topic, this also provides an understanding of which topics does the policy cover. Whether it expresses what information is being collected, what technology is used to collect that information and to whom may the information be disclosed. It does so by processing policies and analyses its content to identify possible ambiguous words. This is achieved with the use of Latent Semantic Analysis (LSA), which is a technique that is used to discover the semantic relation between words. It creates a latent relationship based on the dimensionality of the words, thus words with the lowest dimensionality are closer together. The results show that the system is able to identify the main topics of the policy and the most significant words for each topic together with a collection of words that express ambiguity. On top of this, it provides a score that shows the similarity of the user entered policy with a typical privacy policy.

One of the sparse publications that take internal business policies, as the subject of their investigation, is the research of Li, Wang and Zhao (2007). In their work, the authors turn their attention towards policies that define or constrain business processes, or as they are commonly known as process policies (i.e. order fulfilment policy, travel reimbursement policy, product development policy etc.). Although these policies are of a significant importance to entities such as executives, auditors or process owners, going through the lengthily policies to understand the business process often results in a tedious and time-consuming task. Thus, by employing TM and IE techniques, the authors propose a framework that automatically extracts process models from business policies. This algorithm consist of four major steps, namely:

- Step 1: Process Policy Selection – here process policies are separated from non-process policies. This is thought to be a form of text classification approach.
- Step 2: Process Component Identification - based on the classification in Step 1, this step aims to identify major process component such as task, organisational resources and data item. This is thought to be done with the use of Named Entity Recognition (NER).
- Step 3: Component Relationship Extraction – having identified the components in the previous step, the current step extracts the relationship among them, with a particular interest in relationships such as ordering relationship between tasks, relationship between the resources of the organisation and dependency relationship
- Step 4: Process Model Construction – having identified the process components and their relationships (Step 2 and Step 3 respectively), the final step constructs the process model.

Although theoretically this framework introduces a novelty in extracting processes from business policies, it has yet to be implemented in practice. Hitherto, only Step 1 and Step 2 of the framework has been tested. Additionally, the authors found NER labelling with HMM as unsuitable for their problematic thus turning their focus to customised statistical models for extracting the entity types. Whereas for Step 1, the authors used bag-of-words and tree kernel methodologies. The methods were implemented on sentence base, meaning that each sentence was labelled as "process" or "non-process". Furthermore, for learning the classification model, the authors used Support Vector Machine (SVM) for both methods. The results show that both methodologies are reluctant towards labelling sentences as "process", with tree kernel weighting more on non-process labelling. Overall, bag-of-words showed to be more powerful in identifying process sentences.

In a digital age, where websites collect as much personal data as possible, from their visitors, Costante, den Hartog and Petković (2013) attempt to inform users and raise awareness. They advocate that from all the uncertainties that come with reading a privacy policy, the users should be well aware of which data is being collected and thus judging whether that is acceptable. To settle the dust, the authors provide a solution that analyses policies automatically and identifies which user details are being collected by the provider. On top of this, the authors aim to provide a scoring system for a website based on the privacy data they collect. In order to extract the list of the data that the website collects, the authors turned towards Information Extraction (IE). Through IE architecture, analysis can be done based on the semantics content of the text rather than word frequency and appearance/absence. Initially, they analysed the privacy policies in order to discover the way the data collection description is stated. This revealed that data collection statements usually consist of entities such as Data Provider, Data Collector, Collection Tools, Provider Action, Collector Action, Tools Action and Personal Data Collected. Having identified the entities, the authors manually annotated them in order to create a training set, against which set, the testing set was evaluated. Further on, the authors used extraction rules in order to extract the content of these entities from the text. The constructed system was able to reach an extraction accuracy of 80%. Although the authors see this as a reasonable and beneficial accuracy, they think that this score can be further improved by adding ontologies, thesaurus, co-references and so forth.

While most of the studies focus on the privacy policies of arbitrary web sites, A. I. Anton and Earp (2003) focus on data that financial institutions collect. Driven by the Gramm-Leach-Bliley Act (GLBA), which argues that policies should be "clear and conspicuous", the authors aim to determine whether the financial privacy policies adhere to this act. They investigated the policies using goal-driven requirement engineering and readability metrics. The authors used goal mining heuristics for extracting a total of 1,032 goals from 40 policies. Furthermore, they divided the extracted goals into privacy protection goals and privacy vulnerability goals. Having the heuristics in place together with vulnerability/protection taxonomy, enables the users and the providers to compare and analyse Internet privacy policies. On top of this, the results showed that a high education level was required for one to have a complete understanding of policies. This puts into question whether privacy policies, of financial institutions, are really "clear and conspicuous".

Different from the aforementioned researchers, Xiao, Paradkar and Xie (2011) do not seek to shed light on privacy policies. On the contrary, the authors attempt to extract Access Control Policies (ACP) from requirements documents. In general, ACPs are used to specify to what resources do specific user have access to, and as such, they represent a mechanism that prevents security vulnerabilities. Manual extraction of this specification often results in a tedious task and it exposes the process to human error, making automation even more important. Although they are often found buried under a huge pile of requirements, this specification usually follow a certain declaration style (i.e. [subject] [can/cannot/ is allowed to] [action] [resources]). Thus, the authors propose a three step NLP technique that would automatically extract these instances. Initially, the text is analysed linguistically, with words and phrases being annotated based on their semantic meaning. It then constructs model instances with the use of annotated words and phrases. And finally, it transforms these model instances into a formal specification. For linguistic analysis, the authors rely on Shallow Parsing, to determine the syntactic structure of the sentences. It annotates the words with Part –of – Speech (POS) tags (Voutilainen, 2003), which are later used to construct patterns for extracting phrases and clauses from the text (Semantic-Pattern Matching). On top of this, it uses domain-specific dictionary, which has predefined semantic classes, in order to address words with *Negative* semantics and verb synonym. With the annotated words in place, the authors use these annotations to construct models. The models are created by constructing semantic patterns that extract the entities that construct an ACP. The models are tuned to extract even words with negative semantics if they are part of the ACP. Having the models established the authors transform them to eXtensible Access Control Markup Language (XACML), which is a formal specification language. For the authors, this cycle represents a new system, to which they refer as Text2Policy. The evaluation of this system showed that Text2Policy performs reasonably well and its automation has a significant impact on reducing the effort for extracting security policies from software requirement documents.

Brodie, Karat and Karat (2006) introduce SPARCLE (Service Privacy ARchitecture and CapabiLity Enablement), as a system that will assist organizations in linking privacy policies with their implementation. Seeing that there is a lack of such a system, Brodie, Karat and Karat (2006) see SPARCLE as a potential solution which receives privacy policies in a natural language, processes them to identify policy elements and outputs a machine readable version of the policies (XML). The authors use a similar approach as (Xiao et al., 2011), for identifying important information from the text. The elements are identified with the help of a shallow parser (Grefenstette, 1996) which linguistically analysez and annotates the syntactical structure of the policy text, for the sole purpose of creating grammar rules to identify the necessary information from the text, similar to Xiao et al. (2011)'s approach. In the evaluation phase of the system, SPARCLE expressed the ability to parse policies with an accuracy of 94%, which is reasonably high. In this paper, the author explains the system and its abilities in a more generic level, failing to go into depth of the technical aspect of the system.

Ammar, Wilson, Sadeh and Smith (2012) also investigates the possibilities of automatically categorising privacy policies. By analysing a corpus of 56 privacy policies, the authors found three main concepts that were shared through the policies. Those concepts are related to the "*ability to leave the services*", "*transparency on law enforcement*" and "*notify before changing the terms*". Thus, depending on whether the identified concepts appeared in the policies, the authors used logistic regression to classify them into two categorical labels, namely: "present" and "absent". Nevertheless, since it was still a work in progress at the time it was published, the authors do not provide much description on the technical aspects of achieving this, except for describing their approach in three brief steps:

1. Convert the document into a vector representation, with dimensions corresponding to unigrams, bigrams and trigrams and a value that denotes term frequency
2. Calculate a scalar score as the inner product of the vector and optimised weight vector
3. If the score is above the threshold, it is hypothesised that the concept is present, if not, it is hypothesised that the concept is absent.

The report on the findings showed that the approach was to some extent suitable for identifying "transparency on law enforcement" concept, and not that suitable for "ability to leave the service". There was no trace of reporting for the third concept, that of "notify before changing the terms", making the overall report a bit vague. Nevertheless, their report on future work showed that the authors knew what the potential cause of the weak performance may be and how it can be improved.

Michael, Ong and Rowe (2001) sees the need for organisations to have a "policy workbench". This integrated system, would enable users to access, search and update, policies from a centralised repository, through an interface, with the sole purpose of facilitating policy adherence. Different from the other studies, this research does not focus on privacy policy, instead they distinguish three types of policies: a) meta-policy; b) goal-oriented policy; and c) operational policy. Meta-policies, are policies that have a form of hierarchy and belong to a higher-level policy. Goal-oriented policies are policies that declare the desired goal but do not mention the process towards achieving the goal. Operational policies, define actions but not goals. Their approach towards developing the "policy workbench", was based on the architecture that Sibley, Michael and Wexelblat (1991) describes. This is an extensible architecture that theoretically describes how such a system should be developed and what are the main entities of the architecture. Nevertheless, for this study, the authors only focused on the first entity, that of processing the user input, or as they call it Natural Language Input Processing (NLIP). The system expects a user input that is formulated in a natural human language. It then automatically, tokenizes, parses (LT CHUNK) and tags (LT POS) the inputted text, in order to identify and extract main elements (i.e. subject, object, attributes and verbs) (Grover, Matheson, Mikheev, & Moens, 2000). The initial results of the experiment did not show promising results, thus leading the authors to conclude that the program should be refined so it can handle more complex statements.

In early 2000, initiatives were taken to codify privacy policies for a better understanding. At the peak of events, these initiatives were manifested with the development of policy codifying standards such as "Platform for Privacy Preferences (P3P)" and "Do Not Track" (Sadeh et al., 2013). These and many other standards of the same kind translated the policies from natural language format to a machine readable format (usually XML). Nevertheless, even though the standards have been under a constant improvement throughout the years, organisations seem to be reluctant in adopting them. This lack of cooperation is what motivated Sadeh et al. (2013) to investigate ways of improving the readability of privacy policies with the use of NLP, privacy preference modelling, crowdsourcing and policy interface. With regards to the language processing aspect of the policies, the authors suggest that in analysing privacy policies researchers should go beyond text categorization. Nevertheless, only a framework approach is presented in this publication, with no proven results. As far as NLP goes, the authors suggest that they will experiment with text categorization and semantic parsing, but make no comment on the linguistic and statistical techniques they will rely on.

3.6 TEXT MINING OF LEGAL TEXT

Legal text is composed of domain specific language, thus meaning that a good knowledge of legal terms and phrases is needed in order to fully understand the content of such a text. Nevertheless, since laws concern all citizens, they should be accessible and understandable to the general public, even ones that do not have domain knowledge. Thus, with the use of a TM, aim to create a method that enables users to search and receive legal document using everyday vocabulary. Their approach starts with users input, regarding their legal issues, using everyday vocabulary. This input then goes through a transformation phase where they are mapped to related legal terms. This step goes through two phases, the training phase and the query phase. In the training phase, TM techniques together with expert review, are used to determine a set of k legal terms. This set of terms is used as a vector base for all the documents. Whereas, the query phase, uses Google Similarity Distance(Cilibrasi & Vitányi, 2007) to transform the user input into related legal terms. This is followed by the use of cosine similarity (Mihalcea, Corley, & Strapparava, 2006), which computes the similarity of two documents. The results of this approach were evaluated by experts who deemed this method as an appropriate way to extract relevant legal text and see it as a superior method to existing query systems.

A great deal of attention has also been paid to automatic summarization of legal text. These investigations have yield different methodologies, systems and tools for automatic summarization. An example of this is the work of (Farzindar & Lapalme, 2004). In this paper, the authors, present the LetSum (Legal Text Summarizer) system which summarises justice decisions based on four thematic structures, INTRODUCTION, CONTEXT, JURIDICAL ANALYSIS, CONCLUSION. Their approach follows through five main phases. Initially, the pre-processing phase splits the text into paragraph, sentences and tokens, which are later annotated with their POS tag. With this in place, the thematic structures such as the ones mentioned previously, are determined. This is followed by the filtering phase, where irrelevant text such as citations is excluded. The selection phase finds all the relevant units and extracts the highest scoring units which then leads to the production of the summary. In this work, besides the pre-processing phase, text mining only comes into expression in the selection phase. For extracting relevant entities, the score is computed based on the position on the paragraph in the document, the position of the paragraph relative to the thematic segment, position of the sentence in the paragraph and the word distribution of the document (tf-idf). Nevertheless, no further explanation is given on how these factors are determined and how each of them impacts the score.

In the realms of TM legal documents, a contribution is also made by (Laws, 2008). In parallel with the previously mentioned publication, the authors see the legal text as a document composed of a collection of themes/categories. This also justifying their motivation of identifying the class of sentences. Furthermore, it was observed that sentences in different categories followed a specific construction pattern, this resulting in a total list of 81 patterns. This observation was used to analyse the text and classify each statement based on the pattern they followed. Although the authors do not delft into the details on how this recognition is made, it is safe to say that such a thing can be achieved with the use of POS tags and grammar rules based on those tags (which in itself is a form of IE). The experiment was conducted on Dutch laws, for which laws the results turned out to be quite promising, with an overall classification accuracy of 94%.

A rather unorthodox, but yet interesting research, is that of Dozier and Haschart (2000), who identifies the attorney and judge in case laws and creates a hyperlink from their name to their biographical details which reside in a repository. This is achieved through four main modules, namely the extraction module, the matching module, insertion module and the loading module. TM properties are present in the first module, where the author uses cue phrases and paragraph position to extract details such as an attorney, judge, date and court, from the case laws. These details are identified with the use of semantic parses, which is a commonly used technique in IE. Although TM is more applicable to the first module, it is worth mentioning that for the matching module, the authors use Bayesian Network. The network uses six pieces of evidence and five pieces of evidence to match attorneys and judges respectively. This combination of techniques proved to achieve a precision of 99% and recall of 92% for linking attorney name to biographical records, and a precision of 98% and recall of 90% for linking judge name to biographical records.

Summarization of legal text is also the aspect that Saravanan, Ravindran and Raman (2006) investigated. Their approach follows a combination of graph-based technique and term distribution model for identifying relevant content from the document. The authors introduce Conditional Random Fields (CRFs) as a graphical model which has proven to be one of the best frameworks for text segmentation. This graph model is used for identifying the rhetorical structure of the text, which structure unveils the text segments that can be combined into constructing an entire summary. This approach in combination with K-mixture model for generic sentence extraction, enable the extraction of a well-structured. Which summary entails a better coherence and readability for the user. The evaluation yields impressive performance results of text segmentation, with an accuracy of 90%.

References have a significant role in legal documents, this justifying their extensive appearance in legal content. References provide extensive information in a legal text, like how to interpret a specific rule, to which law it refers, to which decree it refers, to which convention it refers and so forth. These factors are the motivation behind the work of (Martínez-González, 2005) who seeks to extract the references of legal text, for storing in a centralised repository. They follow a three-step approach where initially the content of the document is analysed. These are done in order to detect the references in the document and distinguish them from the rest of the content. The authors advocate that the references in a legal document follow a specific pattern, this also being the reason why the focus of the second step is in recognising these patterns. Their recognition is made by means of the lexical structure of the reference text, which structure also assists in composing extraction grammars. The result of these two initial steps is a set of reference data for each policy, which data are taken to the third step where they are matched to other available legal item. The presented approach is quite simple and straightforward, with no use of domain specific dictionaries, which is also reflected in the recall score of 74%. Although this score may seem moderate, it is quite an impressive one given the fact that no domain specific parsers were used.

Besides the aforementioned publications, there have been numerous other researchers that have investigated the gains that can be derived from using TM on the legal text. Thus, Moens, Uyttendaele and Dumortier (2000) emphasises the importance of using discourse structure knowledge and linguistic cues in IE systems, in order to properly process the specific text of legal documents. On the other hand, Chieze, Farzindar and Lapalme (2010) introduce DecisionExpress system, which system automatically summarises court decision and provides a consisted and informative version of the same. The use of rhetorical structure for automatic summarization, of legal documents, was also noticed in the work of (Grover, Hachey, Hughson, & Korycinski, 2003). It has been applied on judicial transcripts from House of Lords, which is the highest court of England. Brüninghaus and Ashley (2001) investigated ways to simplify the cumbersome process of indexing legal text. They hypothesised that the abstraction of actors and events, capturing multi-word actions and recognising negation can lead to a better representation for automatic case indexing. Their experiment showed that domain specific heuristics and state-of-the-art IE tools can successfully identify actors. Galgani, Compton and Hoffmann (2012) aim towards a summarization of legal text with the use of knowledge acquisition. The acquisition process was conducted with the use of rules that extract specific information. This resulted in a precision of 87.4% for extracting relevant information.

Nevertheless, the presented literature is only the tip of the iceberg when it comes to publication regarding the legal text. The applicability of TM on legal documents has been extensively investigated and has resulted in multiple publications. Scientist have shown interest in this domain for a couple of decades now, this also is reflected in the amount of systems that have been developed for processing legal document. The systems range from complex ones to more simple ones, from systems that use heuristics to extract information to systems that use linguistic to extract information and the countless techniques that have been used to summarise such documents. This mirrors the evolution of the systems which goes in parallel with the maturity of the field.

3.7 NATURAL LANGUAGE PROCESSING

Ling, Maccartney and Penn (2011) define NLP as "*a theoretically motivated range of computational techniques for analysing and representing natural occurring texts at one or more levels of linguistic analysis for purpose of achieving human-like language processing for a range of tasks or applications*". Breaking down this definition we can see that *"range of computational techniques"* refers to the range of techniques that can be chosen to accomplish a particular type of language analysis. *"Natural occurring text"* refers to the fact that the text can be of any format (written or oral), mode, genre and language, as long as it is a language that it is used for communication among humans. *"Level of linguistic analysis"* refers to the variety of language processing types that can be applicable when human produce (or comprehend) language. *"Human-like language processing"* introduces the fact that the roots of NLP are within Artificial Intelligence (AI). Thus, this breakdown, walks us through the requirements, range of applicability, technique and origin of NLP. Turban, Sharda, Aronson and King (2008) state that the goal of NLP is to move beyond a text manipulation that is syntax-driven (also known as bag-of-word), towards a manipulation that will understand the grammar, semantics and context of the natural language text. Kumar (2011) provides a brief introduction to the history of this field and its developments through the four eras, with the Turing Test as the inception point from where the development of various NLP systems has followed.

Although the entire field is referred as Natural Language Processing, (Ling et al., 2011) advocate that there are two main focus points within the field. One is natural language processing and the other one is natural language generation. The former refers to the process of analysing the language in order to compile a meaningful representation. The latter refers to the compilation of language from a presentation. Liddy (1998) and Feldman (1999) advocate that in order to achieve a proper natural language processing, the system should be able to make a distinction among the different levels of language in order to acquire an understanding the same way humans do. These levels of language are phonological level, morphological level, lexical level, syntactic level, semantic level, discourse level and pragmatic level. The literature emphasises the apparent contribution of NLP frameworks in operations, such as information retrieval (IR), information extraction (IE), question answering, automatic summarization, machine translation and so on. Furthermore, these frameworks have been used in developing language engineering tools. Cunningham (2002) is an example of an NLP framework, whose robustness has enabled it to be used in the development of systems such as GATE. Cunningham, Maynard and Bontcheva (2002) introduces some of the systems that have found GATE usable, such as MUSE (Maynard, Tablan, & Ursu, 2001) which is a Named – Entity recognition system capable of processing texts from a wide range of domains and genres, MUMIS is another system that uses GATE in order to produce formal annotation about essential football events. Furthermore, (Mokhov, 2010) introduces MARF which is an NLP framework that has found applicability in applications regarding speaker-identification, language identification, natural language probabilistic parsing and so on. Another NLP framework that the literature introduces is TectoMT (Popel, 2010), which is used for language translation (English - Czech). Although nowadays there are multiple NLP frameworks in circulation, (Collobert & Weston, 2008) advocates that in most of the cases the same NLP tasks are integrated for developing the frameworks. He resumes stating that these tasks are designed to apply a divide and conquer approach to the text for better analysis. Manning et al. (2014) provide a depiction of this approach, which is also visualised in Figure 4. Nonetheless, (Nadkarni, Ohno-Machado, & Chapman, 2012) argues that the execution of these tasks does not always yield the best results, especially in the medical domain, where it is more difficult to tokenize since biomedical text often consist of characters that are usually used as token boundaries (i.e. “30 mg/day” - forward slash is a token boundary).

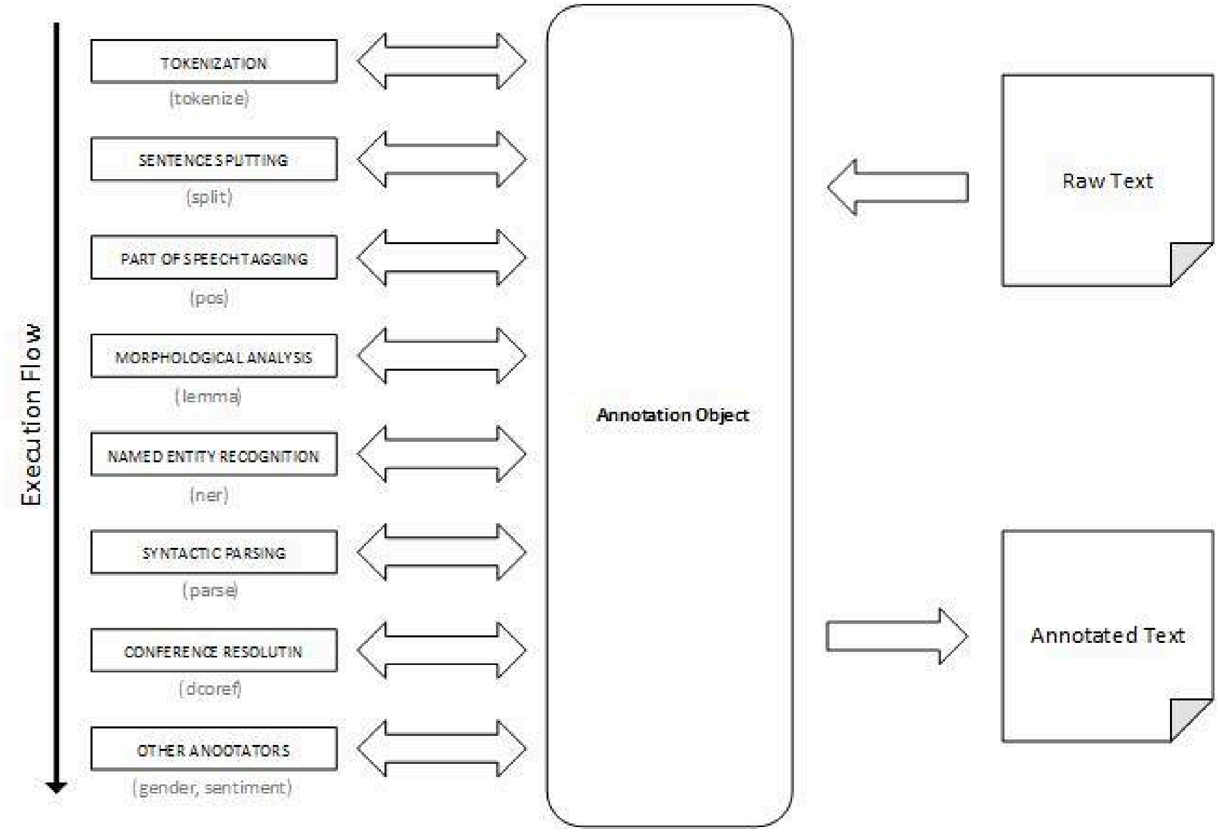


FIGURE 4- NLP TASKS (ADAPTED FROM MANNING ET AL., 2014)

4. CASE STUDY METHODOLOGY

Most of the data that organizations create and collect are found in an unstructured format. As such, this kind of data make up for the majority of information resources, which resources support the decision making process. Although unstructured data represent the prime source of information for most organizations, they still require a considerable amount of processing effort to unveil these information. This comes as a result of both, their extensible amount (which is continuously increasing) and the format in which they are found. Such a scenario is shared across multiple organizations, making it a common bottleneck across industries. The symptoms of unstructured data, and its attributes, have triggered the interest of multiple scientist to investigate processing techniques that can ease the task of information extraction from such data. This claim is also supported by the extensive amount of publications that have treated this phenomena, and at the same time by the continuous introduction of state-of-the-art technology for processing unstructured data. The large pool of publication showed different scenarios and motivations for processing textual data. Whether it was for identifying hidden thematic in text, extracting specific information from text, identifying specific actors in the unstructured content, developing an indexing mechanism and so on. Often, the aftermath of such investigation was a new system that provides technological solution to one or more aspects of unstructured data. Nevertheless, among all of these publications, a shared cause was visible, and that had to do with bringing order to unstructured data and automatically extract information from them.

Even the CS at hand did not shy away from the same problematics. In this digital era, it is quite common to have a designated system for specific purposes. Nowadays, organizations invest heavily in introducing new intra-organizational systems for various purposes such as communication, knowledge sharing, project planning, secure file sharing, and so on. Nevertheless, while all of these innovative initiatives take place, there are other problematics that are left in the shadow. Such a case is evident in the financial industry. In a time when financial organizations work under a large stream of regulations from governing bodies, to the best of our knowledge, there is no digitalized approach for handling such developments. The release of a new regulation often was associated with the introduction of a new policy or update of an existing one. The high frequency by which these regulations are introduce and at the same time the lack of digitalized solution for handling changes, often imposed difficulties in governing policies. A similar scenario was visible in the institution where the case study took place. ABN AMRO has recognised the need for more dexterity when it comes to dealing with policy documents. This especially since policies are being introduced and updated continuously across departments and organisation, which makes the process of managing them quite difficult. At the time, policies are stored in designated folders from where they are accessed and managed. Although so far this approach has proven to be somewhat suitable, with the latest development in the financial industry, ABN AMRO feels the need to have a more digitalized solution in place. An initial step towards digitalization was to move from a decentralized manner of accessing policies, to a more centralized facilitation of the same. As a solution to this was seen the creation of a centralized repository that contained all the policies, together with some descriptive information for each document. Nevertheless, as text-heavy documents, policies hold an extensive amount of important information in themselves, which led to discussing what information should then be stored in the repository. In a discussion with Policy Writers and Advisors to identify this information, but at the same time having in mind that the stored information should be consistent across different documents, three types of data were deemed as necessary to have in the repository.

Within their content, policies often refer to other internal policies, for various purposes. This references can be found dispersed throughout the whole content of policies which can be of numerous pages. Policy writers find these information important, nevertheless searching through the text for them often results in a tedious and time-consuming task. Thus, it was perceived as relevant to have a list of all the policies to which a given document refers to. Besides this, policy writers advocate that for each policy there should be a list of keywords. The keywords will be a form of a tag for each policy, through which tags the repository will be queried in order to retrieve the relevant policies to the user. All of this making up for an easier navigation through the collection of policies. In addition to this, a brief summary for each policy should be stored, enabling the user to get a fast understanding and assessing whether the policy is relevant for the case.

4.1 DATA GATHERING

With a good understanding of how a centralised repository should ideally look like, and what kind of descriptive information it should contain, the process of data gathering was initiated. Typically, in this process, a considerable amount of data is collected, upon which data the experiments is executed. There are different approaches to gathering the relevant data, depending on the study circumstances. Nevertheless, considering the fact that this study is conducted within the settings of an organisation, the data gathering process was straightforward. Within the institution, policies are grouped based on the topic that their content treats. This in itself represents a manual text clustering mechanism. From a wide range of text cluster, the policies that treated the concept of Credit Risk were made available for the study. This document cluster counted roughly 25 policies that treated different aspects of Credit Risk such as how they are identified, measured, mitigate and so forth. As mentioned earlier, these policies were stored in their designated folder, where each folder contained a PDF and Word version of the policy. Nevertheless, upon investigating the group of 25 policies and hence 25 folders, it was unveiled that a couple of policies were not readily available for the study. Taking into consideration that the study only concerns textual documents that are written in English, the manual investigation of the policies revealed that one of the policies was written entirely in the Dutch language. This difference in language disables the uniform and automatic processing of the textual documents since different languages require different methods of analysis. Because of this, it was decided that the policy written in the Dutch language would be excluded from the case study. Additionally, it was revealed that one of the folders was empty, not containing any document. Thus from the collection of 25 policies that were initially made available, only 23 were available for the study, given that one was written in a foreign language and one was inexistent. Additionally, in terms of content, the volume of the available policies ranged from 11 pages up to 45 pages per document including cover pages and appendices. Since these policies are considered as internal documents, and as such they are not meant for the eyes of general public, some anonymization was required, which is further explained in the following subsection.

4.2 DATA PREPARATION

As the name implies, unstructured data do not have a defined structure and as such can be found in different formats. Henceforth, the whole TM methodology revolves around the idea of creating some form of order and structure around these data. Nevertheless, in most of the cases, before any analysis takes place, the unstructured data undergo a process of preparation. This data preparation phase can be of a various kind, all depending on the CS and the intended outcome of the study itself. Thus, shifting the focus to this experiment, it was evidenced that the available data corpus required some preparation. As the first step of data preparation, it was decided that parts of the policy documents such as cover page, table of content, tables and appendices should be excluded from the analysis. Cover pages and table of content were removed since they do not contain information that introduce any level of significance for the study and if left they can present outliers. Additionally, the study only concerns the main written content thus motivating the decision behind removing appendices and other tables in the text.

Furthermore, it is often advocated that when dealing with TM, the most appropriate data format to conduct the analysis, is the plain text format (.txt). Nevertheless, as it was mentioned earlier, the data that were provided for this study were available in PDF and Word format. Due to the difficulties of processing data in such formats, the need to convert the files into appropriate .txt format was apparent. Given the relatively small corpus, the data conversion could have been done manually or by using online platforms for conversion. Nevertheless, since the corpus contains sensitive data that are only meant for internal use, converting them with the use of online platforms, represented the risk of leaking and disclosing confidential data. Due to this reasons, the online conversion was discarded for safety reasons. Additionally, since one of the aims of the study is to automate some processes, the option to manual convert the files was also excluded. Thus the batch converting of the files through a programming script was seen as the best solution towards having appropriate data format. Initially, all of the available documents were copied from their designated folders and were saved in a centralised folder. From there, each of the documents was individually fetched, processed and converted. The conversion was done by constructing a Python script. Python is a high-level general-purpose programming language, whose comprehensive and large amount of standard libraries enable it to be used effectively in scientific computing (Sanner, 1999). Furthermore, this was made possible with the use of a Python package named "*pdfminer*". As such, *"pdfminer"* is a package written in Python language which is designated for processing PDF files and converting them into other formats such as HTML, .txt, XML and so forth.

As the final step of data preparation, all the document titles were anonymized. The real title of the policies was replaced by the prefix "*Policy",* followed by a letter from the English alphabet. As mentioned earlier, these documents are meant only for internal entities and as such, they should be anonymized when presented to the general public. Nevertheless, this was only possible for the title of the available documents and not for other referring policies within their content, since those are the focal point of this study and should not be manipulated. If this data preparation process is summarised, it can be wrapped up in the following manner:

- Removal of cover pages

- Removal of table of content (TOC)
- Removal of tables
- Removal of appendices
- Conversion of the documents to .txt format
- Title anonymization

4.3 TEXT MINING

An additional benefit of data preparation is that one gets familiar with the structure and the content of the documents that will be processed. This step also introduces one of the first attempts to bring uniformity to these data, such as transforming them from different formats to a single common format that is suitable for NLP. Furthermore, there are multiple things that one should account for while preparing the data, since any failure in handling or recognising irrelevant data, may have an impact and have an influence on the results in the later stages. An example of this can be the table of content aspect of a document, which if not handled properly, may skew the results and as such deem the experiment as improper.

Nevertheless, upon having the cluster of documents in the proper format, the implementation of TM experiments can take charge. The literature showed a rich variety of TM techniques, which were mostly constructed to handle specific situations or documents. Thus, even before starting to implement TM techniques, there should be a good understanding of the purpose of using such a technique and what the intended outcome is. The answer to these questions unveils the approach and the necessary techniques that are needed for processing the unstructured data. Same was the case with the study at hand, where initially the purpose of processing the documents was determined. From here, knowing that the overall intention of TM in this study is to form a kind of indexing mechanism for documents, this tells enough about which techniques should be used. Considering the requirements for the repository, three TM applications were considered appropriate for this case. These applications were, namely IE, automatic summarization and automatic keyword extraction. IE is thought to be an adequate approach to extract reference policies, automatic summarization for creating a concise and meaningful summary for each policy, and finally automatic keyword extraction for extracting the relevant and most descriptive keywords from each policy.

The extensive use of TM has resulted in multiple tools, applications, software packages and libraries being developed for this purpose. Furthermore, most of these state-of-the-art tools are open source and free of use. Thus, at this point one can afford to make a decision between using a single tool or a combination of tools, to achieve the desired outcome. Nevertheless, not necessarily all of the available TM techniques are generable and can be used in multiple domains. The literature showed that most of the times the systems, support a specific language dictionary thus not being appropriate for cases that do not consider such a language or that convey a different message. Additionally, although the literature showed a considerable amount of systems, the nature of documents made it impossible to use open source platforms. Given the sensitive data that the policy documents contains, the stakeholders did not approve the use of such free of use systems for security reasons.

Nonetheless, this did not impose any critical difficulties, since solutions can always be found in the pool of packages and libraries that have been developed explicitly for processing textual data and natural language. Additionally, these packages have found support in a variety of programming and scripting languages. Thus, one can easily find a solution by using Perl programming language and all its capacities for processing textual data, such as presented in the book of (Bilisoly, 2011). Or by using the R programming language, and the infrastructure it provides for processing textual data with the use of *tm* package (Feinerer, Hornik, & Meyer, 2013). Nevertheless, besides the aforementioned programming languages, Python is another language that has proven to be quite successful in processing textual data. As mentioned earlier, currently there are numerous libraries developed in Python that make this language more than suitable for complicated and scientific computations. One of the biggest Python libraries, which at the same time is specialized in processing textual data, is the Natural Language Toolkit (NLTK) library by Bird, Klein and Loper (2009). Its simplicity, extensibility and uniformity, complimented by the shallow learning curve of Python programming languages, is what makes this combination advisable (Loper & Bird, 2002). Additionally, it's use has extensively been documented in two different books, "The Natural Language Processing with Python" of Bird et al. (2009) and "Python Text Processing with NLTK Cookbook" by Perkins (2010). Due to this factors, the TM experiments of this study, are conducted with the use of Python and its NLTK library.

4.3.1 INFORMATION EXTRACTION

In itself, TM encapsulates a variety of techniques that can be utilised to extract various form of information from textual data. Given the inconsistency in such data format, the relevant information that one wishes to extract may appear in different forms. There are cases that the most important details of the textual document are not explicitly mentioned, but they are hidden in the large content of the text. These differences, also impose selection criteria for the TM techniques. Henceforth, the techniques for extracting explicit information may not always be appropriate for extracting hidden information from the text. Such a difference in information format was also visible in the policy corpus of this study. Generally speaking, references are pieces of information that are emphasised in the textual document, or formatted in a way that is distinguishable from the rest of the content. This was also true for the collection of textual documents that make up the corpus of this study. In most of the cases, the reference policies in the corpus followed the same manner of citation. This made them recognizable from the rest of the text. Furthermore, this format of information, motivated the use of IE to extract them from the policy text. IE is one of the most utilized techniques for extracting explicit information from unstructured data.

Over the years, this TM technique has been used to find structured information from unstructured text and encode them in a format that is appropriate for populating a database (Cardie, 1997). Its extensive applicability has resulted in IE systems suitable for different purposes such as IE system for analysing life insurance, systems for summarising medical patient records, systems for analysing news and newspaper articles, systems for automatic classification of legal text and so on (Cardie, 1997). The wide range of systems also hints the number of IE architectures that have been utilised. These architectures do not tend to differ much from one another, nevertheless they unveil the two spectrums of IE that are known to researchers. In their core, IE systems either rely on the linguistic processing of text for extracting valuable data, or they rely on keyword matching techniques with little or no linguistic processing. Both techniques have proven to yield appropriate results when applied. Nevertheless, with regards to the corpus of policies, although the references follow the same form of citation for most of the cases, they do not contain the same concepts or terminology. This introduces some difficulties in using an IE architecture that relies on keyword matching techniques. Additionally, considering the research questions and at the same time the use of the NLTK module, this research endorsed the linguistic processing approach of IE. At the same time, NLTK provides its own IE architecture which promotes the linguistic processing aspect. This framework, is initiated with the splitting of raw text into sentences. This splitting is then followed by the division of the sentences into tokens of words. To each word token, a part-of-speech (POS) tag is assigned depending on its semantic and syntactic structure. The added value of this tags comes to light when recognising and constructing named entities to be extracted. The cycle finishes by identifying the relationships among different entities of interest thus enabling the full extraction of relevant information. The visual depiction of this architecture can be found in Figure 5.

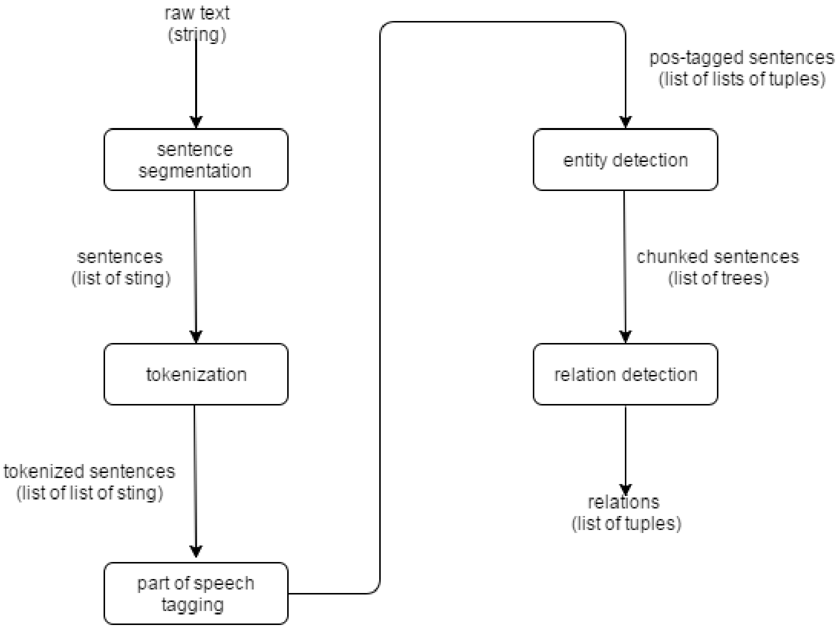


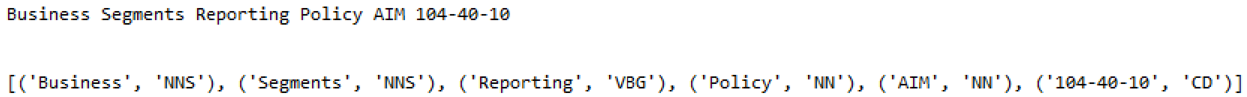
FIGURE 5- IE ARCHITECTURE (ADAPTED FROM BIRD ET AL., 2009)

In the Data Preparation phase, it was mentioned that all the policies were converted into a text format that only consisted of main content, excluding parts like cover pages and appendices. With only the main content in place, sentence splitting and word tokenizing step from the IE architecture were initiated. These two steps, were made possible with the use of NLTK and the modules it provides for segregation, namely NLTK sentence segmenter and NLTK word tokenizer. These two steps initially split the entire textual content into sentences, which sentences were further tokenized into individual words. This drove to the next step of the pipeline, where the tokenized words were analyzed based on their semantic and syntactic representation. This enabled to assign a POS tag to the words, respectful to their grammatical description and position in the sentence. One of the things that POS tagging does, is that it transforms a list of words into a list of tuple that has the following format *(word, POS-tag).* Throughout the years there have been multiple POS tagging methods, with the earliest being the Rule-Based Method, which used a set of human constructed rules for tagging (Martinez, 2012). However, the labor-intensive process together with the need of highly qualified people with excellent linguistic knowledge, made this method inefficient and prone to errors. Nevertheless, much has changed since then, with the introduction of methods such as Transformation-Based Learning where rules are not manually created but they are learned from the corpora, and probabilistic methods such as Markov Model (Charniak, 1996) and Maximum Entropy (Adwait Ratnaparkhi, 1996). In itself, NLTK offers various tagged corpora such as Brown Corpus which is composed of 500 documents from 15 different domains, and Wall Street Journal Corpus (WSJ). Unfortunately, to the best of our knowledge, there are no (available) tagged corpora for policies in general and financial policies in specific. The lack of such annotated corpora does not impose any significant difficulties since NLTK provides its own default POS tagger. This default tagger relies on the Maximum Entropy statistical approach as presented in the work of (Adwait Ratnaparkhi, 1996). When using this module, NLTK utilises the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993) annotation corpus. The tags of this corpus are presented in Table 3. This statistical approach relies heavily on the context of the words which reflects in its ability to achieve 96.6% accuracy in tagging previously unseen text.

Abbreviation	Meaning	Abbreviation	Meaning
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal Number	RB	Adverb
DT	Determiner	RBR	Adverb Comparative
EX	Existential There	RBS	Adverb Superlative
FW	Foreign Word	RP	Particle
IN	Preposition or subordinating conjunction	SYM	Symbol
JJ	Adjective	TO	To
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund or present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd person singular present
NNP	Proper Noun, singular	VBZ	Verb, 3rd person singular present
NNPS	Proper Noun, plural	WDT	Wh-determiner
PDT	Predetermine	WP	Wh-pronoun
POS	Possessive ending	WP$	Possessive wh-pronoun
PRP	Personal Pronoun	WRB	Wh-adverb

TABLE 3- PENN TREEBANK FOR POS TAGS (ADAPTED FROM MARCUS ET AL., 1993)

How the tagger performed can be seen in Figure 6. , which depicts the transformation of a policy reference from a natural representation towards a *(word, POS- tag)* representation. These taggers often support the option of training and assessing them on a held-out set. Although such a step is not included in the framework, it is often suggested as a best-practice approach. Thus, to ensure consistency in tagging across document, this best practice was employed, where the tagger was trained on 90% of the words in the corpus. This tagger was then evaluated on the held-out set of 10% and showed an over 98% consistency in tagging words. Although this step is not included in the IE framework, it was seen as a preventive measure for having consistent tags across the corpus of documents, avoiding any bottlenecks in the future that may come as a result of improper tagging.



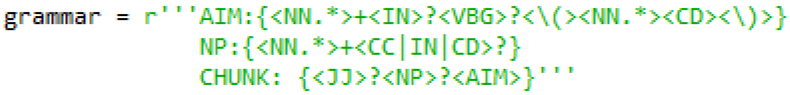
```
Business Segments Reporting Policy AIM 104-40-10

[('Business', 'NNS'), ('Segments', 'NNS'), ('Reporting', 'VBG'), ('Policy', 'NN'), ('AIM', 'NN'), ('104-40-10', 'CD')]
```

FIGURE 6- TRANSFORMATION OF POLICY REFERENCE

The process of mapping the words to their corresponding POS tag took the process a step closer towards completing the IE cycle. Through the collection of words and their respective POS tags, one can identify and group relevant entities together. This process is known as Named Entity Recognition (NER). In some literature, NER is referred to as shallow parser, which is a sentence analysis method that initially identifies the word tags and then it uses those tags to construct higher level entities which have separate grammatical meaning, such as noun phrases (NP), verb phrases (VP), proposition phrases (PP) and so on (Leroy, Chen, & Martinez, 2003). Nevertheless, in essence, both of these terminologies refer to the same practice in the IE pipeline. The significance of this phase is that it enables to separate the signal from the noise, thus identifying relevant data from the entire collection of text. The completeness of NLTK, for processing textual content, is also shown in this phase. NLTK provides a default NE parses which is capable of identifying entities such as institutions, dates, brands, individual names, countries and so forth. Judging from the entities that it is capable to identify, this default parser shows that it can perform properly when processing news articles. Nevertheless, when implemented against the corpus of policy documents, the parser did not yield promising results. This comes as a fact that such documents make use of a domain-specific language, which language may not be familiar to the Named Entity (NE) parser. Since the NLTK parser could not be used, this phase of the pipeline required some user insights in order to achieve appropriate results.

When dealing with textual documents, such as policies, the entity recognition phase requires the construction of custom NE parser. In order to properly construct these parsers, one should have a good understanding of the data that it is being processed and the format in which these relevant information is presented. The process of constructing the custom parsers generally starts with a manual investigation of the tagged corpus, to identify chunks of information that represent some form of interest for the study. This information can be a word or a collection of words adjacent to each other, which together provide some meaningful insight. Knowing the representation of this relevant information, they are grouped together in a single entity through a process known as chunking. Although the default NE parser was not appropriate for identifying relevant entities from the policies, NLTK still supports the user in the process of constructing custom chunkers for domain specific situations. The manual construction of these chunks is achieved through the chunking grammar. Chunking grammar is usually a set of rules, expressed in Regular Expressions, which specify the sequence of words that form an entity. An important attribute of such rules is that they look for POS tag of relevant words, and their order when defining the grammar. The sequence of tags that make up the rule is also known as *tag pattern*. Here we also see the importance of having consistent word tags across the policy corpus, since the chunking grammar highly depends on the tag pattern. In a summed up version, chunking starts with identifying the tag pattern of relevant words, it goes further by constructing the chunking grammar based on those tags and parsing the same grammar on the annotated textual content. The results of parsing a chunk can be either as a tag representation or a tree representation. Nevertheless, this project looks only into the tree representation of chunks, since the use of Regular Expressions only yields a tree representation of the chunking grammar.

A better understanding of all of this and how it has been utilised in the project at hand can be attained with the help of Figure 7. Having in mind the sentence transformation that took place earlier, where a sentence was transformed into a tuple representation, the given figure shows how these tags have been used to form a chunking representation for parsing the textual content. In order to capture and extract the policy name together with its AIM code (ID), we needed to construct chunks that group these words together based on their POS tag. From our visual inspection, we saw that the policy names were usually composed of a collection of nouns (NN) or variations of nouns (i.e. NNP, NNS,NNPS), adjectives (JJ) and verbs (VGB), this also introducing an NE in itself. The other visible entity was the one presented within the brackets, which holds the code (ID) of the respective policy. This entity was a combination of a noun (NN), or its variations, and cardinal number (CD). This information were used in constructing the following NE, which was referred to as ***AIM***. Next to this, some references held tags other than the ones defined in the first NE. In order to account for them, a new NE was constructed (***NP)*** which encapsulated their representation. Both of this entities, were further combined under a single entity that was titled ***CHUNK.*** Parsing this grammar on the tagged textual content returns a graphical representation of the relevant information that can be seen in Figure 8. This figure shows in detail how the relevant information are chunked together depending on the grammar rules that were provided. Thus, initially, this figure shows the natural representation of the information, followed by the chunking of ***NP*** entity and ***AIM*** entity subsequently. Even though in most of the cases, the references followed a uniform representation, there were also cases where a given reference did not adhere to this uniformity. This differences were accounted for when constructing the ***CHUNK*** entity, making it an overarching grammar for the entire corpus.



```
grammar = r'''AIM:{<NN.*>+<IN>?<VBG>?<\(><NN.*><CD><\)>}
              NP:{<NN.*>+<CC|IN|CD>?}
              CHUNK: {<JJ>?<NP>?<AIM>}'''
```

FIGURE 7- CHUNKING GRAMMAR

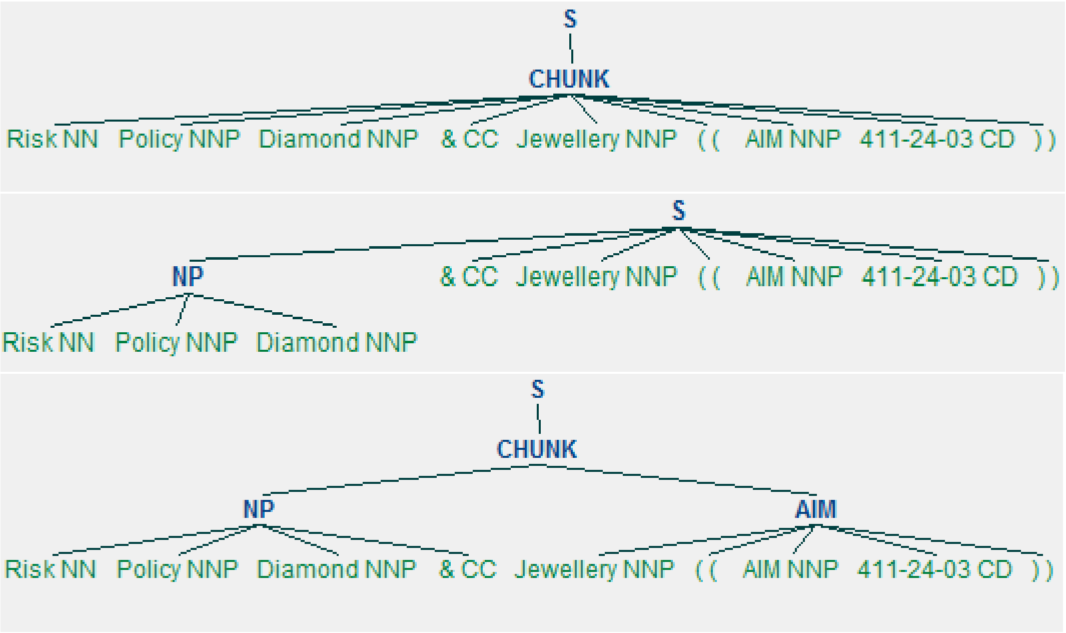


FIGURE 8- DETAILED CHUNKING PROCEDURE

4.3.2 AUTOMATIC SUMMARIZATION

The enormous pool of unstructured data makes it quite challenging to identify and extract valuable information in a timely fashion. A factor in this is also the large volume of unstructured data, which do not follow a uniform approach of representing valuable insights. Because of this, in a number of cases, the compression of the data volume has been seen as a viable solution for such problematics. Creating a succinct representation of the documents, that encapsulate its main content, can have a significant impact on the time and effort required to extract relevant information from the same. Nevertheless, it should be considered that the manual creation of such concise representations is a tedious and time consuming task as well. An answer to such problems can be found in the variety of TM techniques, more precisely, in the automatic summarization technique. As the name implies, this technique can be used to automatically generate a concise representation of the unstructured data, hence having an impact in their volume.

This technique can be beneficial when having a corpus of text-heavy documents that in themselves hold insights that are relevant to the business. The ability to create an overarching representation of the textual documents in a timely fashion is what has made automatic summarization one of the most utilised techniques. This can also be seen from the literature, where automatic summarization has found use in different domains such as biomedicine, law, microblogs, news articles and much more. Its extensive use has contributed in the creation of various summarization algorithms, which Larson (2012) grouped into two classes, namely extractive summarization algorithms and abstractive summarization algorithms. These algorithms differ in the approach that they employ for constructing summary representations. Extractive algorithms construct the summaries by using the most important sentences of the textual document and concatenating them into a consisted summary. Whereas the abstractive summaries, may not always draw on the same concepts as the ones that the original text contains. It usually reuses the main phrases of the document and constructs them in a manner that would convey the message. In addition to this, the author further categorized the algorithms, based on their appliance. This criterion revealed two additional categories, namely multi-document summarization algorithm and single-document summarization algorithm. As the name implies, multi-document summarization algorithm generates an overarching summary based on the content of all the documents in the corpus. This can be thought as a summary of the entire corpus. Whereas, single-document summarization algorithm, generates an individual summary for each document. The summary generated from such an algorithm is a direct output of the textual content of the respective document. A better understanding for some of the most notable summarization algorithm can be obtained by the work of (Inouye & Kalita, 2011). They authors provide a basic introduction and understanding to summarization algorithms such as SumBasic (Vanderwende, Suzuki, Brockett, & Nenkova, 2007) a multi-document extractive algorithm, LexRank (Erkan & Radev, 2004) a multi-document extractive algorithm, TextRank (Mihalcea & Tarau, 2004) single-document extractive algorithm, and MEAD (Radev et al., 2004) a multi-document extractive algorithm.

Considering, the lack of scientific literature that investigates the use of TM on policy documents, complimented with the inexistence of policy summarization literature, gave us no ground to base the algorithm selection for the study at hand. Nevertheless, the types of text summarization approach presented by Larson (2012), together with the list of most notable algorithms mentioned earlier, provided some insights on what the decision should rely on. This was also complimented by the literature regarding the use of TM on legal text, where automatic summarization appeared to be one of the most utilized techniques. Considering the fact that the ultimate goal of this research is to create a repository that contains some descriptive information for each policy, it was obvious that the chosen algorithm should be a single-document summarization algorithm. Additionally, abstractive extraction algorithm was seen as inappropriate for this study given the fact that policies are composed of domain specific text and the automatic abstraction of the text may change the context of the document. Thus, a single document extractive algorithm was seen as most suitable to this research. This condition immediately excluded SumBasic, MEAD and LexRank algorithms, leaving TextRank as the only extractive algorithm that suits the purpose.

TextRank algorithm has been introduced more than a decade ago as a graph-based extractive algorithm (Mihalcea & Tarau, 2004). This algorithm, has its roots embedded in the PageRank algorithm (Page & Brin, 1997), and it utilizes the same logic. It is worth mentioning, that this algorithm laid the foundations for creating Google. PageRank is a graph-based ranking algorithm that ranks web pages by determining the importance of its vertex in a graph, which graph contains computed vertexes of global web pages. This means that the importance of a web page is not determined by analysing its individual content, but by comparing it to the content of other global websites. Similar logic is used by Mihalcea and Tarau (2004) in developing TextRank. TextRank uses the knowledge drawn from the entire text to construct a graph, upon which graph the PageRank formula is applied in order to determine the most important vertices. This knowledge is acquired by the “voting” or “recommendation” concept, where every time a vertex is linked to another peer, it casts a vote for that vertex. Furthermore, the weight of each vote is also dependent on the importance of the vertex that casts it. Thus the more votes a vertex has, the more important it is.

Adhering to the same methodology as presented in the work of (Mihalcea & Tarau, 2004), all the documents available in the corpus were retrieved individually through the algorithm. From here, the content of the document underwent a process of segregation. With the use of the default NLTK sentence segmenter, the policy content was split into separate sentences. Such segmentation enabled the construction of the graph representation of the document, where each sentence represented a vertex (node) in the graph. Furthermore, to complete the graph, these vertices needed to create a link among each other, which is also known as edge. Mihalcea and Tarau (2004) advice that the edges of the graph should be created based on the "recommendation" concept between the vertices. The recommendation concept is built based on the assumption that a given sentence, will recommend another sentence to read, based on their resemblance. The content similarity has been widely used in developing systems such as plagiarism detection, computing language and dialect differences and so on. These systems have utilised similarity measurement formulas such as cosine similarity (Mihalcea et al., 2006), string kernels (Lodhi, Saunders, Shawe-Taylor, Cristianini, & Watkins, 2002), Jaccard Similarity Coefficient (Achananuparp, Hu, & Shen, 2008) etc. Nevertheless, to compute the similarity between vertices in this study, the Levenshtein Distance (LD) of (Levenshtein, 1966) was used. LD is also known as edit distance, since it measures the similarity between entities based on the number of insertion, substitution or deletion that one entity requires, in order to become the same as the entity that it is compared to. The literature showed that LD has been applied in various scenarios and has proven to be a suitable method for measuring linguistic similarities. Examples of its applicability can be found in the work of Soukoreff and MacKenzie (2001), who used the LD to contract a new technique for measuring the error rate in text entry. Additionally, Spruit (2006) and Heeringa (2004) used LD in measuring the syntactic variation of different Dutch dialects. Other examples can be found in the work of Renz, Ficzay and Hitzler (2003), who saw LD as a better keyword extraction algorithm than *tf-idf*, whereas R. C. Wilson and Hancock (2004) found LD as a suitable method for extracting graph represented features. With regards to this study, LD was computed for all the sentences in the policy text and its score was added as an edge between the two vertices in the graph, resulting in a full graph representation of the policy text. With this graph in place, the PageRank formula was applied to its vertices, whose equation is as follows:

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out\ (V_j)|} S(V_j)$$

EQUATION 2: PAGERANK CALCULATION

For a given vertex V_i, $In(V_i)$ is the set of vertices that point to it, also known as predecessor vertices. On the other hand, $Out\ (V_j)$ is the set of vertices to whom the V_i, vertex points to, which are also known as successor vertices. Whereas, d is defined as a dumping factor which may take a value between 0 and 1. Nevertheless, Page and Brin (1997) argue that this dumping factor can be set to a constant value of 0.85. The result from this algorithm is a dictionary of nodes with the PageRank result as their value. When extracting summaries, there is not much pre-processing that can be done, except for segmenting the sentences. This is mainly because of the fact that if entities are removed from the text, once the computations take place, there is a high risk that the extracted sentences for the summary may not provide meaningful representation. On the other hand, since there is no pre-processing, the size of the graph results to be enormous. All the segmented sentences are added as a vertex in the graph and their edges are computed. From this extensive graph, it is important to be able to extract the most important sentences, namely sentences that score highest after the PageRank calculation. In order to do so, the algorithm is tuned to return the vertices in a descending order, putting the highest scoring vertices on top. Depending on the cut-off value, this structuring enables to always have the most important vertices in the summary. With regards to this study, a cut-off value of 250 words was set for summaries. Such a value was decided in accordance with the stakeholders, after some trial and error attempts, where longer representations were concerned. Furthermore, such a cut-off makes a consistent representation of the summaries across the corpus.

4.3.3 AUTOMATIC KEYWORD EXTRACTION

Keywords have shown to be appropriate attributes for searching relevant information and at the same time for indexing documents in a repository. This is also the motivation why keywords were chosen for this study. It is thought that these keywords can be used as a form of tagging system for document, which tags would subsequently comprise a filtering mechanism, contributing to the time and efficiency of retrieving relevant documents. Furthermore, the benefits of using keywords for indexing and querying repositories, are not novel to this studies. The literature shows that such a mechanism has been used extensively over the years, with multiple TM techniques and algorithms developed for this purpose. Such an approach is also presented in the paper of (Mihalcea & Tarau, 2004). Besides depicting it as an automatic summarization algorithm, the authors portray TextRank as an algorithm that can extract keywords as well. They proclaim that by adhering to the same logic as in automatic summarization, but processing in word level rather than sentence level, TextRank can extract relevant keywords from the textual document. These capabilities make TextRank even more appropriate for the case study at hand. Using the same line of logic, but at the same time making the necessary changes as recommended by the authors, the TextRank algorithm was tuned to extract keywords. Thus, starting from what was updated in the algorithm, the sentences were further tokenized into words. This enabled the construction of a graph that carries out the analysis at a word level, where each word represented single vertex. Nevertheless, having all of the words as individual vertices, impacts the density of the graph. To compress its density, some vertex filtering mechanism was implemented.

If during automatic summarization we were reluctant towards the idea of removing unnecessary or irrelevant words, this was not the case in keyword extraction. When dealing with full sentences, removing a word exposes you to the risk of changing or losing the context of the sentence, thus ending up with an incomprehensible summary. This is not the case when dealing with words as entities on their own. To the contrary, removing irrelevant words from the graph increases the chance of extracting relevant keywords. Thus, initially, a stop-word filter was applied to the content of the document prior constructing the graph. Stop words are frequent appearing words that assist in creating an idea, when used in a sentence, but do not represent any significant meaning when alone (Rajaraman & Ullman, 2011). Such words may be “the”, “and”, “or”, “that”, “this” and many other words similar to these. In addition, a filtering option was applied based on the syntactic representation of the words. What this filter does, is that it determines valuable and invaluable words based on their POS representation. Similar as with IE, initially the words were tokenized and annotated with their POS tag. To determine which word should be excluded we manually investigated the POS representation of potential keywords. This investigation showed that the majority of potential keywords appeared in the form of nouns (NN), proper nouns (NNP) and adjectives (JJ). Thus the words that do not belong to this group of tags, were excluded from the list of potential keywords. This type of pre-processing is also advocated by the developers of TextRank. Furthermore, another step that has an impact on the density of the graph, is the removal of duplicate values. It is highly possible that the same set of words will appear throughout the text. Because of this, a third and final filtering was implemented, to account for the duplicate values. While scanning the list of entities, if a given word appeared for the first time, it was added to the final dictionary of words. Any subsequent appearance of the same word, or any other word that was already in the unique value dictionary, was ignored. This set of unique words was subsequently used to construct the graph. An example of how the graph looks, can be seen in Figure 9, where 15 words out of Policy A are depicted. It goes without saying that the graph construction follows the same logic as in automatic summarization. Initially, all the words in the dictionary were added as vertices in the graph. From here, the similarity between the vertices was computed with the help of LD. The outcome of this computation is what created the edge between the vertices, thus enabling the execution of the PageRank algorithm. Furthermore, this algorithm determined the importance of each vertex, with respect to the entire graph. Again, similar to the automatic summarization case, the results were reversed in a descending order, thus having the highest scoring vertices appearing first.

Generally speaking, the content of the policies varied from one document to the other, with some policies having more textual content and some less. This created the assumption that the volume of the content also has an impact on a number of keywords. To put it in other words, it is thought that the more content a document has, the more number of keywords it may have. To account for such a factor, many studies in the literature advise that the number of keywords should be relative to the content volume. Most of the studies shared a similar approach regarding this issue, where they advised to extract a number of keywords that is one-third (1/3) of the entire document content (Mihalcea & Tarau, 2004). This approach mitigates the risk of ending up with an irrelevant list of keywords for documents with less content. Nevertheless, looking at the policies in our corpus, and their volume, it was recognised that adhering to such a ratio will yield an exhausting list of keywords. Even after all the filtering, the graph still contained several thousand vertices. Given such quantities, a better solution was seen if only 1% of the entire list would be considered as potential keywords. Thus given the list of potential keywords, only 1% of the highest scoring entities were selected as the final collection of keywords. Such a ratio resulted in a range of 8 up to 51 keywords per document.

Additionally, up to this point, the list of potential keywords consisted only of single word entities. Nevertheless, Mihalcea and Tarau (2004) argue that it is a good practice to also construct key-phrases from the final list of keywords. Such a practice was also implemented in this study, where given the unique entities, a set of key-phrases were constructed. This was done by matching each entity with one-another. This matching created key-phrases that were composed of two unique keywords. Furthermore, the set of mutated key-phrases was compared against the entire content of the document. This approach provided a form of validation method to determine whether such a key-phrase, indeed represents a potential keyword and it appears in the respective document. Hence, the constructed key-phrases were only a combination of two potential keywords and not more.

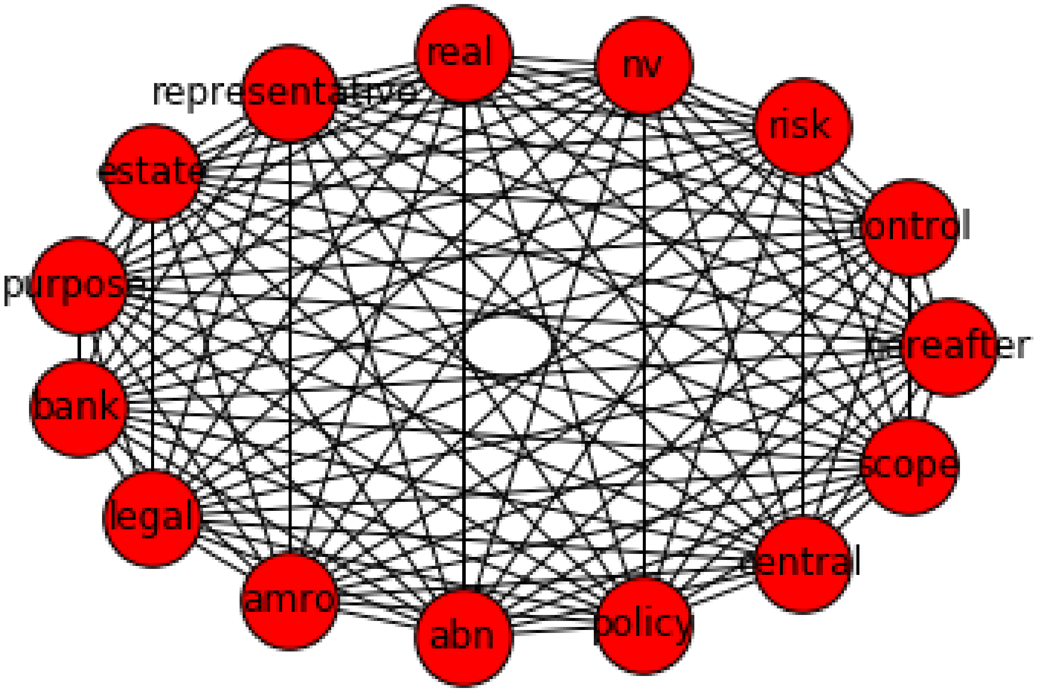


FIGURE 9 - GRAPH REPRESENTATION OF WORDS IN POLICY A

5. RESULTS

It is quite unfortunate that the literature, regarding the use of TM on policy documents, did not provide sufficient references to determine which TM techniques should be used for the CS at hand. Nevertheless, consulting the literature in other domains, and especially the literature regarding legal text, assisted in better understanding the requirements and it subsequently pinpointed towards the relevant techniques. From the collection of numerous established techniques and their implementation, it was more obvious how these techniques fit in this study and furthermore how should they be implemented. Thus upon selecting, constructing and implementing the necessary artefacts, a collection of results were retrieved from each executed module. These results were a direct output of all the statistical and linguistic computations that the textual content underwent. With all of these in place, the generated output was readily available for evaluation. Nevertheless, when it comes to evaluation it is highly important to emphasise the nature of the textual documents. Usually, the content of textual documents depend on the topic it treats and the purpose it is created for. Given such dependencies, a textual document may be designed for the understanding of the general public or it may be designed for the understanding of individuals with specific domain knowledge. The difference between such types of documents can be found in the concepts they use or treat. Thus, when analysing documents that are specific to a domain, one should have extensive knowledge of the concepts that the domain treats in order to be able to properly validate the results. Looking back to the corpus of policies used in this study, it was obvious that such documents were consolidated to exploit issues that concern financial institutions, more precisely banks. Furthermore, the corpus of policies revolved around issues that concerned only the Credit Risk aspect of the banks. Thus, for evaluating the results of this study, a good understanding of both banking industry and Credit Risk domain, was necessary. More specifically, such a knowledge was highly important in evaluating the outcomes of automatic summarization and keyword extraction.

Henceforth, upon having the algorithms designed and executed, their results were automatically stored in separate files. Regarding the extracted summary and keywords, both of these outputs were combined into a single file and were made available to the competent entities for evaluation. The evaluators were asked to review all the results individually and to deem them either as relevant or irrelevant, with respect to the document they belong. In addition to this, concerning the list of keywords, Policy Writers had the chance to also add potential keywords, which based on their perception are relevant but have failed to be recognised by the algorithm as such. On the other hand, the summaries followed more of a Black-Box evaluation approach, where each summary was evaluated based on its level of descriptiveness, with respect to the entire content. The evaluation form that was made available for the Policy Writers, can be found in Appendix A. On the other hand, the reference extraction experiment in the study did not require domain or expert knowledge for evaluation. Given the fact that this experiment concerned parts of the policy that were recognisable for the ordinary reader, its evaluation followed a different approach, which did not require them to be included in the evaluation form. Considering the small corpus of documents, initially, all the referencing policies were manually extracted from each document and were saved in a single file. Such a list provided an overview of the references in the entire corpus and at the same time the references per specific policy. This list forms of *golden source* standard, against which standard the performance of the algorithm was evaluated.

This evaluation form makes it possible to measure the accuracy of the algorithm in extracting necessary information. A common practice in literature, for determining the accuracy of the algorithm regarding the keyword extraction and IE, is the use of the *F-measure* (F_1) formula. This formula represents the harmonic mean of *precision* (P) and *recall* (R), which measurements themselves determine the success rate of the algorithm (Sasaki, 2007). *Precision* and *recall* have been regularly used to measure the performance of information retrieval and information extraction systems (Makhoul & Kubala, 1999). Based on what the algorithm returns, precision (also known as *confidence*) determines how many of the retrieved values are indeed correctly predicted. Whereas, recall (also known as *sensitivity*), determines how many values from the gold standard are correctly predicted by the algorithm. A better understanding of such measurements can be gained with the help of Confusion Matrix in Table 4. This matrix categorises results based on their actual value and the value that the algorithm assigns to them. Thus precision is the fraction of truly predicted values from the eternity of values predicted by the algorithm. Whereas recall represents the fraction of true values predicted by algorithm compared to the entire collection of true values. Although these two measures provide a good understanding of how the algorithm performs and the accuracy of its results, the scientific community felt the need to have a single measure of performance. This led to the introduction of the *F-measure* formula, which is an equally weighting formula of *P* and *R* (Makhoul & Kubala, 1999). A better understanding of *P, R* and *F* formula can be acquired from Equation 3, Equation 4 and Equation 5, respectively. It is worth mentioning that these measurements were used to evaluate the performance of the algorithm for reference extraction and automatic keyword extraction (which is explained later on). To the best of our knowledge, such measurements have not previously been used for evaluating automatic summarization. That is why, as mentioned earlier, the evaluation of the summaries followed the acceptance testing method that was presented earlier (Chapter 2).

		Classified as:	
		True	False
Really is:	True	TP	FN
	False	FP	TN

TABLE 4 - CONFUSION MATRIX

$$precision = \frac{TP}{TP + FP}$$

EQUATION 3: PRECISION

$$recall = \frac{TP}{TP + FN}$$

EQUATION 4: RECALL

$$F = \frac{(2 * precision * recall)}{precision + recall}$$

EQUATION 5: F-MEASURE

5.1 RESULTS FROM REFERENCE EXTRACTION

In order to compute the performance of the algorithm, the algorithmic outcome together with the entire collection of true values was needed. This called for the creation of a golden standard, which holds the entire collection of true values, against which standard the algorithmic outcomes were benchmarked. The small corpus of documents made the creation of such a standard possible, where from each document, all of the references were manually extracted. This standard, besides containing the entire list of true values, it also categorised them based on the document they belonged, enabling to compute the accuracy of the algorithm for each document separately. The so-called golden standard can be found in Appendix B.

Having such a golden standard together with the algorithmic outcome, some simple statistics showed that from the total number of 435 references, the algorithm failed to entirely recognise only 23 of them. This is roughly 5% of the total number of the reference. Additionally, there were cases that the algorithm partially recognised the references. Such was the case for only 44 references, which references were not fully chunked, thus missing a part of the policy title. Investigating the reason why the algorithm failed to properly chunk such entities, it was revealed that most of these entities contained a word in the Dutch language. This language was not accounted for in the algorithm, thus making the module inadequate to recognise foreign words. Such words caused the algorithm to either fail to recognise the reference entirely or to recognise the reference partially. In an attempt to determine what is the proper scientific way of dealing with such imperfect chunks, the literature revealed that the concept of mistake/negative (i.e. partially correct) is not properly defined and it is a subject of change depending on researchers' perception (Hripcsak & Rothschild, 2005). Thus, it was a matter of discussion whether the partially correct references should be denoted as correct or incorrect. A consensus was reached by deciding to follow two evaluation approaches. The initial evaluation approach (E1) had a low tolerance level and it considered the partially correct references as entirely incorrect. If the chunked reference missed a single word, then the entire chunk was discarded and deemed as incorrect. The results of E1 are given in Table 5. This table provides a detailed overview of each policy, like the total number of references it contains, number of references that the algorithm was able to correctly recognize, number of references that the algorithm partially recognized, number of references that the algorithm entirely missed and number of chunks that the algorithm recognized in the document, which chunks were not references. In essence, both of the calculations rely on the same table of values, the difference here was that the second approach (E2) considered the partially correct chunks as entirely correct, thus being less rigor. As such, Table 6, only reports on *P*, *R* and *F* of E2. A better understanding of the difference between these two evaluation forms can be acquired from the confusion matrix. While in both of the cases, the correctly parsed chunks were denoted as *True Positive (TP)* values and the extra chunks were denoted as *False Positive (FP)* values, the difference occurred in the *False Negative (FN)* values. E1 counted both, the partially correct references and the missed references as *FN*, whereas in E2 the partially correct references were considered as *TP*, leaving only the entirely missed references as *FN* values. Such a difference had a subsequent impact on the computation of *P, R* and *F.* Thus while Table 5 provides a full overview of the experiment results, Table 6 only shows the values that have been altered as a result of change in evaluation methodology. Additionally, researchers such as (Powers, 2011), argue that in IE experiments related to Machine Learning and Computational Linguistic, more importance should be put on determining how confident one can be with the rules or classifier. This means that precision should be weighted more than recall when assessing the accuracy. Considering that the extraction of the policy references was also a form of IE, this aspect was also considered in the study. For such cases, a variation of the F formula is used, namely the F_β formula which is shown in Equation 6. The β value determines which measurement is weighted more, where the most famous values that it can take are 0.5 and 2. The former gives more weight to P, whereas the latter to R. Thus next to the normal weighted F value (F_1), the tables also provide insights on what the average F score was when more weight was given to P with the use of $F_{0.5}$ formula. The full calculations, for all the relevant metrics, are provided in Appendix C. These evaluation showed that the IE approach followed in this study managed to yield quite impressive results. In the E1 approach, the algorithm managed to reach an average extraction accuracy of 89% with equally weighted P and R, and a 94% accuracy, when P was weighted more. Additionally, when being less strict with the imperfect chunks, the algorithm resulted in a 95% accuracy with normal weighted metrics, and 97% accuracy when $F_{0.5}$ was measured. Overall this showed that the devised artifact was highly capable to recognize the relevant information in the corpus and extract them with a high success rate.

$$F_\beta = (1+\ \beta^2)\ \cdot\ \frac{precision\ \cdot\ recall}{(\beta^2\ \cdot precision) + recall}$$

EQUATION 6: F_β -MEASURE

Title	Total	Correct	Part.Corr.	Miss	Extra	P	R	F	F_β
Policy A	15	11	3	1	0	1	0,73	0,84	0,93
Policy B	15	15	0	0	0	1	1	1	1
Policy C	13	8	4	1	0	1	0,61	0,76	0,88
Policy D	11	4	1	6	0	1	0,36	0,53	0,73
Policy E	13	12	1	0	2	0,85	0,92	0,88	0,86
Policy F	18	13	4	1	0	1	0,72	0,83	0,92
Policy G	14	12	2	0	1	0,92	0,85	0,88	0,90
Policy H	11	11	0	0	0	1	1	1	1
Policy I	27	21	5	1	0	1	0,77	0,87	0,94
Policy J	31	26	4	1	0	1	0,83	0,9	0,96
Policy K	6	6	0	0	0	1	1	1	1
Policy L	20	14	6	0	0	1	0,7	0,82	0,92
Policy M	13	11	2	0	0	1	0,84	0,91	0,96

Policy N	13	13	0	0	0	1	1	1	1
Policy O	24	22	2	0	0	1	0,91	0,95	0,98
Policy P	17	14	0	3	0	1	0,82	0,90	0,95
Policy Q	8	8	0	0	0	1	1	1	1
Policy R	10	10	0	0	0	1	1	1	1
Policy S	11	10	1	0	0	1	0,90	0,94	0,97
Policy T	77	74	1	2	1	0,98	0,96	0,97	0,97
Policy U	29	21	6	2	0	1	0,72	0,83	0,92
Policy V	22	16	2	4	0	1	0,72	0,84	0,92
Policy W	17	16	0	1	0	1	0,94	0,96	0,98
Total	**435**	**368**	**44**	**23**	**4**	**22,75**	**19,3**	**20,61**	**21,63**
Average	**18,91**	**16**	**1,91**	**1**	**0,17**	**0,99**	**0,84**	**0.89**	**0,94**

TABLE 5 - RESULTS FROM E1.

Title	P	R	F	F_β
Policy A	1	0,73	0,96	0,98
Policy B	1	1	1	1
Policy C	1	0,92	0,96	0,98
Policy D	1	0,45	0,62	0,80
Policy E	0,86	1	0,93	0,88
Policy F	1	0,94	0,97	0,98
Policy G	0,93	1	0,96	0,94
Policy H	1	1	1	1
Policy I	1	0,96	0,98	0,99
Policy J	1	0,96	0,98	0,99
Policy K	1	1	1	1
Policy L	1	1	1	1
Policy M	1	1	1	1
Policy N	1	1	1	1
Policy O	1	1	1	1
Policy P	1	0,82	0,90	0,95
Policy Q	1	1	1	1
Policy R	1	1	1	1
Policy S	1	1	1	1
Policy T	0,98	0,97	0,97	0,97
Policy U	1	0,93	0,96	0,98
Policy V	1	0,81	0,90	0,95
Policy W	1	1	0,94	0,98
Average	**0,99**	**0,93**	**0,95**	**0,97**

TABLE 6 – RESULTS FROM E2.

5.2 RESULTS FROM SUMMARY EXTRACTION

The extracted summaries, from the algorithm, were a concise mirroring of the entire policy content. This means that the summaries were generated by using exact sentences as in the policy itself with no additional mutations. Furthermore, such summaries only included the most important vertices of the entire graphical representation of the content. Given that in most of the cases there is no golden standard against which the summaries can be evaluated, such a technique required a different form of evaluation. In the case of reference extraction, it was possible to transform the qualitative outcome into quantitative annotations. This transformation made it possible to compute the accuracy of the algorithm with the use of the *P, R* and *F*. Unfortunately, this was not the case with the automatic summaries since the content of the summary could not be encoded in quantitative values. Furthermore, given the domain specific dictionary that has been used to compile these documents, the evaluation of the summaries required an equal knowledge for evaluating them. Henceforth, domain experts were charged with the responsibility to evaluate the algorithmic summaries. The experts were asked to assess the summary and to comment on issues such as whether the summary covers the main aspects, whether the summary is a good representation of the policy, what does the summary miss and other comments of the similar nature. As mentioned earlier, such form of evaluation follows a somewhat Black-Box evaluation approach where a comment was provided regarding the accuracy of the outcome. Through an evaluation form, the experts could read the generated summary for the respective policy, and assess its competences by providing a comment about its ability to convey the main message of the document. Furthermore, these comments were denoted in the test plan that was introduced earlier. An example of such a test plan together with an expert evaluation comment is given in Table 7, which table beside providing descriptive details also cites the expert comment which is in harmony with the outcome of the test. Nevertheless, most of the fields in the test plan were consistent throughout the entire collection of tests, differing only in the order and the policy they were designated for. The only distinct feature about these tests is under the outcome field, which corresponds to the expert's comments. Thus Appendix D provides the evaluation form where expert's comments can be found. Bear in mind that the actual summaries consisted sensitive data, which data could not be disclosed because of a confidentiality agreement. Nevertheless, in order to create an idea how these summaries might have looked, Policy A under the appendix, provides the summary of a publicly available policy. The algorithm was able to properly construct extractive summaries, nevertheless, the expert evaluation showed that not always these summaries were adequate. The evaluation disclosed that the extracted summaries had some data quality issues. Such an issue was the fact that in some cases the algorithm generated a more detailed summary than necessary. Looking at the expert's evaluation, it was noticed that the comments followed a somewhat similar pattern across the documents. Thus, these comments were summed up and they revealed three evaluation themes, namely:

1. *"Thu summary covers the main aspect of the policy."*
2. *"The summary is a moderate representation of the policy*."
3. *"The summary is too detailed, thus does not cover the main aspect of the policy.*"

These comments were more or less equally dispersed across the corpus. From the collection of 23 Credit Risk policies, 30.4% of the summaries were found to be a good representation of the policy, thus covering the main aspects of the document. Such a comment was found in 7 evaluation forms. Furthermore, in 34.7% of the cases, the experts evaluated the summaries as a moderate representation of the document. Such comments were found in 8 evaluation forms. The same amount of documents received the comment that their summaries is too detailed and thus does not cover the main aspect of the policy. This comment counts for the remaining 34.7% of the corpus. What this evaluation showed, is that the algorithm was perfectly capable to generate extractive summaries from the text, nevertheless in a number of cases it showed the tendency to retrieve vertices that represented data quality issues when evaluated. Hence, not being a fully appropriate representation of the respective document.

Test No.	Test Type	Test Name	Purpose of Test	Test Date	Expected Results	Outcome
1.	Algorithm	Summary extraction for Policy A.	To determine whether the extracted summary is relevant to the content.	12/05/2017	The algorithm should return a collection of sentences that provide a concise summary of the policy.	The algorithm returns a summary that covers the main aspects of the policy

TABLE 7 - EXAMPLE OF POLICY A SUMMARY EVALUATION

5.3 RESULTS FROM KEYWORD EXTRACTION

The keyword extraction and automatic summarization relied on the same computation logic. As such, the list of potential keywords, similar to the summaries, was a direct output of the words with the highest score after the PageRank computation. Additionally, these words were also a cluster of domain and document specific phrases, meaning that they required some domain expertise to evaluate their relevance. Thus together with the summaries, the expert was requested to also evaluate the extracted keywords. Such an evaluation form gave the opportunity to the expert to denote the potential keywords as either relevant or irrelevant. In addition to these two options, the expert also had the chance to add keywords/key-phrases that they saw as relevant but were not included in the output list. Although in essence, the evaluation method of summaries and keywords was the same, in the case of keywords, this method enabled to translate the qualitative evaluation into quantitative representations. Such an evaluation provided three classes of keywords, namely relevant keywords extracted by the algorithm, irrelevant keywords extracted by the algorithm and relevant keywords *not* extracted by the algorithm. Such classes gave the necessary parameters needed to calculate the accuracy of the algorithm. Thus, having the confusion matrix in mind, the correctly predicted keywords represented the *TP* values*,* the incorrectly predicted keywords represented the *FP* values*,* and finally, the keywords suggested by the experts represented the *FN* values*.* Encoding the keyword evaluation in such a way enabled to calculate the *P, R* and *F* for each document individually. Such an evaluation was conducted on the list of keywords that derived from the algorithm. Furthermore, Table 8 provides a breakdown of all the policies together with the correct, incorrect and suggested keywords from the expert evaluation, as well as *P*, *R* and *F* for each document. These evaluation showed that when it came to keyword extraction, the constructed algorithm reached an 83% extraction accuracy across the whole corpus. Furthermore, from an average of 19 keywords per document, the module predicted on average 15 correct keywords per document, resulting on an average incorrect value of 3,6 per policy. The full calculation description of these metrics is given in in Appendix E.

Name	Total	Correct	Incorrect	Suggested	P	R	F
Policy A	8	4	4	2	0,50	0,67	0,57
Policy B	13	8	5	0	0,62	1	0,76
Policy C	20	18	2	3	0,90	0,86	0,88
Policy D	29	23	6	2	0,79	0,92	0,85
Policy E	16	12	4	7	0,75	0,63	0,69
Policy F	14	10	4	2	0,71	0,83	0,77
Policy G	10	8	2	1	0,80	0,89	0,84
Policy H	19	11	8	0	0,58	1	0,73
Policy I	29	24	5	3	0,83	0,89	0,86
Policy J	51	43	8	10	0,84	0,81	0,83
Policy K	14	10	4	6	0,71	0,63	0,67
Policy L	13	10	3	2	0,77	0,83	0,80
Policy M	12	11	1	0	0,92	1	0,96
Policy N	13	11	2	0	0,85	1	0,92
Policy O	35	29	6	0	0,83	1	0,91
Policy P	13	12	1	0	0,92	1	0,96
Policy Q	10	7	3	0	0,70	1	0,82
Policy R	11	9	2	6	0,82	0,60	0,69
Policy S	17	16	1	0	0,94	1	0,97
Policy T	48	40	8	0	0,83	1	0,91
Policy U	13	12	1	0	0,92	1	0,96
Policy V	19	16	3	0	0,84	1	0,91
Policy W	11	9	2	0	0,82	1	0,90
Average	**19**	**15**	**3,6**	**1,91**	**0,79**	**0,89**	**0,83**

TABLE 8- RESULTS FROM KEYWORD EXTRACTION

As mentioned earlier, the list of extracted keywords corresponded with the entire content of the policies. Thus, the number of potential keywords was set to be 1% of the entire density of the graph. As it can be seen from the table above, applying such a ratio resulted in a dispersed number of keywords per policies, starting from 8 up to 51. Nevertheless, one may argue that this range is not realistic and it exceeds the reasonable amount of keywords per documents. Taking this into account, the focus shifted towards the literature and how similar situations were handled. Looking into publications that treated similar technique, it was revealed that studies such as (Rose, Engel, Cramer, & Cowley, 2010), (Yang, Chen, Cai, Huang, & Leung, 2016), (Hulth & Megyeesi, 2006) and (Mihalcea & Tarau, 2004), based their research on a range of keywords between 9 and 15 per document. Relying on these findings, it was decided that the performance of the algorithm should also be measured for such a range. Thus, every list of keywords that had more than 15 suggestions, was reduced to the underlying value. Henceforth, any positive or negative value beyond the given threshold was discarded. The expert suggestions were immune to these changes, thus they were kept the same as in the previous evaluation. Applying such a cut-off value meant that the list only contained the 15 highest ranking keywords. The scores of this experiment are provided in Table 9 and the calculations in Appendix E. In this experiment the average extraction accuracy resulted to be around 82%, which showed a relatively small difference from the initial evaluation. Furthermore, from an average number of 13,2 keywords per policy, roughly 10 of them resulted to be predicted correctly. This ratio did not differ much from the ratio that the original evaluation had between the total number of predicted keywords and the actual correct keywords. This indicated that even if the volume of keywords is reduced, it will not have a significant impact on the overall results. To get a better understanding of how the algorithm performed in these two cases, we benchmarked the generated outcome with other relevant studies. In this benchmarking attempt, we consider studies, such as the work of Rose et al. (2010) who used an unsupervised, domain-independent and language independent method to extract keywords from a corpus of 2000 abstracts of journal papers. Another, benchmarking study was the work of Mihalcea and Tarau (2004) who used the same graph representation and PageRank computation to extract keywords. Yang et al. (2016) used a machine learning algorithm to extract keywords from a corpus of 2000 abstracts, whereas Liu, Pennell and Liu (2009) evaluated the keyword extraction on a corpus of 27 meeting summaries. Hulth and Megyeesi (2006) evaluated their keyword extraction approach on a corpus containing nearly 20 000 news articles. And finally, Zhang et al. (2008) extract keywords from a corpus of 600 academic papers concerning the field of economics. Most of these studies reported on multiple experiments, providing a *P, R* and *F* for each of the experiments. Nevertheless, in our comparison, only the experiments with the highest F value were considered. Additionally, this comparison may not be fully scientific since they do not explicitly concern policy documents, nevertheless, it was seen as a suitable method for providing an indication on how similar studies have performed and where does our study stand compared to them. This benchmarking mechanism is presented in Table 10, which depicts the results that each of the aforementioned studies have achieved, together with the outcomes of this study. This benchmarking showed that the outcome of this experiment represent one of the highest accuracy levels that have been reached when extracting keywords.

Title	Total	Correct	Incorrect	Suggested	P	R	F
Policy A	8	4	4	2	0,50	0,67	0,57
Policy B	13	8	5	0	0,62	1	0,76
Policy C	15	11	2	3	0,85	0,79	0,81
Policy D	15	13	2	2	0,87	0,87	0,87
Policy E	15	12	3	7	0,80	0,63	0,71
Policy F	14	10	4	2	0,71	0,83	0,77
Policy G	10	8	2	1	0,80	0,89	0,84
Policy H	15	9	6	0	0,60	1	0,75
Policy I	15	12	3	3	0,80	0,80	0,80
Policy J	15	12	3	10	0,80	0,55	0,65
Policy K	14	10	4	6	0,71	0,63	0,67

Policy L	13	10	3	2	0,77	0,83	0,80
Policy M	12	11	1	0	0,92	1	0,96
Policy N	13	11	2	0	0,85	1	0,92
Policy O	15	15	0	0	1	1	1
Policy P	13	12	1	0	0,92	1	0,96
Policy Q	10	7	3	0	0,70	1	0,82
Policy R	11	9	2	6	0,82	0,60	0,69
Policy S	15	14	1	0	0,93	1	0,97
Policy T	15	12	3	0	0,80	1	0,89
Policy U	13	12	1	0	0,92	1	0,96
Policy V	15	12	3	0	0,80	1	0,89
Policy W	11	9	2	0	0,82	1	0,90
Average	**13,2**	**10,56**	**2,60**	**1,91**	**0,79**	**0,87**	**0,82**

TABLE 9- RESULTS FROM KEYWORD EXTRACTION OF MAX. 15 KEYWORDS

Study	**P**	**R**	**F**
Automatic Keyword extraction from individual documents (RAKE) – (Rose et al., 2010)	33,7%	41,5%	37,2%
TextRank: Bringing order into text – (Mihalcea & Tarau, 2004)	31,2%	43,1%	36,2%
Improved automatic keyword extraction given more linguistic knowledge – (Yang et al., 2016)	22,5%	51,7%	33,9%
Unsupervised approach for automatic keyword extraction using meeting transcripts – (Liu, Pennell & Liu, 2009)	Na	Na	19,6%
A study on automatically extracted keywords in Text Categorization – (Hulth & Megyeesi, 2006)	92,89%	72,94%	81,72%
Automatic keyword extraction from documents using conditional random fields – (Zhang et al., 2008)	66,3%	41,9%	51,25%
Our Study	**79,1%**	**89,3%**	**83,2%**
Out Study (max. 15)	**79,5%**	**87,3%**	**82,3%**

TABLE 10- STUDY BENCHMARKING

5.4 CASE STUDY IMPLICATIONS

What started as an attempt to determine the use of TM on business policies, more specifically bank policies, evolved into a novel approach of processing such text-heavy; domain-specific documents. The literature clearly indicated the capabilities of TM, to process and scrutinise unstructured data of different domains. Its prominence was mostly shown in domains such as biomedicine, national security, legal, fraud detection and so forth. The use of TM in these industries has moved from designing theoretical frameworks to the adoption and implementation of these frameworks into a working system. Nevertheless, overlooking at its use on policy documents, let it be privacy or business, the lack of proper scientifically validated frameworks and systems was noticeable. The limitations in this literature ranged from the disparity of TM techniques that they utilise, up to the validation of the frameworks. Additionally, much of the literature findings derived from the use of TM on privacy policies, and the implications that come with them, casting a shadow on the possibilities of processing business policies. Nevertheless, this literature deficit, unveils an opening for scientific contribution that can be fulfilled by the outcome of the study at hand.

Even though, the available publications on policy documents gave little to no indication on which techniques should be used for the case study requirements, the rich body of literature in other domains pinpointed techniques that would attain the study objectives. This literature showed that when dealing with unstructured data, the processing can rely on a single or a harmony of TM techniques, depending on the intended outcome. Such was the nature of the case study at hand, where the set of requirements called for the use of multiple TM techniques. Looking thoroughly at the adaption of these techniques into programmable artefacts and the execution of the artefact on the study corpus, the algorithmic results did not fail to impress. The ability of the algorithm to reach an over 89% accuracy in extracting relevant information, hints the potentials of such techniques to process policy documents. A distinctive feature of this artefact was its ability to miss a relatively small portion of relevant information from the corpus. On a similar note, the automatic keyword extraction technique proved to be highly capable of recognising and extracting the most distinctive words of a document. Although the summarization of the documents did not share the same level of success, it still managed to reach a moderate success, thus not being a complete failure. Furthermore, its tendency to generate more detailed summaries than necessary shows that the issues in these are approach were mostly data quality related rather than methodological.

The results of this study are the first of its kind when it comes to internal business policies. While most of the literature studies have yet to execute their conceptual frameworks and ratify their claims, the results of this study and their statistical evaluation validate the use of such an approach on business policies and make it distinguishable from the rest. Referring back to the gap in the literature that was mentioned earlier, the contributions that this framework offers for filling this gap are numerous. Starting with the fact that the given approach followed an unprecedented set of TM techniques when it comes to analysing policy documents in general. Additionally, it is one of the rare studies that has been fully implemented on business policies, more specifically bank policies. And finally, it is validated by experts and scientific methods. In addition to all of this, given the fact that it utilises three different TM techniques independently from each other, this framework offers two versions of itself. The first version of the framework, shown in Figure 10, depicts a complete overview of the followed approach. All the utilised techniques are combined into a single framework representation, which provides a step-by-step description of the followed approach. It guides from the data preparation phase towards actions that concern reference extraction, keyword extraction and automatic summarization. At the same time, such phases can be recognised in the framework as the main activities, each having a collection of sub-activities. In addition to these, the *"Graph"* activity is a shared entity among keyword extraction and automatic summarization. The latter version of the framework shows how the different TM techniques of the study can be used discretely. Given the fact that these techniques yield different results from each other, it may be the case that only one of them is relevant for a given study. Considering this, and the fact that they work independently from each other, they can be represented as a stand-alone framework. Each of these stand-alone representations follows on the major activities that concern the relevant technique, sharing only the data preparation activity as a commonality among them. A visual depiction of this framework version can be found in Appendix F.

Furthermore, the design of this frameworks draws upon the meta-algorithmic representation of the approach followed in the case study. M. Spruit and Jagesar (2016) define meta-algorithmic modelling (MAM) as an *"…engineering discipline where sequences of algorithm selection and configuration activities are specified deterministically for performing analytical task based on problem-specific data input characteristics and process preferences*". This approach relies on the activity recipes of the case study, which are modelled with the use of method –engineering notations. Method-engineering is defined as "*the engineering discipline to design construct and adapt methods, techniques and tools for the development of information systems*" (Brinkkemper, 1996). By following this approach, the framework is depicted as a Process Deliverable Diagram (PDD), which diagram uses UML activity diagram to represent the processes and UML class diagram to represent the deliverables (van de Weerd & Brinkkemper, 2008). In general activity diagram falls in the behavioural class of UML diagrams, thus being a representation that depicts the behaviour of a module. Whereas, class diagram belongs to the structural class of UML diagrams, thus representing the structure of the entities in the module. Nevertheless, this framework only represents the processes, thus composed of only UML activity diagram. The annotations of such a meta-algorithmic modelling approach together with the description of the activities and sub-activities can be found in Appendix G.

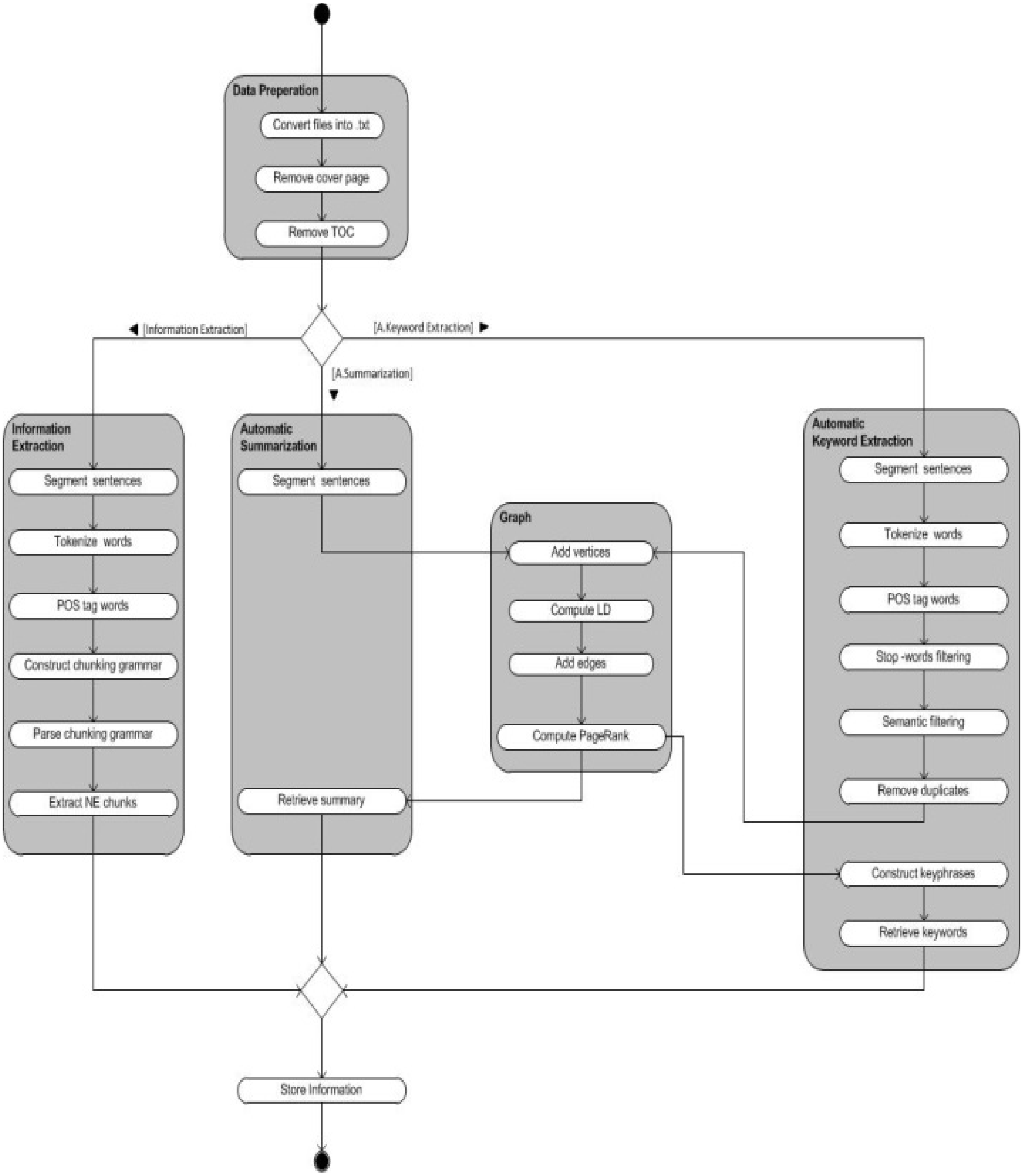


FIGURE 10 - META-ALGORITHMIC FRAMEWORK FOR PROCESSING BUSINESS POLICIES

6. CONCLUSION

Looking at TM as a well-established discipline, it is hard to find a domain where its capabilities are not exploited to its full potential. This comes as a result of the relatively high number of unstructured data that organisations create and collect in today's digital era. The need to extract valuable and relevant information from, these collection of data, has motivated scientist to introduce a wide range of TM techniques that rely on statistical and linguistic computation. An undeviating reflection of these claims is the scientific body of literature, which is a solid evidence that TM is applicable in nearly all industries. These convictions fostered the objective to contribute to the literature, either in a form of scientific literature review or through a novel artefact, where a particular attention was paid to the use of TM in the financial industry, more specifically bank policies. To institutionalise such objectives, the voyage started by examining the literature. This was accompanied by the set of RQ, where most of the answers relied on a full traversing of the literature. Using the snowballing method to answer the RQ, two spectrums of the literature were unveiled. While the majority of industries leveraged extensively from TM techniques and systems, the use of TM on policy text failed to reach the same pedestal. The following answers give an in-depth description of the current state of literature, implying that next to the systematic literature review, the biggest contribution of this research comes from the introduced framework. First and foremost, being a fully validated framework, is what makes it novel in the first place. Additionally, its use of unprecedented techniques, on business policies, and at the same time its ability to achieve appropriate results, introduces the potentials that can be derived from using TM on business policies. An aspect that has often been overlooked by previous studies. At the same time, this answers one of the research questions (*"To what extent can the followed case study approach be generalised in a new TM framework"*). In a time when the processing of policy documents suffers from a shortage in scientifically validated approaches, this framework introduces a novelty in the domain.

"RQ 1 - To what extent has TM been applied on policy documents?"

The fact that policies are installed across all organisations, engendered a general perception that TM can extensively contribute in processing these textual documents. This perception was additionally supported by the fact that such documents lack in standardisation and often are ambiguous and incomprehensible for the masses. Nonetheless, what the literature review showed was quite the opposite picture. In a time when the use of TM on domains such as biomedicine, finance, law and national security is flourishing, the use of TM on policy documents falls short in both, qualitative and quantitative aspect. The amount of publications that revolved around this idea was quite limited and failed to match the prominence of other domains. In addition to this, the available publications only considered privacy policies in their study, which policies are also known as "Terms and Conditions". It was apparent that most of the publications neglected internal policies. This small corpus of publications introduced numerous conceptual frameworks, nevertheless, the majority of them lacked a full scientific validation. This is also reflected in the ability to create a system that uses TM techniques for processing policy documents. Even though most of the literature revolved around the idea of creating such systems, only a small portion embodied the existing frameworks into working modules. Furthermore, as far as the qualitative aspect is concerned, the processing of policy documents often failed to go beyond text categorization/classification, IE, and Topic Modelling technique. Most of the researchers experimented with identifying thematic areas in the privacy policies and associating the part of the policy text to the thematic area it belongs. This made Topic Modeling one of the most utilised techniques in the literature. Next to Topic Modeling, IE was another technique that most of the researchers experimented with. They saw this technique as a suitable way to extract specific information from the text. This was often complemented with a linguistic analysis of the text and use of grammar rules to extract the necessary information. Nevertheless, the potential of other TM techniques, such as automatic text summarization, keyword extraction, QA, was not exploited on policy documents.

"SQ 1.1 – To what extent has TM been applied on bank policy documents?"

In a small corpus of publication regarding TM on policy document, even smaller was the amount of research that had financial policies as a subject of investigation. The presence of financial policies was evident in two publications, that of (A. I. Anton & Earp, 2003) and (Brodie et al., 2006). In their attempt to create the SPARCLE system, Brodie et al. (2006) used a large corpus of policies, which corpus had financial privacy policies in its composition. SPARCLE used IE to process policies in natural language and identify important policy elements which were then transformed into a machine readable format. Whereas, (A. I. Anton & Earp, 2003) created their corpus entirely from privacy policies of financial websites. Nevertheless, in their attempt to standardise and provide clarity in financial privacy policies, they did not rely on TM techniques. They utilised goal-driven requirement engineering and goal-mining heuristics to extract privacy vulnerability and privacy protection goals. Thus to the best of our knowledge, the literature lacks studies that utilise TM techniques on internal bank policies or internal financial policies in general.

"RQ 2 – Which TM techniques or frameworks have been applied to policy documents?"

The literature was quite repetitive regarding this aspect. Looking at the corpus of publications, it was evident that most of them shared the same scope of TM techniques in their examination. The most utilised techniques were text categorization/classification, IE and Topic Modeling. Most of the publications in the corpus relied solely on one TM technique to process privacy policies. Furthermore, when utilised, the authors made use of the general framework that this techniques rely upon, thus failing to introduce a novel processing framework. Nevertheless, it is worth mentioning that an exception is seen in the work of (Li et al., 2007). What makes this research exclusive, is the fact that it is exercised upon a corpus of business policies and at the same time, it utilises two TM techniques. In an attempt to identify process models from business policies, the authors initially use text classification to separate process policies from non-process policies. In addition to this, the author advocates that IE should be used to identify and extract the relation between major process components, such as date/time, resources, organisation, tasks etc. This study introduces a novel TM framework that is built upon two TM techniques. Nevertheless, what this framework lacks, is a full scientific validation, since up to date only the text classification aspect of the framework has been implemented and validated.

"SQ 2.1 – Which TM techniques or frameworks have been applied on bank policy documents?"

The processing of bank policies also suffered from the lack of attention that business policies have received so far. Up to date, there was no evidence of a technique or framework that has been applied on bank policies. A minor exception can be made for the two studies that were mentioned earlier, that of (Brodie et al., 2006), nevertheless, the presented approach lacked a full scientific validation.

"SQ 2.2- Which linguistically oriented techniques have been applied on bank policy documents?"

In general, the use of IE framework was often associated with the use of linguistically oriented techniques. Nearly in all the cases when IE was used, researchers used linguistically oriented techniques to easier identify and extract relevant information. The most utilised approach was the use of a shallow parser, which uses its linguistic knowledge to identify the syntactic structure of the words in the text and assign them a Part-of-Speech (POS) tag. Subsequently, these POS tags assist in constructing grammars for identify the necessary information based on their POS pattern. This technique was also present in processing financial privacy policies. Unfortunately, such a linguistically oriented technique has not been applied on internal bank policies so far.

"SQ 2.3 - What is the level of similarity between the bank and non-bank policies, in terms of TM techniques and frameworks that have been applied on them?"

The relatively small corpus that treats bank policies and policy documents in general, makes it impossible to provide an elaborate answer to this research question. The literature showed that when dealing with privacy policies, the same techniques and frameworks have been used to process financial and non-financial policies accordingly. Nevertheless, such a statement cannot be generalised for internal bank policies and other internal policies. This is mostly due to the fact that privacy policies across different industries share similar concepts, whereas the content of internal policies fully relies on the industry it is applicable to.

"RQ3 – Which TM techniques can be used to obtain information that would enable an easier navigation through the policies?"

In general, all existing TM techniques are designated to process unstructured data and retrieve some form of information from them. These techniques are capable to extract information directly from the data, can find hidden thematic of the unstructured data, can create a concise portrayal of the entire assembly of unstructured data, and much more. In essence, all TM techniques are capable and specialised for processing noisy data and manifest a uniform representation of the same. The literature is a good exemplar of the fitness that these techniques showed when applied to privacy policies. Whereas, the case study outcome demonstrated that TM techniques are capable to properly process business policies and extract relevant information from them. Thus depending on what policies one wants to process, IE, keyword extraction and automatic summarization proved to be adequate techniques for processing business policies and retrieve information that can enable an easier navigation and comprehension of the same. Whereas, techniques such as IE, text classification and Topic Modelling, were capable of bringing some understanding to privacy policies and understand their restraints.

7. DISCUSSION

The case study, together with the results that derived from its implementation, provide a glimpse into the possibilities that TM offers in analysing and simplifying text heavy documents such as policies. A testimony of its success is the framework that was introduced earlier. A framework that gives a step-by-step description on how three different techniques can be utilised independently on policy-like documents. The better part of the framework showed to be highly capable in analysing such unstructured data, yielding results that are quite impressive even when compared to similar studies. Nevertheless, trying to think about the bigger picture, it is a matter of discussion whether such an approach can be used in other circumstances.

Regarding this issue, one should consider that although the framework was evaluated on domain specific documents, its composition relies entirely on generic NLP modules. The fact that this framework only relies on such artefacts, and still manages to yield promising results, is an indication of what the answer might be. A potential generalizability also comes from the fact that textual content from the legal domain shares some similarities with policy text. This makes it reasonable to expect similar success if the given approach is adapted in the legal domain. Nevertheless, in a time when the benefits of this framework for policy documents are apparent, the question is how significant is such a framework for other domains. Looking at the literature on legal text, the summarization of such lengthy documents was a common practice. This, together with the use of IE for identifying relevant entities, were commonly used to acquire a timelier understanding of such documents. Additionally, the continuous increase in biomedical publication makes it quite difficult to be up to date with the latest developments in the field. Thus, the automatic summarization of these documents, together with the ability to extract relevant entities from the text (i.e. enzyme reaction, drug names, chemical reactions etc.) is seen as a smart solution that enables to follow the domain developments in a contemporary manner.

Moving towards a more technical discussion, from the cluster of different techniques it can be seen that in the case of reference extraction an important factor was the chunking grammar. The consistency in POS tags across the corpus of unstructured data made it possible to construct an overarching extraction module for significant information. Thus as long as these requirements are met, and the relevant information follows a similar representation throughout the corpus, the use of the given IE approach can derive optimistic results for other domains as well. The rest of the techniques also proved that, to a large extent, they are capable of distinguishing the most important entities of a document, even without any indication of what the relevant information for the domain might be. This domain independence dims the applicability constraints of the introduced framework. Nevertheless, the followed approach may not always be the single version of the truth. Changes or updates can be made to this approach that may (not) fine tune the results, yet still, adhere to the same framework design. Examples of such changes may be the use of a different similarity measure instead of LD. There are numerous other similarity measures that one can choose, such as cosine similarity (Mihalcea et al., 2006) or Google distance similarity (Cilibrasi & Vitányi, 2007), that can replace the role of LD in the graph, and still follow the same framework logic. Furthermore, the key-phrase constructor can be updated depending on the length of the relevant phrases without impacting the overall framework. Additionally, the semantic filters can always be adapted, based on the POS representation of the domain words, yet not causing any mutation to the overall framework. It is still debatable whether these changes can have an impact on the results for the better or worst. Nevertheless, what this discussion indicates, is that the framework is capable of handling methodological changes without sustaining significant design change.

7.1 LIMITATIONS

As far as limitations are concerned, there are some areas that introduce some form of limitation to the study. Most of the limitations come as a result of the dataset, upon which the experiments were conducted. Taking into consideration that organisations, especially the ones in the financial domain, operate under a large number of internal policies, one may argue that a dataset of 23 internal policies is relatively small. This paves the way for discussing whether similar results will be acquired in a larger dataset. Next, to this, the result evaluation process also faced some limitations. Due to time constraint, the summaries could only be evaluated in a form of acceptance testing. Thus, one may argue that such an evaluation approach may yield bias results. A better approach would have been to rely on statistical methods for summary evaluation, nevertheless for the time being such a thing was not possible. Furthermore, given the fact that this research was designated only for internal documents, it was limited only to a single case study. This means that the derived results are applicable only for the organisation where the experiment took place, thus impacting the external validity of the framework. Nevertheless, as it was mentioned earlier, framework gives enough freedom to tune it, depending on the applicability circumstances. Furthermore, based on the expert evaluation, limitations were also found in the implementation of the artefacts. While key-phrases were only constructed from two words, a considerable amount of expert suggestions exceeded this word limit.

7.2 FUTURE WORK

Given that the proposed framework is on its early days, we initially intend to test it on a larger dataset, thus also overcoming one of the limitations. It can be argued that as the dataset increases, the representation of relevant information tends to differ more and more. Thus, by testing it on a larger dataset, we want to see the impact of the dataset on the results. Next, to this, we call for the framework to be tested in different settings. What we mean by this is that the framework should be tested in both: a) a different industry and b) different bank. This will not only impact its validity, but it will also show whether the framework is still capable of deriving promising results by relying solely on generic modules. Furthermore, in this study, the IE aspect was only concerned with the references between policies. Considering that internal policies hold an extensive amount of valuable information, we consider testing this part of the framework on different types of relevant information as well. At the same time, given that the aspect using TM on internal policy documents is not that cultivated, we suggest that the future work in this domain should also consider techniques other than the ones utilized in the framework. This will not only contribute in the field, but it will also show whether other TM techniques can be implemented in harmony with the introduced framework. Additionally, since a motivation for this research was to determine a way that will enable an easier navigation through policy documents. We have yet to create an interface that will enable the users to search for policies, using the set of information that were extracted in the study, and evaluate its efficiency.

8. BIBLIOGRAPHY

A.Hearst, M. (1999). Untangling Text Data Mining. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 3–10.

Achananuparp, P., Hu, X., & Shen, X. (2008). The evaluation of sentence similarity measures. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5182 LNCS*, 305–316. http://doi.org/10.1007/978-3-540-85836-2_29

Adwait Ratnaparkhi. (1996). A maximum entropy model for part-of-speech tagging. *In Proceedings of the Empirical Methods in Natural Language Processing Conference*, *1*(49), 133–142. Retrieved from http://acl.ldc.upenn.edu/W/W96/W96-0213.pdf

Ammar, W., Wilson, S., Sadeh, N., & Smith, N. A. (2012). Automatic categorization of privacy policies: A pilot study. *School of Computer Science, Language Technology Institute, Technical Report CMU-LTI-12-019*.

Andreessen, M. (2011). Why Software Is Eating The World. *Wall Street Journal*, 1–5. Retrieved from http://online.wsj.com/article/SB10001424053111903480904576512250915629460.html

Anton, A., & Earp, J. (2004). A requirements taxonomy for reducing Web site privacy vulnerabilities. *Requirements Engineering*, *9*, 169–185. http://doi.org/10.1007/s00766-003-0183-z

Anton, A. I., & Earp, J. (2003). The Lack of Clarity in Financial Privacy Policies and the Need for Standardization, (August), 1–12. Retrieved from http://www.truststc.org/wise/articles2009/article4.pdf

Baars, H., & Kemper, H.-G. (2008). Management Support with Structured and Unstructured Data - An Integrated Business Intelligence Framework. *Information Systems Management*, *25*(2), 132–148. http://doi.org/10.1080/10580530801941058

Badr, Y. (2011). Journal of Emerging Technologies in Web Intelligence, *3*(1), 104–114.

Berkeley, U. C. (2003). What Is Text Mining ? Marti Hearst, 1–3.

Berkouwer, C. (2009). Master Thesis The Reflection of Foresight in Defence Policy Making : A Comparative Study of the United Kingdom and the United States, (March).

Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text Mining for Central Banks. *Centre for Central Banking Studies, Handbook*, *33*, 1–19.

Bilisoly, R. (2011). *Practical text mining with Perl (Vol. 2)*. John Wiley & Sons.

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

Blei, D. M. (2012). Introduction to Probabilistic Topic Modeling. *Communications of the ACM*, *55*, 77–84. http://doi.org/10.1145/2133806.2133826

Boukus, E., & Rosenberg, J. V. (2006). The Information Content of FOMC Minutes. *SSRN Electronic Journal*, (February). http://doi.org/10.2139/ssrn.922312

Brinkkemper, S. (1996). Method engineering: Engineering of information systems development methods and tools. *Information and Software Technology*, *38*(4 SPEC. ISS.), 275–280. http://doi.org/10.1016/0950-5849(95)01059-9

Brodie, C. A., Karat, C.-M., & Karat, J. (2006). An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench. *Proceedings of the Second Symposium on Usable Privacy and Security - SOUPS '06*, 8. http://doi.org/10.1145/1143120.1143123

Brüninghaus, S., & Ashley, K. D. (2001). Improving the representation of legal case texts with information extraction methods. *Proceedings of the International Conference on Artificial Intelligence and Law*, 42–51. http://doi.org/10.1145/383535.383540

Bruno, G. (2016). Text Mining and Sentiment Extraction in Central Bank Documents.

Cardie, C. (1997). Empirical Methods in Information Extraction. *AI Magazine*, *18*(4), 65–79. http://doi.org/10.1.1.20.8120

Chang, C.-H., Kayed, M., Girgis, M. R., & Shaalan, K. F. (2006). A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1411–1428. http://doi.org/10.1109/TKDE.2006.152

Charniak, E. (1996). Statistical language learning, 170.

Checklist, I. R., & Management, S. I. (2008). Institutional Repository, (April), 1–4.

Chieze, E., Farzindar, A., & Lapalme, G. (2010). An Automatic System for Summarization and Information Extraction of Legal Information, 216–234. http://doi.org/10.1007/978-3-642-12837-0_12

Cilibrasi, R. L., & Vitányi, P. M. B. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, *19*(3), 370–383. http://doi.org/10.1109/TKDE.2007.48

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25 Th International Confer- Ence on Machine Learning*. Retrieved from http://dl.acm.org/citation.cfm?id=1390177

Copeland, L. (2003). *A Practitioner's Guide to Software Test Design*. Artech House.

Costante, E., den Hartog, J., & Petković, M. (2013). What Websites Know About You. *Data Privacy Management and Autonomous Spontaneous Security*, 146–159. http://doi.org/10.1007/978-3-642-35890-6

Costante, E., Sun, Y., Petković, M., & den Hartog, J. (2012). A machine learning solution to assess privacy policy completeness. *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society - WPES '12*, 91. http://doi.org/10.1145/2381966.2381979

Costantino, M., Collingham, R., & Richard, G. (1996). Qualitative Information in Finance : Natural Language Processing and Information Extraction.

Count, F. O. G., Reading, F., & Personnel, E. (1975). Derivation of new readability formulas (automated readability inde, fog count and flesch reading ease formula) for navy enlisted personel.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the*

Humanities, *36*(2), 223–254. http://doi.org/10.1023/A:1014348124664

Cunningham, H., Maynard, D., & Bontcheva, K. (2002). DGATE: an Architecture for Development of Robust HLT Applications, *19*(July), 168–175.

Das, S. R. (2014). *Text and Context: Language Analytics in Finance*. *Foundations and Trends® in Finance* (Vol. 8). http://doi.org/10.1561/0500000045

Denicia-Carral, C., Montes-y-Gomez, M., Villasenor-Pineda, L., & Hernandez, R. G. (2006). A text mining approach for definition question answering. *Advances in Natural Language Processing, Proceedings*, *4139*, 76–86. Retrieved from <Go to ISI>://WOS:000240270400008

Dictionary, O. E. (2007). *Oxford English dictionary online*.

Dijrre, J., Gerstl, P., & Seiffert, R. (1999). Finding Text Mining: Nuggets in Mountains of Textual Data. *KDD '99 Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 398–401. http://doi.org/10.1145/312129.312299

Dozier, C., & Haschart, R. (2000). Automatic extraction and linking of person names in legal text. *In Proceedings of RIAO '2000; Content Based Multimedia Information Access*, (September), 1305–1321. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.7431&amp;rep=rep1&amp;type=pdf

Drucker, H., Wu, D., & Vapnik, V. N. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, *10*(5), 1048–1054. http://doi.org/10.1109/72.788645

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479. http://doi.org/10.1613/jair.1523

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006b). Tapping the power of text mining. *Communications of the ACM*, *49*(9), 76–82. http://doi.org/10.1145/1151030.1151032

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006a). Tapping the power of text mining. *Communications of the ACM*. http://doi.org/10.1145/1151030.1151032

Farzindar, A., & Lapalme, G. (2004). Letsum, an automatic legal text summarizing system. *Legal Knowledge and Information Systems, JURIX*, 11–18.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27–34. http://doi.org/10.1145/240455.240464

Feinerer, I., Hornik, K., & Meyer, D. (2013). *Text Mining Infrastructure in R*. Retrieved from http://www.slideshare.net/waqasbcs/institutional-repository-16510631?qid=56cf08fa-51a3-449e-8894-fe9f0c44c28f&v=default&b=&from_search=12

Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). *International Conference on Knowledge Discovery and Data Mining (KDD)*, 112–117. http://doi.org/10.1.1.47.7462

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., … Zamir, O. (1998). Text Mining at the Term Level. *Second European Symposium, PKDD '98*, (September), 65–73. http://doi.org/10.1007/BFb0094806

Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. *Online - Weston Then Wilton*, *23*(May), 62–73.

Flesch, R. (1948). A New Readability Yardstick. *The Journal of Applied Psychology*, *32*(3), 221–233. http://doi.org/10.1037/h0057532

Frawley, W. J., Piatetsky-shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases : An Overview. *AI Magazine*, *13*(3), 57–70. http://doi.org/10.1609/aimag.v13i3.1011

Friedman, C., Johnson, S. B., Forman, B., & Starren, J. (1995). Architectural requirements for a multipurpose natural language processor in the clinical environment. *Proceedings of the Symposium on Computer Applications in Medical Care*, 347–351.

Galgani, F., Compton, P., & Hoffmann, A. (2012). Combining Different Summarization Techniques for Legal Text. *EACL 2012 Hybrid 2012 Innovative Hybrid Approaches to the Processing of Textual Data*, 115–123.

Grefenstette, G. (1996). Light parsing as finite-state filtering.

Grimes, S. (2008). Unstructured data and the 80 percent rule. *Carabridge Bridgepoints*.

Grover, C., Hachey, B., Hughson, I., & Korycinski, C. (2003). Automatic summarisation of legal documents. *Proceedings of the 9th International Conference on Artificial Intelligence and Law - ICAIL '03*, 243. http://doi.org/10.1145/1047788.1047839

Grover, C., Matheson, C., Mikheev, A., & Moens, M. (2000). LT TTT - A Flexible Tokenisation Tool. *Proc. LREC 2000*. Retrieved from http://www.ltg.ed.ac.uk/papers/

Gunning, R. (1952). The technique of clear writing.

Gupta, R., & Gill, N. S. (2012). Financial Statement Fraud Detection using Text Mining. *Editorial Preface*, *3*(12), 189–191.

Gupta, V., & Lehal, G. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76. http://doi.org/10.4304/jetwi.1.1.60-76

Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, *55*(3), 685–697. http://doi.org/10.1016/j.dss.2013.02.006

Haug, P. J., Ranum, D. L., & Frederick, P. R. (1990). Computerized extraction of coded findings from free-text radiologic reports. Work in progress. *Radiology*, *174*(2), 543–8. http://doi.org/10.1148/radiology.174.2.2404321

Heeringa, W. (2004). Measuring dialect pronunciation differences using Levenshtein distance. *Dissertations.Ub.Rug.Nl*. http://doi.org/10.1.1.222.285

Hendry, S., & Madeley, A. (2010). Text Mining and the Information Content of Bank of Canada

Communications.

Hevner, A. R., March, S. T., Park, J., Ram, S., & Ram, S. (2004). Research Essay Design Science in Information. *MIS Quarterly*, *28*(1), 75–105.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296–298. http://doi.org/10.1197/jamia.M1733

Hulth, A., & Megyeesi, B. B. (2006). A study on automatically extracted keywords in text categorization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 353–360). http://doi.org/10.3115/1220175.1220220

Humphreys, K., Demetriou, G., & Gaizauskas, R. (2000). Two Applications of Information Extraction to Biological Science Journal Articles : Enzyme Interactions and Protein Structures. *Pacific Symposium on Biocomputing*, *5*, 502–513.

Iiritano, S., & Ruffolo, M. (2001). Managing the knowledge contained in electronic documents: A clustering method for text mining. *Proceedings - International Workshop on Database and Expert Systems Applications, DEXA*, *2001-Janua*, 454–458. http://doi.org/10.1109/DEXA.2001.953103

Inouye, D., & Kalita, J. K. (2011). Comparing twitter summarization algorithms for multiple post summaries. *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 298–306. http://doi.org/10.1109/PASSAT/SocialCom.2011.31

Juárez-González, A., Téllez-Valero, A., Denicia-Carral, C., Montes-y-Gómez, M., & Villaseñor-Pineda, L. (2007). Using machine learning and text mining in question answering. *Evaluation of Multilingual and Multimodal Information Retrieval*, *4730/2007*(4730), 415–423. Retrieved from http://www.springerlink.com/index/NQ1106525234L680.pdf

Junqu De Fortuny, E., De Smedt, T., Martens, D., & Daelemans, W. (2014). Evaluating and understanding text-based stock price prediction models. *Information Processing and Management*, *50*(2), 426–441. http://doi.org/10.1016/j.ipm.2013.12.002

Kamaruddin, S. S., Hamdan, A. R., & Bakar, A. A. (2007). Text Mining for Deviation Detection in Financial Statement, 446–449.

Karanikas, H., Tjortjis, C., & Theodoulidis, B. (2000). An approach to text mining using information extraction. *Proc. Workshop Knowledge Management Theory Applications (Kmta 00).*, (Dm). Retrieved from http://eric.univ-lyon2.fr/~pkdd2000/Download/WS5_13.pdf

Kloptchenko, A., Eklund, T., Karlsson, J., Back, B., Vanharanta, H., & Visa, A. (2004). Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance and Management*, *12*(1), 29–41. http://doi.org/10.1002/isaf.239

Kloptchenko, A., Magnusson, C., Back, B., Visa, A., & Vanharanta, H. (2004). Mining Textual Contents of Financial Reports. *The International Journal of Digital Accounting Research*, *4*(47), 1–29. http://doi.org/10.4192/1577-8517-v4_1

Kongthon, A. (2004a). A text Mining Framework for Dicovery Technological Intelligence, (April).

Kongthon, A. (2004b). A Text Mining Framework for Discovering Technological Intelligence to Support Science and TechnologyManagement, (April).

Kumar, E. (2011). *Natural Language Processing*. IK International PVT Ltd.

Larkey, L. S. (1999). A patent search and classification system. *Proceedings of the Fourth ACM Conference on Digital Libraries - DL '99*, *1*(212), 179–187. http://doi.org/10.1145/313238.313304

Larson, M. (2012). *Automatic Summarization*. *Foundations and Trends® in Information Retrieval* (Vol. 5). http://doi.org/10.1561/1500000020

Laws, D. (2008). Automatic Classification of Sentences in.

Leroy, G., Chen, H., & Martinez, J. D. (2003). A shallow parser based on closed-class words to capture relations in biomedical text. *Journal of Biomedical Informatics*, *36*(3), 145–158. http://doi.org/10.1016/S1532-0464(03)00039-X

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. http://doi.org/citeulike-article-id:311174

Li, J., Wang, H. J., & Zhao, J. L. (2007). Mining Business Policy Texts for Discovering Process Models : A Framework and Some Initial Results. *Policy*, 1–12.

Liddy, E. (1998). Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science*. Retrieved from http://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/MRI 06 - Liddy, ED - 1998.pdf\nhttp://onlinelibrary.wiley.com/doi/10.1002/bult.91/full

Ling, C. S. N., Maccartney, B., & Penn, G. (2011). Natural Language Processing.

Liu, F., Pennell, D., & Liu, Y. (2009). Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics* (pp. 620–628).

Liu, F., Ramanath, R., Sadeh, N., & Smith, N. (2014). A Step Towards Usable Privacy Policy: Unsupervised Alignment of Privacy Statements. *Proceedings of The 25th International Conference on Computational Linguistics (COLING 2014)*, 884–894. Retrieved from http://www.coling-2014.org/accepted-papers/585.php

Liu, Y. (2016). *Question answering for biomedicine*. University of Alberta.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text Classification using String Kernels. *Journal of Machine Learning Research*, *2*, 419–444. http://doi.org/10.1162/153244302760200687

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. http://doi.org/10.3115/1118108.1118117

Makhoul, J., & Kubala, F. (1999). Performance measures for information extraction. *In Proceedings of DARPA Broadcast News Workshop*, 249–252. Retrieved from http://books.google.com/books?hl=en&lr=&id=uuR3mpBI5ksC&oi=fnd&pg=PA249&dq=Performance+measures+for+information+extraction&ots=DN2Am7TIcT&sig=47QBj2yOUssENGUPKOnbJETOPBI

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit (52nd ed., pp. 55 – 66). Maryland , USA: Baltimore, System Demonstrations.

Mantrach, A. (2008). Text Mining. *Foundations*, 1–32. http://doi.org/10.1017/CBO9781107415324.004

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. http://doi.org/10.1162/coli.2010.36.1.36100

Martinez, A. R. (2012). Part-of-speech tagging. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(1), 107–113. http://doi.org/10.1002/wics.195

Martínez-González, M. (2005). Reference extraction and resolution for legal texts. *PReMI'05 Proceedings of the First International Conference on Pattern Recognition and Machine Intelligence*, 218–221. http://doi.org/10.1007/11590316_29

Massey, A. K., Eisenstein, J., Antón, A. I., & Swire, P. P. (2013). Automated Text Mining for Requirements Analysis of Policy Documents, 4–13. Retrieved from http://www.cc.gatech.edu/~jeisenst/papers/re13rt-p085-p-18125-preprint.pdf

Maynard, D., Tablan, V., & Ursu, C. (2001). Named entity recognition from diverse text types. *Recent Advances in Natural Language Processing*, (August 2002), 257–274. http://doi.org/10.1007/978-3-540-85287-2_42

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, *12*(8), 639–646. http://doi.org/10.1039/b105878a

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, *35*, 128–144.

Michael, J. B., Ong, V. L., & Rowe, N. C. (2001). Natural-language processing support for developing policy-governed software systems. *Proceedings 39th International Conference and Exhibition on Technology of Object-Oriented Languages and Systems. TOOLS 39*, (July), 263–274. http://doi.org/10.1109/TOOLS.2001.941679

Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Proceedings of the 21st National Conference on Artificial Intelligence*, *1*, 775–780. http://doi.org/10.1.1.65.3690

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. *Association for Computational Linguistics*.

Moens, M.-F., Uyttendaele, C., & Dumortier, J. (2000). Intelligent Information Extraction from Legal Texts, *1*. http://doi.org/10.1525/sp.2007.54.1.23.

Mokhov, S. A. (2010). Evolution of {MARF} and its {NLP} Framework. *Proceedings of C3S2E'10*, 118–122. http://doi.org/10.1145/1822327.1822344

Moniz, A., & De Jong, F. (2014). Predicting the impact of central bank communications on financial market investors' interest rate expectations. *CEUR Workshop Proceedings*, *1240*, 75–86. http://doi.org/10.1007/978-3-319-11955-7_12

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2012). Natural language processing: an introduction. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 544–51. http://doi.org/10.1136/amiajnl-2011-000464

Nenkova, A., & Bagga, A. (2003). Facilitating Email Thread Access by Extractive Summary Generation. *Recent Advances in Natural Language Processing III, Selected Papers from RANLP'03*, *260*, 287–296.

Niharika, S., Latha, V. S., & Lavanya, D. R. (2012). A survey on text categorization. *International Journal of Computer Trends and Technology*, *3*(1), 39–45. Retrieved from http://www.ijcttjournal.org/Volume3/issue-1/IJCTT-V3I1P108.pdf

Nilsson, F. (2013). REBOK V1.0 Requirements Engineering Body Of Knowledge. *REQB® Certified Professional for Requirements Engineering*, *1.0*, 59.

Page, L., & Brin, S. (1997). PageRank: Bringing order to the web. *Stanford Digital Libraries Working Paper*, *72*.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. *J. of Management Information Systems*, *24*(3), 45–77. Retrieved from http://dblp.uni-trier.de/db/journals/jmis/jmis24.html#PeffersTRC08

Perkins, J. (2010). *Python text processing with NLTK 2.0 cookbook*. *Processing*. Packt Publishing Ltd. http://doi.org/10.1017/CBO9781107415324.004

Popel, M. (2010). Popel - TectoMT - Modular NLP Framework, 293–304.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63. http://doi.org/10.1.1.214.9232

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., … Zhang, Z. (2004). MEAD - A platform for multidocument multilingual text summarization. *Conference on Language Resources and Evaluation (LREC)*, 699–702. Retrieved from http://clair.si.umich.edu/~radev/papers/lrec-mead04.pdf

Rajaraman, A., & Ullman, J. D. (2011). Data Mining. *Mining of Massive Datasets*, *18 Suppl*, 114–142. http://doi.org/10.1007/978-1-4419-1280-0

Rajman, M. (1997). Text mining: Natural language techniques and text mining applications. *In Proceedings of the 7 Th IFIP*, *1998*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.4827

Ramanath, R., Liu, F., Sadeh, N., & Smith, N. (2014). Unsupervised alignment of privacy policies using hidden Markov models, 605–610. Retrieved from http://repository.cmu.edu/lti/150/

Renz, I., Ficzay, A., & Hitzler, H. (2003). Keyword Extraction for Text Characterization. *8th International Conference on Applications of Natural Language to Information Systems*, 228–234.

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1–277. http://doi.org/10.1002/9780470689646.ch1

Sadeh, N., Acquisti, A., Breaux, T. D., Cranor, L. F., Mcdonald, A. M., Reidenberg, J. R., … Wilson, S. (2013). The Usable Privacy Policy Project : Combining Crowdsourcing , Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About. *Tech. Report CMU-ISR-13-119*, (1). Retrieved from http://ra.adm.cs.cmu.edu/anon/usr0/ftp/home/anon/isr2013/CMU-ISR-13-119.pdf

Sanner, M. F. (1999). Python: a programming language for software integration and development. *Journal of Molecular Graphics & Modelling*, *17*(1), 57–61. http://doi.org/10.1016/S1093-3263(99)99999-0

Saravanan, M., Ravindran, B., & Raman, S. (2006). Improving Legal Document Summarization Using Graphical Models. *Jurix*, *152*(April 2014), 51–60.

Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*, 1–5. Retrieved from http://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf

Schumaker, R. P., & Chen, H. (2009). A quantitative stock prediction system based on financial news. *Information Processing and Management*, *45*(5), 571–583. http://doi.org/10.1016/j.ipm.2009.05.001

Sibley, E. ., Michael, J. ., & Wexelblat, R. . (1991). Use of an experimental policy workbench: Description and preliminary results. *Results of the IFIP WG 11.3 Workshop on Database Security V: Status and Prospects*, 47–76.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA.

Soukoreff, R. W., & MacKenzie, I. S. (2001). Measuring errors in text entry tasks. *CHI ’01 Extended Abstracts on Human Factors in Computing Systems - CHI '01*, 319. http://doi.org/10.1145/634067.634256

Spence, D. (2010). Data, data everywhere: a special report on managing information. *The Economist*, 1–10. http://doi.org/10.1136/bmj.f725

Spruit, M. . (2006). Measuring syntactic variation in Dutch dialects. *Literary and Linguistic Computing*, *21*(4), 493–506.

Spruit, M., & Jagesar, R. (2016). Power to the People! - Meta-Algorithmic Modelling in Applied Data Science. *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, (January 2016), 400–406. http://doi.org/10.5220/0006081604000406

Stake, R. (1994). Case Studies en NK DENZIN. y YS LINCOLN.(eds.). *Handbook of Qualitative Research*, 236–247.

Stake, R. E. (1995). *The art of case study research*. Sage.

Stamey, J. W., & Rossi, R. a. (2009). Automatically identifying relations in privacy policies. *Proceedings of the 27th ACM International Conference on Design of Communication - SIGDOC '09*, 233. http://doi.org/10.1145/1621995.1622041

Street, C. (1986). Amount : Currency : Value Date : Sender name : Sender city : Sender ref :, 1089–1093.

Street, N. (2000). Exploring the forecasting potential of company annual reports, (1993).

Tan, A. (1999). Text Mining : The state of the art and the challenges Concept-based. *In Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases*, 65–70. Retrieved from http://www.ntu.edu.sg/home/asahtan/Papers/tm_pakdd99.pdf

Teufel, S. (2001). Task-based evaluation of summary quality: Describing relationships between scientific papers. *In Workshop Automatic Summarization, NAACL*, *102*.

Tucker, S., & Whittaker, S. (2008). No TitleTemporal compression of speech: An evaluation. *IEEE Transactions on Audio, Speech, and Language Processing, 4*(16), 790–796.

Turban, E., Sharda, R., Aronson, J. E., & King, D. (2008). Text and Web Mining. In *Business intelligence: A managerial approach* (pp. 189–230). New Jersey: Pearson Prentice Hall.

van de Weerd, I., & Brinkkemper, S. (2008). Meta-Modeling for Situational Analysis and Design Methods. *Handbook of Research on Modern Systems Analysis and Design Technologies and Applications*, 35–54. http://doi.org/10.4018/978-1-59904-887-1

Vanderwende, L., Suzuki, H., Brockett, C., & Nenkova, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management*, *43*(6), 1606–1618. http://doi.org/10.1016/j.ipm.2007.01.023

Voutilainen, A. (2003). *Part-of-speech tagging*. The Oxford handbook of computational linguistics.

Weiss, S., Apte, C., & Damerau, F. (1999). Maximizing text-mining performance. *IEEE Intelligent Systems*, *14*(4), 63–69. Retrieved from http://dl.acm.org/citation.cfm?id=630474

Wilson, D. J. H., Irwin, G. W., & Lightbody, G. (1997). Neural networks and multivariate SPC. *IEEE Colloquium on Faults Diagnosis in Process Systems*, *1/5-5/5*(MAY), 20080258. http://doi.org/10.1049/ic

Wilson, R. C., & Hancock, E. R. (2004). Levenshtein distance for graph spectral features. *Proceedings - International Conference on Pattern Recognition*, *2*(C), 489–492. http://doi.org/10.1109/ICPR.2004.1334272

Witten, I. H. (2004). Text Mining. *International Journal of Computational Biology and Drug Design*, 198. http://doi.org/10.1504/IJCBDD.2011.041412

Wohlin, C. (2014). Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. *18th International Conference on Evaluation and Assessment in Software Engineering (EASE 2014)*, 1–10. http://doi.org/10.1145/2601248.2601268

Xiao, X., Paradkar, A., & Xie, T. (2011). Automated extraction and validation of security policies from natural-language documents. *Perspective*, 11. http://doi.org/10.1145/2393596.2393608

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval - SIGIR '03*, 267. http://doi.org/10.1145/860484.860485

Yang, K., Chen, Z., Cai, Y., Huang, D. P., & Leung, H. fung. (2016). Improved automatic keyword extraction given more semantic knowledge. *Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9645*(April), 112–125. http://doi.org/10.1007/978-3-319-32055-7_10

Yin, R. K. (2009). *Case study research : design and methods*. *Applied social research methods series ;* (Vol. 5.). http://doi.org/10.1097/FCH.0b013e31822dda9e

Yin, R. K. (2014). Case study: design and methods. Retrieved from https://books.google.de/books?hl=de&lr=&id=AjV1AwAAQBAJ&oi=fnd&pg=PP1&dq=Yin,+R.K.+(2003)+Case+Study+Research:+Design+and+Methods+(3rd+edn).+Thousand+Oaks,+CA:+Sage.&ots=gkThlGHQcN&sig=nvB2euds1JwBmtjwd-o4TZbcVmQ#v=onepage&q&f=false

Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., & Wang, B. (2008). Automatic Keyword Extraction from Documents Using Conditional Random Fields. *Journal of Computational Information*, *43*, 1169–1180. Retrieved from http://www.jofci.org

Zhao, Y. (2013). Analysing Twitter Data with Text Mining and Social Network Analysis. *11th Australasian Data Mining Conference (AusDM 2013)*, (DECEMBER 2013), 41–47.

Zhong, S., & Ghosh, J. (2005). Generative model-based document clustering: A comparative study. *Knowledge and Information Systems*, *8*(3), 374–384. http://doi.org/10.1007/s10115-004-0194-1

Zonabend, F. (1992). The monograph in European ethnology. *Current Sociology*, *40*(1), 49–54.

Appendix A – Evaluation Form

Guidance on how to evaluate the algorithm outcome

The algorithm creates two types of outputs for each policy. It creates a summary of the policy with around 250 words and it creates a range of 8-51 keywords (key phrases) for each policy.

As an expert, you have to evaluate whether the algorithmic outputs are indeed relevant for the given policy.

For keywords:

1. Each new-line represents a separate keyword/key phrase
2. Highlight with green the keywords that you assess to be relevant (e.g. income)
3. Highlight with red the keywords that you assess to be irrelevant (e.g. income)
4. If there is a missing keyword add it and highlight it with yellow (e.g. revenue) – you can add it in the "**Comment**" section, or add it with the rest of the keywords.
5. Note : The number of keywords is relative to the content of the policy. The more content it has, the more keywords are generated. Hence, different number of keywords for each policy.
6. Note 2: In case that the extracted key phrase has the following format "**currencies/exchange**". If at least one of the words is relevant, mark the whole phrase as relevant (green).

For summary:

1. Read the automatically generated summary.
2. In the "**Comment**" section provide a comment for the given summary. The comment should be in a form of an assessment, like:
   a. whether the summary covers the main aspect of the policy;
   b. whether there is something missing;
   c. how descriptive is the summary with respect to the policy;
   d. any comment that you see necessary

Evaluation Form

Policy Details	
Policy Title:	
Date:	
Reviewer:	

Algorithm Keywords	Comments

Algorithm Summary	Comments

Appendix B – Policy Reference Golden Standard

Policy A

1. Central Credit Risk Policy (AIM 101-21-75)
2. Residential Real Estate Policy (AIM 101-21-64)
3. Business Risk Policy Onroerend Goed C&MB Bedrijven en Corporate Clients (AIM 101-22-07)
4. “Agrarisch & Visserij (AIM 101-21-38)”
5. “Medici en Vrije Beroepers (Maatwerkfinancieringen)” (AIM 101-21-11)
6. “Credit Risk Real Estate Finance Krediet Policy” (AIM 101-21-36)
7. Business Risk Policy Onroerend Goed C&MB Bedrijven en Corporate Clients (AIM 101-22-07)
8. “Investment Real Estate Policy” (AIM 101-21-66)
9. Business Risk Policy Onroerend Goed van business line Particulieren (maatwerkfinancieringen) (AIM 101-21-10)
10. Central Credit Risk Policy (AIM 101-21-75).
11. (Central Collateral Policy (AIM 101-21-09)).
12. Central Collateral Policy (AIM 101-21-09),
13. Central Collateral Policy (AIM 101-21-09)
14. Central Collateral Policy (AIM 101-21-09),
15. Non-Retail Credit Risk Model Hierarchy (AIM 101-21-56)

Policy B

1. Central Credit Risk Policy (AIM 101-21-75).
2. Bank Risk Policy (AIM 101-23-40).
3. IFRS – Policy on netting (AIM 108-07-11).
4. Central Collateral Policy (AIM 101-21-09).
5. Credit Risk Monitoring Policy (AIM 101-21-76).
6. IFRS – Policy on Netting (AIM 108-07-11)
7. IFRS – Policy on Netting (AIM 108-07-11).
8. Central Collateral Policy (AIM 101-21-09)
9. Central Collateral Policy (AIM 101-21-09).
10. Central Collateral Policy (AIM 101-21-09).
11. Central Collateral Policy (AIM 101-21-09).
12. Central Collateral Policy (AIM 101-21-09)
13. General Product Approval Policy (AIM 101-23-80),
14. Credit Risk Monitoring Policy (AIM 101-21-76).
15. Central Collateral Policy (AIM 101-21-09)

Policy C

1. Policy for Capital Treatment of Securitisations of Own Originated Assets (AIM 101-21-14)
2. Policy for Capital Treatment of Securitisations of Own Originated Assets (AIM 101-21-14)
3. Economic Capital Policy (AIM 101-22-37)
4. Economic Capital Policy, (AIM 101-22-37)
5. Central Credit Risk Policy (AIM 101-21-75)

6. Policy for Capital Treatment of Securitisations of Own Originated Assets (AIM 101-21-14).
7. Securities Financing Risk Policy (AIM 101-21-54)4.
8. Basel II Reporting - CR SEC IRB Instructions (AIM 108-05-78)
9. Economic Capital Policy (AIM 101-22-37),
10. ICAAP Policy (AIM 101-22-36),
11. Firm Wide Stress Testing Framework Policy (AIM 101-21-59)
12. Pillar 3 Disclosure Policy (AIM 101-21-04).

Policy D

1. Economic Capital Policy (AIM 101-22-37)2.
2. Funding Policy (AIM 101-22-16)3.
3. Economic Capital Policy, AIM 101-22-37.
4. Funding Policy, AIM 101-22-16.
5. Policy for Capital Treatment of Positions in Third-Party Securitisations (AIM 101-21-79)
6. AIM 101-21-09, Central Collateral Policy.
7. Economic Capital Policy, AIM 101-22-37.
8. ICAAP policy, AIM 101-22-36.
9. Firm Wide Stress Testing Framework Policy, AIM 101-21-59.
10. Pillar 3 disclosure (AIM 101-21-04).
11. Pillar 3 Disclosure Policy (AIM 101-21-04).

Policy E

1. Back-to-Back Facility Policy (AIM 101-21-31).
2. Policy on Issue of Intercompany Guarantees (AIM 101-15-25)
3. Bank Risk Policy (AIM 101-23-40)
4. Central Credit Risk Policy (AIM 101-21-75).
5. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
6. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
7. ABN AMRO Clearing Market & Credit Risk Policy (AIM 101-21-51).
8. ABN AMRO Clearing Market & Credit Risk Policy (AIM 101-21-51).
9. Central Collateral Policy (AIM 101-21-09).
10. CAAML Policy (AIM 102-20-21)
11. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
12. Credit Risk Monitoring Policy (AIM 101-21-76)
13. Central Credit Risk Policy (AIM 101-21-75).

Policy F

1. Central Credit Risk Policy (AIM 101-21-75)
2. Credit Risk Monitoring Policy (AIM 101-21-76).
3. Fiduciary Duties (AIM 102-25-28)
4. Conflicts of Interest Policy (AIM 102-25-21).
5. ABN AMRO Product Approval Policy (AIM 101-23-80)

6. Product Approval For Markets Policy (AIM 101-21-49).
7. Legal Documentation & CSA Collateral for OTC Derivatives Risk Policy (AIM 101-21-52).
8. OBSI for Non-Professional Risk Policy (AIM 101-22-35).
9. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
10. "Legal Documentation & CSA Collateral for OTC Derivatives Risk Policy" (AIM 101-21-52)
11. OBSI for Non-Professionals Risk Policy (AIM 101-22-35).
12. Intercompany Credit Facilities Policy (AIM 101-22-30).
13. Non-Market Price Transactions Policy and Procedures (AIM 102-25-12).
14. Central Credit Risk Policy (AIM 101-21-75)
15. OBSI for Non-Professionals Risk Policy (AIM 101-22-35).
16. Overdispositiebeleid Bedrijven, Corporate Clients, Private Banking NL en Particulieren (AIM: 101-22-08).;
17. Credit Risk Monitoring Policy (AIM 101-21-76).
18. Legal Documentation & CSA Collateral for OTC Derivatives Risk Policy (AIM 101-21-52)".

Policy G

1. Central Credit Policy (AIM 101-21-75)
2. Central OBSI Credit Risk Policy (AIM 101-21-77)
3. Credit Risk Monitoring Policy (AIM 101-21-76).
4. Fiduciary Duties (AIM 102-25-28)
5. Conflicts of Interest Policy (AIM 102-25-21).
6. ABN AMRO Product Approval Policy (AIM 101-23-80)
7. Product Approval For Markets Policy (AIM 101-21-49).
8. Legal Documentation & CSA Collateral for OTC Derivatives Risk Policy (AIM 101-21-52).
9. Credit Valuation Adjustment Policy for OTC Derivatives (AIM 101-24-20)
10. International Lombard Lending Policy (AIM 101-21-62)
11. Overall Credit Policy Private Banking (AIM 101-21-60)
12. Fiduciary Duties (AIM 102-25-28)
13. Legal Documentation & CSA Collateral for OTC Derivatives Risk Policy (AIM 101-21-52).
14. Central OBSI Credit Risk Policy (AIM 101-21-77)

Policy H

1. Risk Taxonomy (AIM 101-21-01),
2. Risk Appetite Framework (AIM 101-21-07)
3. Risk Taxonomy (AIM 101-21-01)
4. Risk Taxonomy (AIM 101-21-01).
5. Business Segments Reporting Policy (AIM 104-40-10)).
6. Risk Governance Charter ABN AMRO (AIM 101-21-00).
7. Risk Governance Charter ABN AMRO (AIM 101-21-00).
8. Risk Governance Charter ABN AMRO (AIM 101-21-00)
9. International Risk Charter (AIM 101-21-05)
10. Risk Governance Charter ABN AMRO (AIM 101-21-00).
11. Risk Governance Charter ABN AMRO (AIM 101-21-00).

Policy I

1. Bank Risk Policy (AIM 101-23-40)
2. Central Credit Risk Policy (AIM 101-21-75)
3. Central Trading Risk Policy (AIM 101-24-02)
4. Central Liquidity Risk Policy (AIM 101-22-10)
5. Liquidity Coverage Ratio Policy (AIM 104-20-50)
6. Collateral Management Policy (AIM 101-22-50)
7. Financial Institutions Risk Policy (AIM 101-21-28)
8. ALM/Treasury Risk Limit Policy (AIM 101-22-32)
9. Legal Documentation & CSA Collateral for OTC Derivatives Risk Policy (AIM 101-21-52)
10. Country Risk Policy (AIM 101-21-20).
11. Country Risk Policy (AIM 101-21-20).
12. Country Risk Policy (AIM 101-21-20).
13. Financial Institutions Risk Policy (AIM 101-21-28)
14. Central Liquidity Risk Policy (AIM 101-22-10).
15. Funding Cost Policy (AIM 101-22-11).
16. Liquidity Coverage Ratio Policy (AIM 104-20-50)).
17. ALM/Treasury Risk Limit Policy (AIM 101-22-32).
18. Market Risk Capital Policy (101-24-10).
19. Market Risk Monitoring and Limit Policy for Trading Book (AIM 101-24-05).
20. Stress Testing and Scenario Analysis Policy for Trading Books (AIM 101-21-53).
21. Trading Model Governance Policy (AIM 101-21-45).
22. IFRS - Policy on Classification of Financial Instruments (AIM 108-07-04).
23. Legal Documentation & CSA Collateral for OTC Derivatives Risk Policy (AIM 101-21-52).
24. Product Approval for Markets Policy (AIM 101-21-49).
25. Product Approval for Markets Policy (AIM 101-21-49)
26. Change Risk Assessment Policy (AIM 101-23-79)
27. Operational Risk Policy (AIM 101-23-01)

Policy J

1. Legal Documentation & CSA Collateral for OTC Derivatives Risk Policy (AIM 101-21-52).
2. Bank Risk Policy (AIM 101-23-40)
3. Central Credit Risk Policy (AIM 101-21-75).
4. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
5. Credit Risk Monitoring Policy (AIM 101-21-76).
6. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
7. Economic Capital Policy (AIM number 101-22-37).
8. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84)
9. Back-to-Back Facility Policy (AIM 101-21-31)
10. Business Risk Policy MBO/MBI (AIM 101-22-06))
11. Single Stock Financing Policy for PBI Asia and Middle East (AIM 101-21-70)
12. Lombard Lending: The Netherlands Risk Policy (AIM 101-21-61),
13. International Lombard Lending Policy (AIM 101-21-62).

14. Lombard Lending: The Netherlands Risk Policy (AIM 101-21-61)
15. international Lombard Lending Policy (AIM 101-21-62),
16. Central Credit Risk Policy (AIM 101-21-75).
17. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
18. Credit Risk Monitoring Policy (AIM 101-21-76).
19. Business Risk Policy Borrowing Base (AIM 101-21-86).
20. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84
21. Central Real Estate Policy (AIM 101-21-69),
22. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84)
23. Risk Policy Diamond & Jewellery (AIM 411-24-03).
24. ECT Clients/Transportation Clients/Shipping Risk Policy (AIM 101-21-39).
25. Business Risk Policy Binnenvaart (AIM 101-22-09)
26. Yacht Finance Policy Private Banking (AIM 101-21-68)19
27. Business Risk Policy Agrarisch & Visserij (AIM 101-21-38)
28. Business Risk Policy Binnenvaart (AIM 101-22-09).
29. ECT Clients/Transportation Clients/Shipping Risk Policy (AIM 101-21-39)
30. Central Credit Risk Policy (AIM 101-21-75).

Policy K

1. • Bank Risk Policy (AIM 101-23-40)
2. Central Credit Risk Policy (AIM 101-21-75)
3. Risk Governance Charter ABN AMRO (AIM 101-21-00).
4. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
5. Central Credit Risk Policy (AIM 101-21-75),
6. Sustainability Risk Policy (AIM 101-25-00).

Policy L

1. Bank Risk Policy (AIM 101-23-40)
2. Central Credit Risk Policy (AIM 101-21-75)
3. Risk Governance Charter ABN AMRO (AIM 101-21-00)
4. International Risk Charter (AIM 101-21-05).
5. Policy on Issue of Intercompany Guarantees (AIM 101-15-25)
6. Policy on Issue of Intercompany Guarantees (AIM 101-15-25).
7. Policy on Issue of Intercompany Guarantees (AIM 101-15-25).
8. Policy on Issue of Intercompany Guarantees (AIM 101-15-25).
9. Back-to-Back Facility Policy (AIM 101-21-31).
10. Risk Governance Charter ABN AMRO (AIM 101-21-00)
11. International Risk Charter (AIM 101-21-05)
12. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
13. Risk Governance Charter ABN AMRO (AIM 101-21-00).
14. International Risk Charter (AIM 101-21-05).
15. Country Risk Policy (AIM 101-21-20)
16. Credit Risk Monitoring Policy (AIM 101-21-76).
17. Central Restructuring and Forbearance Policy (AIM 101-21-17).

18. Policy on Issue of Intercompany Guarantees (AIM 101-15-25).
19. Central Guarantee Policy (AIM 101-21-78).
20. Policy on Issue of Intercompany Guarantees (AIM 101-15-25).

Policy M

1. Bank Risk Policy (AIM 101-23-40)
2. Central Credit Risk Policy (AIM 101-21-75)
3. Central Restructuring and Forbearance Policy (AIM 101-21-17)
4. Central Loan Loss Provisioning and Write-off Policy (AIM 101-21-13)
5. Risk Governance Charter ABN AMRO (AIM 101-21-00)
6. IFRS - Policy on Recognition and Derecognition (AIM 108-07-03).
7. IFRS - Policy on Distinction Between Debt and Equity (AIM 108-07-09)
8. Central Loan Loss Provisioning and Write-off Policy (AIM 101-21-13)).
9. Central Restructuring and Forbearance Policy (AIM 101-21-17)).
10. Conflict of Interest Policy (AIM 102-25-21).
11. Risk Governance Charter ABN AMRO (AIM 101-21-00)
12. Risk Governance Charter ABN AMRO (AIM 101-21-00)
13. Risk Governance Charter ABN AMRO (AIM 101-21-00)

Policy N

1. Risk Governance Charter ABN AMRO (AIM 101-21-00)
2. Bank Risk Policy (AIM 101-23-40)
3. Central Credit Risk Policy (AIM 101-21-75)
4. Risk Taxonomy (AIM 101-21-01).
5. Risk Governance Charter ABN AMRO (AIM 101-21-00)
6. ABN AMRO Clearing Market & Credit Risk Policy (AIM 101-21-51
7. 'Overdispositiebeleid (AIM 101-22-08)',
8. Central Credit Risk Policy (AIM 101-21-75)
9. Central Collateral Policy (AIM 101-21-09).
10. Central Collateral Policy (AIM 101-21-09).
11. Central OBSI Credit Risk Policy (AIM 101-21-77).
12. Central Credit Risk Policy (AIM 101-21-75)
13. Central Restructuring and Forbearance Policy (AIM 101-21-17).

Policy O

1. Bank Risk Policy (AIM 101-23-40),
2. Risk Governance Charter ABN AMRO (AIM 101-21-00),
3. Central Credit Risk Policy (AIM 101-21-75)
4. Program Lending Credit Risk Policy (AIM 301-21-20).
5. Central Restructuring and Forbearance Policy (AIM 101-21-17)
6. Central Restructuring and Forbearance Policy (AIM 101-21-17)
7. Central Loan Loss Provisioning and Write-Off Policy (AIM 101-21-13)

8. Sustainability Risk Policy (AIM 101-25-00));
9. Change Risk Assessment Policy (AIM 101-23-79).
10. Program Lending Scoring Policy (AIM 301-21-35).
11. CAAML Policy (AIM 102-20-20).
12. Central Collateral Policy (AIM 101-21-09).
13. Central Collateral Policy (AIM 101-21-09).
14. Global Legal Mandate (AIM 108-55-01).
15. Policy for Capital Treatment of Securitisations of Own Originated Assets (AIM 101-21-14).
16. Central Restructuring and Forbearance Policy (AIM 101-21-17).
17. Global Legal Mandate (AIM 108-55-01)
18. Central Collateral Policy (AIM 101-21-09).
19. Central Restructuring and Forbearance Policy (AIM 101-21-17).
20. Central Restructuring and Forbearance Policy (AIM 101-21-17).
21. Risk Event Management Policy (AIM 101-23-20).
22. Security & Intelligence Management Charter (AIM 108-75-01).
23. Corporate Insurance Policy (AIM 108-50-05)
24. Central Loan Loss Provisioning and Write-off Policy (AIM 101-21-13)

Policy P

1. Risk Governance Charter ABN AMRO (AIM 101-21-00)
2. Bank Risk Policy (AIM 101-23-40)
3. Central Credit Risk Policy (AIM 101-21-75)
4. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
5. Program Lending Credit Risk Cycle Policy (AIM 301-21-19)
6. Program Lending Scoring Policy (AIM 301-21-35)
7. Program Lending Indirect Lending Policy (AIM 301-21-32).
8. Program Lending Indirect Lending Policy (AIM 301-21-32)).
9. (Central Credit Risk Policy (AIM 101-21-75
10. Risk Appetite Framework (AIM 101-21-07);
11. Risk Taxonomy (AIM 101-21-01)
12. Program Lending Credit Risk Cycle Policy (AIM 301-21-19);
13. ABN AMRO Product Approval Policy (AIM 101-23-80).
14. Risk Governance Charter ABN AMRO (AIM 101-21-00).
15. Risk Governance Charter ABN AMRO (AIM 101-21-00).
16. Risk Governance Charter ABN AMRO (AIM 101-21-00),
17. Sustainability Risk Policy (AIM 101-25-00).

Policy Q

1. Risk Governance Charter ABN AMRO (AIM 101-21-00),
2. Bank Risk Policy (AIM 101-23-40),
3. Central Credit Risk Policy (AIM 101-21-75)
4. Program Lending Credit Risk Policy (AIM 301-21-20).
5. Intermediaries Policy (AIM 102-25-35).
6. Intermediaries Policy (AIM 102-25-35)

7. ABN AMRO Product Approval Policy (AIM 101-23-80)

Policy R

1. ABN AMRO Credit Risk Model Use Framework (AIM 101-21-55).
2. Bank Risk Policy (AIM 101-23-40),
3. Risk Governance Charter ABN AMRO (AIM 101-21-00),
4. Program Lending Credit Risk Policy (AIM 301-21-20)
5. ABN AMRO Credit Risk Model Use Framework (AIM 101-21-55).
6. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
7. ABN AMRO Credit Risk Model Use Framework (AIM 101-21-55)
8. ABN AMRO Credit Risk Model Use Framework (AIM 101-21-55).
9. ABN AMRO Credit Risk Model Use Framework (AIM 101-21-55).
10. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84),

Policy S

1. Bank Risk Policy (AIM 101-23-40)
2. Risk Governance Charter ABN AMRO (AIM 101-21-00)
3. Central Credit Risk Policy (AIM 101-21-75)
4. Central Collateral Policy (AIM 101-21-09)
5. Securities Financing Risk Policy (AIM 101-21-54)
6. Central OBSI Credit Risk Policy (AIM 101-21-77)
7. Central Trading Risk Policy (AIM 104-24-02)
8. Product Approval Policy for Capital Markets Solutions & Treasury (AIM 101-21-49).
9. Risk Governance Charter ABN AMRO (AIM 101-21-00).
10. Records Management Policy (AIM 108-25-25).
11. Risk Governance Charter ABN AMRO (AIM 101-21-00).

Policy T

1. Risk Taxonomy (AIM 101-21-01).
2. Bank Risk Policy (AIM 101-23-40)
3. Program Lending Credit Risk Policy (AIM 301-21-20)
4. Risk Governance Charter ABN AMRO (AIM 101-21-00)
5. International Risk Charter (AIM 101-21-05)
6. Risk Governance Charter ABN AMRO (AIM 101-21-00).
7. Risk Appetite Policy (AIM 101-21-07).
8. Risk Appetite Policy (AIM 101-21-07).
9. Risk Governance Charter ABN AMRO (AIM 101-21-00).
10. Bank Risk Policy (AIM 101-23-40)
11. Program Lending Credit Risk Policy (AIM 301-21-20).
12. ABN AMRO Methodology on Credit Risk Measurement AIRB Approach (AIM 101-21-84).
13. Risk Taxonomy (AIM 101-21-01).
14. Country Risk Policy (AIM 101-21-20).

15. Central OBSI Credit Risk Policy (AIM 101-21-77).
16. Risk Taxonomy (AIM 101-21-01)
17. White Knight Facilities Risk Policy (AIM 101-21-23)).
18. Secured Inventory Products – Icestar Risk Policy (AIM 101-21-42).
19. ABN AMRO Clearing Credit & Market Risk Policy (AIM 101-21-51).
20. Central Trading Risk Policy (AIM 101-24-02).
21. Large Exposure Reporting Policy (AIM 104-20-25).
22. ABN AMRO Clearing Credit & Market Risk Policy (AIM 101-21-51).
23. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
24. Risk Governance Charter ABN AMRO (AIM 101-21-00)
25. International Risk Charter (AIM 101-21-05).
26. ECT/Commodities Clients Risk Policy (AIM 101-21-44)
27. ECT/Transportation Clients/ Shipping Risk Policy (AIM 101-21-39)
28. ECT/Transportation Clients/Intermodal Finance Risk Policy (AIM 101-21-32)
29. ECT/Energy Clients Risk Policy (AIM 101-21-33)
30. Global Export Finance Credit Risk Policy (AIM 101-21-90)
31. Project Finance Credit Risk Policy (AIM 101-21-91)
32. ECT/Energy Clients Risk Policy (AIM 101-21-33)
33. Project Finance Credit Risk Policy (AIM 101-21-91)
34. Financial Institutions Risk Policy (AIM 101-21-28)
35. Central Real Estate Policy (AIM 101-21-69)
36. Risk Policy Diamond & Jewellery (AIM 411-24-03)
37. Public Sector Risk Policy (AIM 101-21-35)
38. Business Risk Policy Binnenvaart (AIM 101-22-09)
39. Subordinated Debt Finance Credit Risk Policy (AIM 101-21-43)
40. Subordinated Debt Finance Credit Risk Policy (AIM 101-21-43)
41. Acquisition and Leveraged Finance Credit Risk Policy (AIM 101-21-40)
42. ABN AMRO Clearing Credit & Market Risk Policy (AIM 101-21-51)
43. Sanctions Policy (AIM 102-20-35)).
44. Sustainability Risk Policy (AIM 101-25-00).
45. Sustainability Risk Policy for Lending (AIM 101-25-02)
46. Bank Risk Policy (AIM 101-23-40)
47. Central Collateral Policy (AIM 101-21-09)
48. Central Guarantee Policy (AIM 101-21-78).
49. Risk Taxonomy (AIM 101-21-01),
50. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84)
51. Business Risk Policy Borrowing Base (AIM 101-21-86)
52. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84)
53. Risk Governance Charter ABN AMRO (AIM 101-21-00)
54. International Risk Charter (AIM 101-21-05)
55. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
56. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
57. Credit Risk Monitoring Policy (AIM 101-21-76)
58. Central Collateral Policy (AIM 101-21-09
59. Risk Governance Charter ABN AMRO (AIM 101-21-00)
60. ABN AMRO Data Quality Policy (AIM 101-23-41).
61. ABN AMRO Clearing Credit & Market Risk Policy (AIM 101-21-51)
62. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84)

63. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84).
64. Central Policy on Impairments, Impairment Allowances and Write off (AIM 101-21-13).
65. Central Restructuring and Forbearance Policy (AIM 101-21-17).
66. Central Restructuring and Forbearance Policy (AIM 101-21-17).
67. Central Restructuring and Forbearance Policy (AIM 101-21-17).
68. Risk Governance Charter ABN AMRO (AIM 101-21-00).
69. Central Policy on Impairments, Impairment Allowance and Write off (AIM 101-21-13).
70. Central Restructuring and Forbearance Policy (AIM 101-21-17).
71. ABN AMRO Clearing Credit & Market Risk Policy (AIM 101-21-51)
72. Acquisition and Leveraged Finance Credit Risk Policy (AIM 101-21-48).
73. ABN AMRO Clearing Credit & Market Risk Policy (AIM 101-21-51)
74. ABN AMRO Clearing Credit & Market Risk Policy (AIM 101-21-51).
75. Economic Capital Policy (AIM 101-22-37).
76. Country Risk Policy (AIM 101-21-20).

Policy U

1. Risk Governance Charter ABN AMRO (AIM 101-21-00
2. Bank Risk Policy (AIM 101-23-40)
3. Central Credit Risk Policy (AIM 101-21-75)
4. Market Risk in the Banking Book Policy (AIM 101-22-15)
5. Central Liquidity Risk Policy (AIM 101-22-10)
6. Funds Transfer Pricing Risk Policy (AIM 101-22-14)
7. Funding Cost Policy (AIM 101-22-11)
8. Country Risk Policy (AIM 101-21-20).
9. Policy on Issue of Intercompany Guarantees (AIM 101-15-25)
10. Central Liquidity Risk Policy (AIM 101-22-10)
11. Funding Cost Policy (AIM 101-22-11).
12. Policy on Issue of Intercompany Guarantees (AIM 101-15-25).
13. Central Liquidity Risk Policy (AIM 101-22-10).
14. Central Liquidity Risk Policy (AIM 101-22-10
15. Funding Cost Policy (AIM 101-22-11),
16. Risk Governance Charter ABN AMRO (AIM 101-21-00))
17. Central Liquidity Risk Policy (AIM 101-22-10).
18. Market Risk in the Banking Book Policy (AIM 101-22-15).
19. Central OBSI Credit Risk Policy AIM 101-21-77)).
20. Country Risk Policy (AIM 101-21-20)).
21. Funding Cost Policy (AIM 101-22-11)
22. Transfer Pricing Policy (AIM 108-05-90).
23. Risk Governance Charter ABN AMRO (AIM 101-21-00).
24. Risk Governance Charter ABN AMRO (AIM 101-21-00).
25. Credit Risk Monitoring Policy (AIM 101-21-76)
26. Central Liquidity Risk Policy (AIM 101-22-10)
27. Risk Governance Charter ABN AMRO (AIM 101-21-00)).
28. Risk Governance Charter ABN AMRO (AIM 101-21-00).
29. Risk Governance Charter ABN AMRO (AIM 101-21-00)

Policy V

1. Risk Taxonomy (AIM 101-21-01):
2. Risk Governance Charter ABN AMRO (AIM 101-21-00)
3. Bank Risk Policy (AIM 101-23-40)
4. Central Credit Risk Policy (AIM 101-21-75)
5. Credit Risk Monitoring Policy (AIM 101-21-76)
6. Program Lending Credit Risk Cycle Policy (AIM 301-21-19)
7. Central Credit Risk Policy (AIM 101-21-75).
8. Risk Governance Charter ABN AMRO (AIM 101-21-00)
9. Central Credit Risk Policy (AIM 101-21-75).
10. Central Credit Risk Policy (AIM 101-21-75).
11. Central Credit Risk Policy (AIM101-21-75)).
12. Central Credit Risk Policy (AIM 101-21-75)
13. the Credit Risk Monitoring Policy (AIM 101-21-76).
14. Credit Risk Monitoring Policy (AIM 101-21-76).
15. Debt to Equity Swap Policy (AIM 101-21-02).
16. Program Lending Credit Risk Cycle Policy (AIM 301-21-19)
17. Central Credit Risk Policy (AIM 101-21-75):
18. Central Credit Risk Policy (AIM 101-21-75).
19. Central Credit Risk Policy (AIM 101-21-75).
20. IFRS - Policy on Impairment of Financial Assets (AIM 108-07-06)
21. Central Policy on Impairments, Impairment Allowances and Write off (AIM 101-21-13).
22. Risk Governance Charter ABN AMRO (AIM 101-21-00).

Policy W

1. Central Credit Risk Policy (AIM 101-21-75)
2. Bank Risk Policy (AIM 101-23-40)
3. International Risk Allocation Policy (AIM 101-21-29)
4. Central Collateral Policy (AIM 101-21-09)
5. Central Guarantee Policy (AIM 101-21-78).
6. Minimum Requirements for On-Balance Sheet Netting Policy (AIM 101-21-26).
7. International Risk Allocation Policy (AIM 101-21-29).
8. Central Collateral Policy, AIM 101-21-09).
9. Central Collateral Policy (AIM 101-21-09)
10. Central Guarantee Policy (AIM 101-21-78).
11. ABN AMRO Methodology for Credit Risk Measurement AIRB Approach (AIM 101-21-84)
12. Sustainability Risk Policy (AIM 101-25-00)
13. Sustainability Risk Policy for Lending (AIM 101-25-02).
14. CAAML Policy (AIM 102-20-20).
15. Tax Policy (AIM 101-21-03).
16. Central Credit Risk Policy (AIM 101-21-75)
17. Credit Risk Monitoring Policy (AIM 101-21-76),

Appendix C – Reference Extraction Calculation

Excluding Partially Correct Chunks

Policy A

$$precision = \frac{11}{11+0} = 1 \qquad recall = \frac{11}{11+3+1} = 0{,}733$$

$$F = \frac{(2*1*0{,}733)}{1+0{,}733} = 0{,}85 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\cdot\frac{1\ \cdot\ 0{,}733}{(0{,}5^2\ \cdot 1)+0{,}733} = 0{,}932$$

Policy B

$$precision = \frac{15}{15+0} = 1 \qquad recall = \frac{15}{15+0+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\cdot\frac{1\ \cdot\ 1}{(0{,}5^2\ \cdot 1)+1} = 1$$

Policy C

$$precision = \frac{8}{8+0} = 1 \qquad recall = \frac{8}{8+4+1} = 0{,}615$$

$$F = \frac{(2*1*0{,}615)}{1+0{,}615} = 0{,}76 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\cdot\frac{1\ \cdot\ 0{,}615}{(0{,}5^2\ \cdot 1)+0{,}615} = 0{,}88$$

Policy D

$$precision = \frac{4}{4+0} = 1 \qquad recall = \frac{4}{4+1+6} = 0{,}363$$

$$F = \frac{(2*1*0{,}363)}{1+0{,}363} = 0{,}53 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\cdot\frac{1\ \cdot\ 0{,}363}{(0{,}5^2\ \cdot 1)+0{,}363} = 0{,}74$$

Policy E

$$precision = \frac{12}{12+2} = 0{,}85 \qquad recall = \frac{12}{12+1+0} = 0{,}92$$

$$F = \frac{(2*0{,}85*0{,}92)}{0{,}85+0{,}92} = 0{,}89 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\cdot\frac{0{,}85+0{,}92}{(0{,}5^2\ \cdot 0{,}85)+0{,}92} = 0{,}86$$

Policy F

$$precision = \frac{13}{13+0} = 1 \qquad recall = \frac{13}{13+4+1} = 0{,}722$$

$$F = \frac{(2*1*0{,}722)}{1+0{,}722} = 0{,}84 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\cdot\frac{1+0{,}722}{(0{,}5^2\ \cdot 1)+0{,}722} = 0{,}92$$

Policy G

$$precision = \frac{12}{12+1} = 0{,}92 \qquad recall = \frac{12}{12+2+0} = 0{,}85$$

$$F = \frac{(2*0{,}92*0{,}85)}{0{,}85+0{,}92} = 0{,}89 \qquad F_{0,5} = (1+\ 0{,}5^2)\ \cdot\ \frac{0{,}92+0{,}85}{(0{,}5^2\ \cdot 0{,}92)+0{,}85} = 0{,}90$$

Policy H

$$precision = \frac{11}{11+0} = 1 \qquad recall = \frac{11}{11+0+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0,5} = (1+\ 0{,}5^2)\ \cdot\ \frac{1+1}{(0{,}5^2\ \cdot 1)+1} = 1$$

Policy I

$$precision = \frac{21}{21+0} = 1 \qquad recall = \frac{21}{21+5+1} = 0{,}777$$

$$F = \frac{(2*1*0{,}777)}{1+0{,}777} = 0{,}88 \qquad F_{0,5} = (1+\ 0{,}5^2)\ \cdot\ \frac{1+0{,}777}{(0{,}5^2\ \cdot 1)+0{,}777} = 0{,}94$$

Policy J

$$precision = \frac{26}{26+0} = 1 \qquad recall = \frac{26}{26+4+1} = 0{,}838$$

$$F = \frac{(2*1*0{,}838)}{1+0{,}838} = 0{,}91 \qquad F_{0,5} = (1+\ 0{,}5^2)\ \cdot\ \frac{1+0{,}838}{(0{,}5^2\ \cdot 1)+0{,}838} = 0{,}96$$

Policy K

$$precision = \frac{6}{6+0} = 1 \qquad recall = \frac{6}{6+0+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0,5} = (1+\ 0{,}5^2)\ \cdot\ \frac{1+1}{(0{,}5^2\ \cdot 1)+1} = 1$$

Policy L

$$precision = \frac{14}{14+0} = 1 \qquad recall = \frac{14}{14+6+0} = 0{,}7$$

$$F = \frac{(2*1*0{,}7)}{1+0{,}7} = 0{,}82 \qquad F_{0,5} = (1+\ 0{,}5^2)\ \cdot\ \frac{1+0{,}7}{(0{,}5^2\ \cdot 1)+0{,}7} = 0{,}92$$

Policy M

$$precision = \frac{11}{11+0} = 1 \qquad recall = \frac{11}{11+2+0} = 0{,}84$$

$F = \frac{(2 * 1 * 0{,}84)}{1 + 0{,}84} = 0{,}92$ $\quad F_{0{,}5} = (1 + \ 0{,}5^2) \ \cdot \ \frac{0{,}84 + 0{,}92}{(0{,}5^2 \ \cdot 0{,}84) + 0{,}92} = 0{,}96$

Policy N

$precision = \frac{13}{13 + 0} = 1$ $\quad recall = \frac{13}{13 + 0 + 0} = 1$

$F = \frac{(2 * 1 * 1)}{1 + 1} = 1$ $\quad F_{0{,}5} = (1 + \ 0{,}5^2) \ \cdot \ \frac{1 + 1}{(0{,}5^2 \ \cdot 1) + 1} = 1$

Policy O

$precision = \frac{22}{22 + 0} = 1$ $\quad recall = \frac{22}{22 + 2 + 0} = 0{,}91$

$F = \frac{(2 * 1 * 0{,}91)}{1 + 0{,}91} = 0{,}96$ $\quad F_{0{,}5} = (1 + \ 0{,}5^2) \ \cdot \ \frac{1 + 0{,}91}{(0{,}5^2 \ \cdot 1) + 0{,}91} = 0{,}98$

Policy P

$precision = \frac{14}{12 + 0} = 1$ $\quad recall = \frac{14}{14 + 0 + 3} = 0{,}82$

$F = \frac{(2 * 1 * 0{,}82)}{1 + 0{,}82} = 0{,}90$ $\quad F_{0{,}5} = (1 + \ 0{,}5^2) \ \cdot \ \frac{1 + 0{,}82}{(0{,}5^2 \ \cdot 1) + 0{,}82} = 0{,}95$

Policy Q

$precision = \frac{8}{8 + 0} = 1$ $\quad recall = \frac{8}{8 + 0 + 0} = 1$

$F = \frac{(2 * 1 * 1)}{1 + 1} = 1$ $\quad F_{0{,}5} = (1 + \ 0{,}5^2) \ \cdot \ \frac{1 + 1}{(0{,}5^2 \ \cdot 1) + 1} = 1$

Policy R

$precision = \frac{10}{10 + 0} = 1$ $\quad recall = \frac{10}{10 + 0 + 0} = 1$

$F = \frac{(2 * 1 * 1)}{1 + 1} = 1$ $\quad F_{0{,}5} = (1 + \ 0{,}5^2) \ \cdot \ \frac{1 + 1}{(0{,}5^2 \ \cdot 1) + 1} = 1$

Policy S

$precision = \frac{10}{10 + 0} = 1$ $\quad recall = \frac{10}{10 + 1 + 0} = 0{,}90$

$F = \frac{(2 * 1 * 0{,}90)}{1 + 0{,}90} = 0{,}95$ $\quad F_{0{,}5} = (1 + \ 0{,}5^2) \ \cdot \ \frac{1 + 0{,}90}{(0{,}5^2 \ \cdot 1) + 0{,}90} = 0{,}98$

Policy T

$precision = \frac{74}{74+1} = 0{,}98$ $recall = \frac{74}{74+1+2} = 0{,}96$

$F = \frac{(2 * 0{,}98 * 0{,}96)}{0{,}98 + 0{,}96} = 0{,}97$ $F_{0,5} = (1 + \ 0{,}5^2) \cdot \frac{0{,}98 + 0{,}96}{(0{,}5^2 \ \cdot 0{,}98) + 0{,}96} = 0{,}981$

Policy U

$precision = \frac{21}{21+0} = 1$ $recall = \frac{21}{21+6+2} = 0{,}72$

$F = \frac{(2 * 1 * 0{,}72)}{1 + 0{,}72} = 0{,}84$ $F_{0,5} = (1 + \ 0{,}5^2) \cdot \frac{1 + 0{,}72}{(0{,}5^2 \ \cdot 1) + 0{,}72} = 0{,}92$

Policy V

$precision = \frac{16}{16+0} = 1$ $recall = \frac{16}{16+2+4} = 0{,}727$

$F = \frac{(2 * 1 * 0{,}727)}{1 + 0{,}727} = 0{,}84$ $F_{0,5} = (1 + \ 0{,}5^2) \cdot \frac{1 + 0{,}727}{(0{,}5^2 \ \cdot 1) + 0{,}727} = 0{,}93$

Policy W

$precision = \frac{16}{16+0} = 1$ $recall = \frac{16}{16+0+1} = 0{,}94$

$F = \frac{(2 * 1 * 0{,}94)}{1 + 0{,}94} = 0{,}97$ $F_{0,5} = (1 + \ 0{,}5^2) \cdot \frac{1 + 0{,}94}{(0{,}5^2 \ \cdot 1) + 0{,}94} = 0{,}98$

Including Partially Correct Chunks

Policy A

$precision = \frac{11+3}{11+3+0} = 1$ $recall = \frac{11+3}{11+3+1} = 0{,}933$

$F = \frac{(2 * 1 * 0{,}933)}{1 + 0{,}933} = 0{,}97$ $F_{0,5} = (1 + \ 0{,}5^2) \cdot \frac{1 \ \cdot \ 0{,}933}{(0{,}5^2 \ \cdot 1) + 0{,}933} = 0{,}98$

Policy B

$precision = \frac{15}{15+0} = 1$ $recall = \frac{15}{15+0+0} = 1$

$F = \frac{(2 * 1 * 1)}{1 + 1} = 1$ $F_{0,5} = (1 + \ 0{,}5^2) \cdot \frac{1 \ \cdot \ 1}{(0{,}5^2 \ \cdot 1) + 1} = 1$

Policy C

$precision = \frac{8+4}{8+4+0} = 1$ $recall = \frac{8}{8+4+1} = 0{,}92$

$F = \frac{(2 * 1 * 0{,}92)}{1 + 0{,}92} = 0{,}96$ $F_{0{,}5} = (1 + \ 0{,}5^2) \cdot \frac{1 \ \cdot 0{,}92}{(0{,}5^2 \ \cdot 1) + 0{,}92} = 0{,}98$

Policy D

$precision = \frac{4 + 1}{4 + 1 + 0} = 1$ $recall = \frac{4 + 1}{4 + 1 + 6} = 0{,}45$

$F = \frac{(2 * 1 * 0{,}45)}{1 + 0{,}45} = 0{,}63$ $F_{0{,}5} = (1 + \ 0{,}5^2) \cdot \frac{1 \ \cdot 0{,}45}{(0{,}5^2 \ \cdot 1) + 0{,}45} = 0{,}80$

Policy E

$precision = \frac{12 + 1}{12 + 1 + 2} = 0{,}86$ $recall = \frac{12 + 1}{12 + 1 + 0} = 1$

$F = \frac{(2 * 0{,}86 * 1)}{0{,}86 + 1} = 0{,}93$ $F_{0{,}5} = (1 + \ 0{,}5^2) \cdot \frac{1 + 0{,}93}{(0{,}5^2 \ \cdot 1) + 0{,}93} = 0{,}89$

Policy F

$precision = \frac{13 + 4}{13 + 4 + 0} = 1$ $recall = \frac{13 + 4}{13 + 4 + 1} = 0{,}944$

$F = \frac{(2 * 1 * 0{,}944)}{1 + 0{,}944} = 0{,}97$ $F_{0{,}5} = (1 + \ 0{,}5^2) \cdot \frac{1 + 0{,}944}{(0{,}5^2 \ \cdot 1) + 0{,}944} = 0{,}98$

Policy G

$precision = \frac{12 + 2}{12 + 2 + 1} = 0{,}93$ $recall = \frac{12 + 2}{12 + 2 + 0} = 1$

$F = \frac{(2 * 0{,}93 * 1)}{0{,}93 + 1} = 0{,}97$ $F_{0{,}5} = (1 + \ 0{,}5^2) \cdot \frac{0{,}93 + 1}{(0{,}5^2 \ \cdot 0{,}93) + 1} = 0{,}94$

Policy H

$precision = \frac{11 + 0}{11 + 0 + 0} = 1$ $recall = \frac{11}{11 + 0} = 1$

$F = \frac{(2 * 1 * 1)}{1 + 1} = 1$ $F_{0{,}5} = (1 + \ 0{,}5^2) \cdot \frac{1 + 1}{(0{,}5^2 \ \cdot 1) + 1} = 1$

Policy I

$precision = \frac{21 + 5}{21 + 5 + 0} = 1$ $recall = \frac{21 + 5}{21 + 5 + 1} = 0{,}96$

$F = \frac{(2 * 1 * 0{,}96)}{1 + 0{,}96} = 0{,}98$ $F_{0{,}5} = (1 + \ 0{,}5^2) \cdot \frac{1 + 0{,}96}{(0{,}5^2 \ \cdot 1) + 0{,}96} = 0{,}99$

Policy J

$$precision = \frac{26+4}{26+4+0} = 1 \qquad recall = \frac{26+4}{26+4+1} = 0{,}96$$

$$F = \frac{(2*1*0{,}96)}{1+0{,}96} = 0{,}98 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\ \cdot \frac{1+0{,}96}{(0{,}5^2\ \cdot 1)+0{,}96} = 0{,}99$$

Policy K

$$precision = \frac{6+0}{6+0+0} = 1 \qquad recall = \frac{6+0}{6+0+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\ \cdot \frac{1+1}{(0{,}5^2\ \cdot 1)+1} = 1$$

Policy L

$$precision = \frac{14+6}{14+6+0} = 1 \qquad recall = \frac{14+6}{14+6+0} = 1$$

$$F = \frac{(2*1*1)}{1+0{,}7} = 1 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\ \cdot \frac{1+1}{(0{,}5^2\ \cdot 1)+1} = 1$$

Policy M

$$precision = \frac{11+2}{11+2+0} = 1 \qquad recall = \frac{11+2}{11+2+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\ \cdot \frac{1+1}{(0{,}5^2\ \cdot 1)+1} = 1$$

Policy N

$$precision = \frac{13+0}{13+0+0} = 1 \qquad recall = \frac{13+0}{13+0+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\ \cdot \frac{1+1}{(0{,}5^2\ \cdot 1)+1} = 1$$

Policy O

$$precision = \frac{22+2}{22+2+0} = 1 \qquad recall = \frac{22+2}{22+2+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\ \cdot \frac{1+1}{(0{,}5^2\ \cdot 1)+1} = 1$$

Policy P

$$precision = \frac{14+0}{14+0+0} = 1 \qquad recall = \frac{14+0}{14+0+2} = 0{,}82$$

$$F = \frac{(2*1*0{,}82)}{1+0{,}82} = 0{,}90 \qquad F_{0{,}5} = (1+\ 0{,}5^2)\ \cdot \frac{1+0{,}82}{(0{,}5^2\ \cdot 1)+0{,}82} = 0{,}95$$

Policy Q

$$precision = \frac{8+0}{8+0+0} = 1 \qquad recall = \frac{8+0}{8+0+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0,5} = (1+\ 0,5^2)\ \cdot \frac{1+1}{(0,5^2\ \cdot 1)+1} = 1$$

Policy R

$$precision = \frac{10+0}{10+0+0} = 1 \qquad recall = \frac{10+0}{10+0+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0,5} = (1+\ 0,5^2)\ \cdot \frac{1+1}{(0,5^2\ \cdot 1)+1} = 1$$

Policy S

$$precision = \frac{10+1}{10+1+0} = 1 \qquad recall = \frac{10+1}{10+1+0} = 1$$

$$F = \frac{(2*1*1)}{1+1} = 1 \qquad F_{0,5} = (1+\ 0,5^2)\ \cdot \frac{1+1}{(0,5^2\ \cdot 1)+1} = 1$$

Policy T

$$precision = \frac{74+1}{74+1+1} = 0,9 \qquad recall = \frac{74+1}{74+1+2} = 0,97$$

$$F = \frac{(2*0,98*0,97)}{0,98+0,97} = 0,98 \qquad F_{0,5} = (1+\ 0,5^2)\ \cdot \frac{0,98+0,97}{(0,5^2\ \cdot 0,98)+0,97} = 0,984$$

Policy U

$$precision = \frac{21+6}{21+6+0} = 1 \qquad recall = \frac{21+6}{21+6+2} = 0,93$$

$$F = \frac{(2*1*0,93)}{1+0,72} = 0,96 \qquad F_{0,5} = (1+\ 0,5^2)\ \cdot \frac{1+0,93}{(0,5^2\ \cdot 1)+0,93} = 0,98$$

Policy V

$$precision = \frac{16+2}{16+2+0} = 1 \qquad recall = \frac{16+2}{16+2+4} = 0,81$$

$$F = \frac{(2*1*0,81)}{1+0,81} = 0,90 \qquad F_{0,5} = (1+\ 0,5^2)\ \cdot \frac{1+0,81}{(0,5^2\ \cdot 1)+0,81} = 0,95$$

Policy W

$$precision = \frac{16+0}{16+0+0} = 1 \qquad recall = \frac{16+0}{16+0+1} = 0,94$$

$$F = \frac{(2 * 1 * 0{,}94)}{1 + 0{,}94} = 0{,}97 \qquad F_{0{,}5} = (1 + \ 0{,}5^2) \ \cdot \ \frac{1 + 0{,}94}{(0{,}5^2 \ \cdot 1) + 0{,}94} = 0{,}98$$

Appendix D – Expert Evaluation

Policy A

Policy Details	
Policy Title:	Policy A
Date:	10_6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. valuation/appraisal/review	Finance
2. relationship	Real Estate
3. counterparty	
4. representative	
5. international	
6. sub-department	
7. introduction	
8. construction	

Algorithm Summary	Comments
these include the united nations environment programme finance initiative (unep fi), the ten principles of the un global compact, the principles for responsible investments (pri), the equator principles, the oecd guidelines for multinational enterprises, the un guiding principles for business and human rights, and the ilo tripartite declaration of principles concerning multinational enterprises and social policy. the bank's sustainability risk management framework consists of operational and sector policies, guidelines and tools through which the bank assures itself that the sustainability risks of its activities are adequately identified, analysed, mitigated, managed, monitored and reported, in accordance with the bank's sustainability risk principles. 3. the bank strives for an inclusive approach and will enter into a dialogue with its business relations abn amro believes that a good sustainability performance results in better corporate performance for its business relations, and the bank aims to help these partners in addressing those risks and opportunities relevant to them. abn amro bank aims to have a prominent position as a sustainable bank that takes responsibility for its actions and engagements, as a member of society with its own impact on the planet and on people, but also as a provider of financial services with an indirect impact through the activities of its clients and investments. 1. sustainability risk management is a driver for quality improvement abn amro is convinced that a structural incorporation of sustainability risk management in the bank's decision-making processes genuinely improves the quality of the bank's business and enables the bank to make	Think too detailed for intro.

Policy B

Policy Details	
Policy Title:	Policy B
Date:	10_6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. product-specific documentation 2. multicurrency 3. sustainability 4. crd-framework 5. representative 6. calculations 7. jurisdictions 8. back-to-back 9. compensation 10. client-oriented 11. risk-weighted 12. market-driven **13.** documentation	

Algorithm Summary	Comments
	the summary is a moderate representation of the document

Policy C

Policy Details	
Policy Title:	Policy C
Date:	28/5/2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. external/inferred 2. over-the-counter 3. responsibility 4. identification 5. non-compliance 6. differently 7. disclosures 8. post-enforcement 9. non-securitisation	Standardised Internal Ratings Based Risk exposure

10. securitisation 11. mark-to-market 12. sub-participation 13. requirements 14. interpretation 15. organisational 16. administrative 17. re-securitisation 18. implementation 19. currencies/exchange **20.** applicability securitisation	

Algorithm Summary	Comments
	The summary covers more detailed elements rather than just providing a good idea of what the document contains.

Policy D

Policy Details	
Policy Title:	Policy D
Date:	28/5/2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. external/inferred 2. pre-securitisation 3. over-the-counter 4. securitization 5. corresponding 6. stratification 7. post-securitisation 8. methodology 9. identification 10. quantification 11. risk-weighted 12. re-securitisation 13. unsecuritised 14. consolidation 15. non-securitised	Hierarchy of methods Tranche

16. subordination 17. mark-to-market 18. sub-participation 19. effectiveness 20. organizational 21. non-granular 22. calculations 23. requirements 24. administrative 25. comprehensive 26. applicability securitisation 27. implementation 28. currencies/exchange **29.** documentation	

Algorithm Summary	Comments
	Provides a definition but the text selected -→ red too detailed. Part of formula

Policy E

Policy Details	
Policy Title:	Policy E
Date:	28/5/2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. responsibility 2. guarantee-like 3. counter-guarantor 4. relationship/group 5. protection/investment 6. insured-of-record 7. facility/these 8. maturity/termination 9. support/comfort/intent/awareness 10. quasi-governmental 11. facility/exposure 12. government/bank 13. facility-specific 14. clearly-defined 15. counter-guarantee **16.** post-credit-eventg	If you ask me to define what guarantees from Basel perspective are about and what is the risk and it's mitigants I'd use these words. They might not be in the policy: - Risk mitigation - Any and all amounts - Direct - Explicit - Irrevocable - Tenor- - Maturity mismatch

Algorithm Summary	Comments
	Good summary. Captured the most important thing. How is it viewed by the regulator and what are the most important forms. Would be good if it had the most important requirements for eligibility in there because that is one of the main purposes of the document. These are the headings. 4.1.1 to 4.1.7: Direct Explicit Enforceable etc

Policy F

Policy Details	
Policy Title:	Policy F
Date:	28/5/2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. responsibility 2. non-performance 3. classification 4. netting-friendly 5. counterparty/client 6. classification/treatment professional/non-professional 7. non-negotiable 8. appropriateness 9. shareholder/owner 10. collateralisation	OBSI Counterparty credit risk

11. non-professional 12. mark-to-market 13. creditworthiness **14.** locked-to-exposure	

Algorithm Summary	Comments
	Good summary. Captured the most important thing.

Policy G

Policy Details	
Policy Title:	Policy G
Date:	28/5/2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. responsibility 2. non-predictable 3. professional/eligible 4. professional/non-professional 5. non-professional 6. mark-to-market 7. collateralisation 8. uncollateralised 9. execution-only **10.** non-collateralised	Counterparty Credit Risk

Algorithm Summary	Comments
	Good summary. Captured the most important thing.

Policy H

Policy Details	
Policy Title:	Policy H
Date:	16_6_2017

Reviewer:	M de Vries

Algorithm Keywords	Comments
1. responsibility 2. operationalization 3. element/principle 4. approval/sign-off 5. confidentiality 6. responsibility identification 7. communication 8. well-balanced 9. applicability 10. representative 11. note/executive 12. misinterpretation 13. effectiveness 14. implementation 15. understanding 16. non-significant 17. appetite/strategy 18. summary/cover **19.** product/service	

Algorithm Summary	Comments
	the summary is a moderate representation of the document

Policy I

Policy Details	
Policy Title:	Policy I
Date:	28/5/2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. cross-currency 2. classification 3. capitalisation 4. explicit/hard 5. representative 6. authorisation 7. country-specific	Matched book Liquidity Buffer Leverage Ratio

8. responsibility 9. counterparty-originated 10. un-correlated 11. identification 12. deterioration 13. sustainability 14. pre-specified 15. sub-investment 16. enforceability 17. administration 18. infrastructure 19. pre-condition 20. liquifiability 21. treasury/cms 22. macro-economic 23. concentration 24. collateralisation 25. incorporation 26. issuer/custodian 27. comprehensive 28. pre-determined **29.** documentation	

Algorithm Summary	Comments
	Not very descriptive wrt policy. Selected too granular rules to really capture a good summary of what the policy is about.

<u>Policy J</u>

Policy Details	
Policy Title:	Policy J
Date:	12-6-2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. complementary	Capital reduction
2. inventory/stock	Credit risk mitigation
3. environmental	Valuation
4. non-residential	Haircut

5. representative 6. non-investment 7. assignment/encumbrance 8. creditworthiness 9. marketability 10. fashion-sensitive 11. authorisation 12. responsibility 13. identification 14. deterioration 15. clients/transportation 16. receivables 17. back-to-back 18. effectiveness enforceability 19. supplementary 20. cash/deposits/cds 21. non-affiliated 22. enforceability 23. marked-to-market 24. administration 25. recommendation 26. securitisation 27. characteristic 28. apportionment 29. non-eligibility 30. non-marketable 31. transportation 32. diversification 33. establishment 34. semi-finished 35. company specific 36. non-collateralised 37. collateralisation 38. acceptability 39. non-physical 40. inter-company 41. comprehensive 42. loan/exposure 43. commodity-type 44. specification documentation 45. manufacturing 46. work-in-progress 47. restructuring 48. documentation 49. transactional 50. group-company **51.** machinery/equipment	Recovery rate - egally valid, duly perfected, enforceable priority right security right

Algorithm Summary	Comments
	the summary is a moderate representation of the document

Policy K

Policy Details	
Policy Title:	Policy K
Date:	12-6-2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. prevent/postpone 2. distributable 3. communication 4. subordination 5. sustainability 6. representation 7. non-profitability 8. representative 9. non-repayment 10. effectiveness 11. non-financial 12. unconditional 13. non-withdrawal **14.** insolvency/illiquidity	If you ask me to define what white knight facilities are about and what is the risk and it's mitigants I'd use these words. They might not be in the policy: - Legal risk - Reputational risk - Case by case analysis - Tailormade solution - CCC approval - One draw down

Algorithm Summary	Comments
	Good summary. Captured the most important thing: what is it and how do we want it to look (1 drawdown, maybe 2) There is some repetition here, that could be improved. Also good to add some more about

	the risks involved and mitigation. Meaning: Risk = legal and reputational risk. Mitigant: always approval by legal

Policy L

Policy Details	
Policy Title:	Policy L
Date:	16_6_2017
Reviewer:	M de Vries

Algorithm Keywords	**Comments**
1. responsibility 2. recommendation 3. accountability 4. capital-usage 5. letter/indemnity 6. representative 7. subsidiary-facing 8. credit-substitution 9. limit-increase 10. subsidiary-facility 11. company-facility 12. aforementioned **13.** counter-indemnity	Collateral Back-to-back facility

Algorithm Summary	**Comments**
	the summary is a moderate representation of the document

Policy M

Policy Details	
Policy Title:	Policy M

Date:	16_6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. sub-department restructuring 2. consolidation 3. debt-to-equity 4. derecognition 5. equity-like 6. participations/abn 7. representation 8. non-consensual 9. sub-department 10. restructuring 11. documentation **12.** consideration	

Algorithm Summary	Comments
	the summary is a moderate representation of the document

Policy N

Policy Details	
Policy Title:	Policy N
Date:	16__6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. responsibility 2. excess/limietoverschrijdingen 3. perfect/imperfect 4. identification 5. comprehensive 6. representative 7. creditworthiness 8. watch-functionality 9. effectiveness 10. permanent/continuous	

11. reviews/overdue 12. asset-by-asset 13. debt/ongeregeld	

Algorithm Summary	Comments
	the summary is a moderate representation of the document

Policy O

Policy Details	
Policy Title:	Policy O
Date:	16_6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. trader/proprietor/free 2. re-instatement/termination 3. risk-conscious 4. dealer/merchant 5. post-approval 6. loss-incurring 7. misappropriation 8. reproducibility 9. creditworthiness 10. champion/challenger 11. prioritisation 12. non-ambiguous 13. cost-effective 14. responsibility 15. cards/personal identification 16. identification 17. socio-economic 18. sustainability 19. legal/regulatory 20. cradle-to-grave 21. approval/decline 22. quantification 23. cost-justified 24. enforceability 25. responsiveness 26. product-planning 27. attorney/agency	

28. interdependent 29. product-specific 30. bankruptcy/skip/fraud 31. anti-discrimination 32. repossession/foreclosure 33. product/customer 34. implementation 35. decision-making	

Algorithm Summary	Comments
	summary covers the main aspects of the policy

Policy P

Policy Details	
Policy Title:	Policy P
Date:	12-06-2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. responsibility 2. predictability 3. review/advice 4. approval/risk 5. sustainability 6. representative 7. cross-product 8. semi-automated 9. implementation 10. decision-making 11. standardisation 12. abovementioned 13. product/business	

Algorithm Summary	Comments
	somewhat excellent summary covers the main aspects of the policy

	Add: The PL business covers lending to: □ Private individuals (typically referred to as consumers); □ Small and Medium-sized Enterprises (SME).

Policy Q

Policy Details	
Policy Title:	Policy Q
Date:	12-06-2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. responsibility 2. intermediary/broker 3. post-approval 4. representative 5. prevention/control 6. effectiveness 7. implementation 8. reputation/financial 9. aforementioned 10. impact/exposure	

Algorithm Summary	Comments
	summary covers the main aspects of the policy

Policy R

Policy Details	
Policy Title:	Policy R
Date:	12-06-2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. responsibility 2. accountability 3. classification 4. facility/line 5. communication 6. program/business 7. representative 8. implementation 9. non-delinquent 10. accept/reject 11. determination	Application scoring models; Behavioural scoring models; Other models (e.g. expert scoring models). Model Scoring variables Parameters

Algorithm Summary	Comments
	The summary covers more detailed elements rather than just providing a good idea of what the document contains.

Policy S

Policy Details	
Policy Title:	Policy S
Date:	16_6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. charters/constitutional 2. 4-eyes/6-eyes 3. enforceability interpretation 4. operations/back 5. documentation	

6. convention-cadre 7. representative 8. relationship/business 9. mark-to-market 10. over-the-counter 11. cross-defaults 12. non-defaulting 13. reorganisation 14. proceedings/litigation 15. marked-to-market 16. constitutional 17. collateralisation	

Algorithm Summary	Comments
	The summary covers more detailed elements rather than just providing a good idea of what the document contains.

Policy T

Policy Details	
Policy Title:	Policy T
Date:	16-6-2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. sector-specific sustainability 2. interdependency 3. classification 4. capitalisation 5. representative 6. materialisation 7. creditworthiness 8. securitization 9. non-governmental 10. collateralised 11. country/geographic 12. originator/servicer 13. responsibility 14. relationship	

15. self-governance 16. ect/transportation clients/intermodal 17. corporations 18. ect/transportation 19. facility/product 20. sub-consolidated 21. non-negotiable 22. sustainability 23. non-retail/npl 24. money-laundering 25. counterparty/facility 26. jurisdictional 27. retail/non-retail 28. unused/undrawn 29. exit/termination 30. government-owned 31. country/region 32. non-financing 33. transportation 34. sub-participation 35. organisational 36. non-communicated 37. macroeconomic self-governance 38. post-forbearance 39. government-related 40. products-icestar 41. requirements 42. intermediation 43. business-specific 44. non-performing 45. decision-making 46. implementation 47. pre-determined 48. notwithstanding	

Algorithm Summary	Comments
	The summary covers more detailed elements rather than just providing a good idea of what the document contains.

Policy U

Policy Details	
Policy Title:	Policy U
Date:	16_6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. recommendation 2. responsibility 3. consolidation 4. calculations 5. intercompany-credit 6. increased/reduced 7. concentration 8. implementation 9. opinion/approval 10. entity/branch 11. requirements 12. documentation 13. banking/international	

Algorithm Summary	Comments
	the summary is a moderate representation of the document

Policy V

Policy Details	
Policy Title:	Policy V
Date:	16_6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. restructuring 2. classification 3. qualification 4. representative 5. change/restructure 6. modification/refinance	

7. interest-free 8. pre-assessment 9. responsibility 10. identification 11. dnb/eba/esma 12. contract/counterparty 13. debt-to-equity 14. non-performing qualification 15. contract/client 16. non-performing 17. counterpartys 18. visualisation 19. non-commercial	

Algorithm Summary	Comments
	The summary covers more detailed elements rather than just providing a good idea of what the document contains.

Policy W

Policy Details	
Policy Title:	Policy W
Date:	16_6_2017
Reviewer:	M de Vries

Algorithm Keywords	Comments
1. conditions/criteria back-to-back 2. objectionable 3. sustainability 4. representative 5. back-to-back 6. international 7. borrower/counterparty 8. implementation 9. background/motivation back-to-back 10. documentation 11. counter-guarantee	

Algorithm Summary	Comments

	the summary is a moderate representation of the document

Appendix E – Keyword Evaluation

Entire List of Keywords

Policy A

$$precision = \frac{4}{4+4} = 0{,}50 \qquad recall = \frac{4}{4+2} = 0{,}67$$

$$F = \frac{(2 * 0{,}50 * 0{,}67)}{0{,}50 + 0{,}67} = 0{,}57$$

Policy B

$$precision = \frac{8}{8+5} = 0{,}62 \qquad recall = \frac{8}{8+0} = 1$$

$$F = \frac{(2 * 0{,}62 * 1)}{0{,}62 + 1} = 0{,}76$$

Policy C

$$precision = \frac{18}{18+2} = 0{,}90 \qquad recall = \frac{18}{18+3} = 0{,}86$$

$$F = \frac{(2 * 0{,}90 * 0{,}86)}{0{,}90 + 0{,}86} = 0{,}88$$

Policy D

$$precision = \frac{23}{23+6} = 0{,}79 \qquad recall = \frac{23}{23+2} = 0{,}92$$

$$F = \frac{(2 * 0{,}79 * 0{,}92)}{0{,}79 + 0{,}92} = 0{,}85$$

Policy E

$$precision = \frac{12}{12+4} = 0{,}75 \qquad recall = \frac{12}{12+7} = 0{,}63$$

$$F = \frac{(2 * 0{,}75 * 0{,}63)}{0{,}75 + 0{,}63} = 0{,}69$$

Policy F

$$precision = \frac{10}{10+4} = 0{,}71 \qquad recall = \frac{10}{10+2} = 0{,}83$$

$$F = \frac{(2 * 0{,}71 * 0{,}83)}{0{,}71 + 0{,}83} = 0{,}77$$

Policy G

$precision = \frac{8}{8+2} = 0{,}80$ $recall = \frac{8}{8+1} = 0{,}89$

$$F = \frac{(2 * 0{,}80 * 0{,}89)}{0{,}80 + 0{,}89} = 0{,}84$$

Policy H

$precision = \frac{11}{11+8} = 0{,}58$ $recall = \frac{11}{11+0} = 1$

$$F = \frac{(2 * 0{,}58 * 1)}{0{,}58 + 1} = 0{,}73$$

Policy I

$precision = \frac{24}{24+5} = 0{,}83$ $recall = \frac{24}{24+3} = 0{,}89$

$$F = \frac{(2 * 0{,}83 * 0{,}89)}{0{,}83 + 0{,}89} = 0{,}86$$

Policy J

$precision = \frac{43}{43+8} = 0{,}84$ $recall = \frac{43}{43+10} = 0{,}81$

$$F = \frac{(2 * 0{,}84 * 0{,}81)}{0{,}84 + 0{,}81} = 0{,}83$$

Policy K

$precision = \frac{10}{10+4} = 0{,}71$ $recall = \frac{10}{10+6} = 0{,}63$

$$F = \frac{(2 * 0{,}71 * 0{,}63)}{0{,}71 + 0{,}63} = 0{,}67$$

Policy L

$precision = \frac{10}{10+3} = 0{,}77$ $recall = \frac{10}{10+2} = 0{,}83$

$$F = \frac{(2 * 0{,}77 * 0{,}83)}{0{,}77 + 0{,}83} = 0{,}80$$

Policy M

$precision = \frac{11}{11+1} = 0{,}92$ $recall = \frac{11}{11+0} = 1$

$$F = \frac{(2 * 0{,}92 * 1)}{0{,}92 + 1} = 0{,}96$$

Policy N

$$precision = \frac{11}{13 + 2} = 0{,}85 \qquad recall = \frac{11}{11 + 0} = 1$$

$$F = \frac{(2 * 0{,}85 * 1)}{0{,}85 + 1} = 0{,}92$$

Policy O

$$precision = \frac{29}{29 + 6} = 0{,}83 \qquad recall = \frac{29}{29 + 0} = 1$$

$$F = \frac{(2 * 0{,}83 * 1)}{0{,}83 + 1} = 0{,}91$$

Policy P

$$precision = \frac{12}{12 + 1} = 0{,}92 \qquad recall = \frac{12}{12 + 0} = 1$$

$$F = \frac{(2 * 0{,}92 * 1)}{0{,}92 + 1} = 0{,}96$$

Policy Q

$$precision = \frac{7}{7 + 3} = 0{,}70 \qquad recall = \frac{7}{7 + 1} = 1$$

$$F = \frac{(2 * 0{,}70 * 1)}{0{,}70 + 1} = 0{,}82$$

Policy R

$$precision = \frac{9}{9 + 2} = 0{,}82 \qquad recall = \frac{9}{9 + 6} = 0{,}60$$

$$F = \frac{(2 * 0{,}82 * 0{,}60)}{0{,}82 + 0{,}60} = 0{,}69$$

Policy S

$$precision = \frac{16}{16 + 1} = 0{,}94 \qquad recall = \frac{16}{16 + 0} = 1$$

$$F = \frac{(2 * 0{,}94 * 1)}{0{,}94 + 1} = 0{,}97$$

Policy T

$$precision = \frac{40}{40 + 8} = 0{,}83 \qquad recall = \frac{40}{40 + 1} = 1$$

$$F = \frac{(2 * 0{,}83 * 1)}{0{,}83 + 1} = 0{,}91$$

Policy U

$$precision = \frac{12}{12+1} = 0{,}92 \qquad recall = \frac{12}{21+0} = 1$$

$$F = \frac{(2 * 0{,}92 * 1)}{0{,}92 + 1} = 0{,}96$$

Policy V

$$precision = \frac{16}{16+3} = 0{,}84 \qquad recall = \frac{16}{16+0} = 1$$

$$F = \frac{(2 * 0{,}84 * 1)}{0{,}84 + 1} = 0{,}91$$

Policy W

$$precision = \frac{9}{9+2} = 0{,}84 \qquad recall = \frac{9}{9+0} = 1$$

$$F = \frac{(2 * 0{,}84 * 1)}{0{,}84 + 1} = 0{,}90$$

Cut-off list of Keywords

Policy A

$$precision = \frac{4}{4+4} = 0{,}50 \qquad recall = \frac{4}{4+2} = 0{,}67$$

$$F = \frac{(2 * 0{,}50 * 0{,}67)}{0{,}50 + 0{,}67} = 0{,}57$$

Policy B

$$precision = \frac{8}{8+5} = 0{,}62 \qquad recall = \frac{8}{8+0} = 1$$

$$F = \frac{(2 * 0{,}62 * 1)}{0{,}62 + 1} = 0{,}76$$

Policy C

$$precision = \frac{11}{11+2} = 0{,}85 \qquad recall = \frac{11}{11+3} = 0{,}79$$

$$F = \frac{(2 * 0{,}85 * 0{,}79)}{0{,}85 + 0{,}79} = 0{,}81$$

Policy D

$$precision = \frac{13}{13+2} = 0{,}87 \qquad recall = \frac{13}{13+2} = 0{,}87$$

$$F = \frac{(2 * 0{,}87 * 0{,}87)}{0{,}87 + 0{,}87} = 0{,}87$$

Policy E

$$precision = \frac{12}{12+3} = 0{,}80 \qquad recall = \frac{12}{12+7} = 0{,}63$$

$$F = \frac{(2 * 0{,}80 * 0{,}63)}{0{,}80 + 0{,}63} = 0{,}71$$

Policy F

$$precision = \frac{10}{10+4} = 0{,}71 \qquad recall = \frac{10}{10+2} = 0{,}83$$

$$F = \frac{(2 * 0{,}71 * 0{,}83)}{0{,}71 + 0{,}83} = 0{,}77$$

Policy G

$$precision = \frac{8}{8+2} = 0{,}80 \qquad recall = \frac{8}{8+1} = 0{,}89$$

$$F = \frac{(2 * 0{,}80 * 0{,}89)}{0{,}80 + 0{,}89} = 0{,}84$$

Policy H

$$precision = \frac{9}{9+6} = 0{,}60 \qquad recall = \frac{9}{9+0} = 1$$

$$F = \frac{(2 * 0{,}60 * 1)}{0{,}60 + 1} = 0{,}75$$

Policy I

$$precision = \frac{12}{12+3} = 0{,}80 \qquad recall = \frac{12}{12+3} = 0{,}80$$

$$F = \frac{(2 * 0{,}80 * 0{,}80)}{0{,}80 + 0{,}80} = 0{,}80$$

Policy J

$$precision = \frac{12}{12+3} = 0{,}80 \qquad recall = \frac{12}{12+10} = 0{,}55$$

$$F = \frac{(2 * 0{,}80 * 0{,}55)}{0{,}80 + 0{,}55} = 0{,}65$$

Policy K

$precision = \frac{10}{10+4} = 0{,}71$ $recall = \frac{10}{10+6} = 0{,}63$

$$F = \frac{(2 * 0{,}71 * 0{,}63)}{0{,}71 + 0{,}63} = 0{,}67$$

Policy L

$precision = \frac{10}{10+3} = 0{,}77$ $recall = \frac{10}{10+2} = 0{,}83$

$$F = \frac{(2 * 0{,}77 * 0{,}83)}{0{,}77 + 0{,}83} = 0{,}80$$

Policy M

$precision = \frac{11}{11+1} = 0{,}92$ $recall = \frac{11}{11+0} = 1$

$$F = \frac{(2 * 0{,}92 * 1)}{0{,}92 + 1} = 0{,}96$$

Policy N

$precision = \frac{11}{13+2} = 0{,}85$ $recall = \frac{11}{11+0} = 1$

$$F = \frac{(2 * 0{,}85 * 1)}{0{,}85 + 1} = 0{,}92$$

Policy O

$precision = \frac{15}{15+0} = 1$ $recall = \frac{15}{15+0} = 1$

$$F = \frac{(2 * 1 * 1)}{1 + 1} = 1$$

Policy P

$precision = \frac{12}{12+1} = 0{,}92$ $recall = \frac{12}{12+0} = 1$

$$F = \frac{(2 * 0{,}92 * 1)}{0{,}92 + 1} = 0{,}96$$

Policy Q

$precision = \frac{7}{7+3} = 0{,}70$ $recall = \frac{7}{7+1} = 1$

$$F = \frac{(2 * 0{,}70 * 1)}{0{,}70 + 1} = 0{,}82$$

Policy R

$$precision = \frac{9}{9+2} = 0{,}82 \qquad recall = \frac{9}{9+6} = 0{,}60$$

$$F = \frac{(2 * 0{,}82 * 0{,}60)}{0{,}82 + 0{,}60} = 0{,}69$$

Policy S

$$precision = \frac{14}{14+1} = 0{,}93 \qquad recall = \frac{16}{16+0} = 1$$

$$F = \frac{(2 * 0{,}93 * 1)}{0{,}93 + 1} = 0{,}97$$

Policy T

$$precision = \frac{12}{12+3} = 0{,}8 \qquad recall = \frac{12}{12+0} = 1$$

$$F = \frac{(2 * 0{,}80 * 1)}{0{,}80 + 1} = 0{,}89$$

Policy U

$$precision = \frac{12}{12+1} = 0{,}92 \qquad recall = \frac{12}{21+0} = 1$$

$$F = \frac{(2 * 0{,}92 * 1)}{0{,}92 + 1} = 0{,}96$$

Policy V

$$precision = \frac{12}{12+3} = 0{,}80 \qquad recall = \frac{12}{12+0} = 1$$

$$F = \frac{(2 * 0{,}80 * 1)}{0{,}80 + 1} = 0{,}90$$

Policy W

$$precision = \frac{9}{9+2} = 0{,}84 \qquad recall = \frac{9}{9+0} = 1$$

$$F = \frac{(2 * 0{,}84 * 1)}{0{,}84 + 1} = 0{,}90$$

Appendix F – Stand-alone framework version

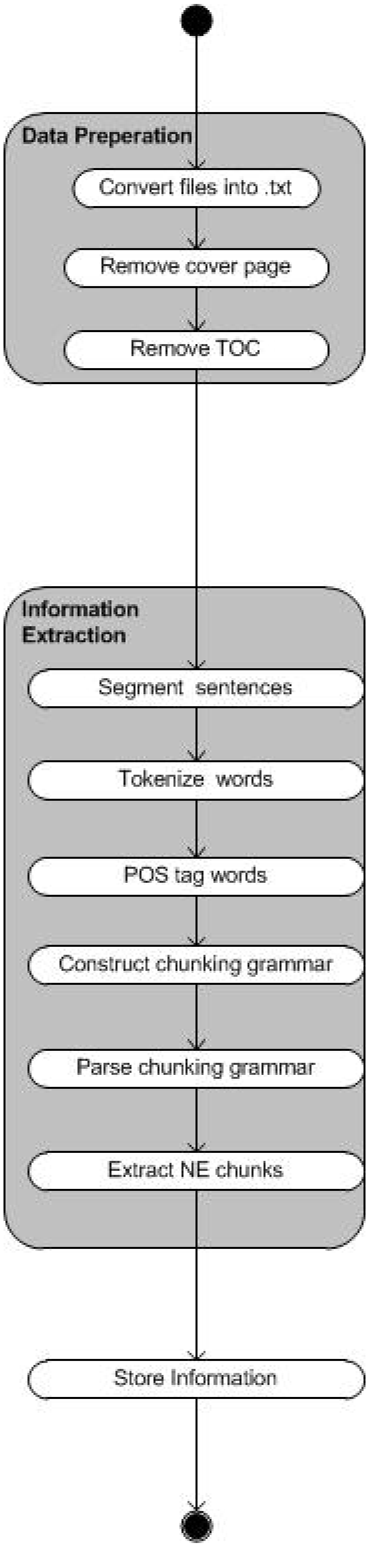


FIGURE 11 - INFORMATION EXTRACTION FRAMEWORK

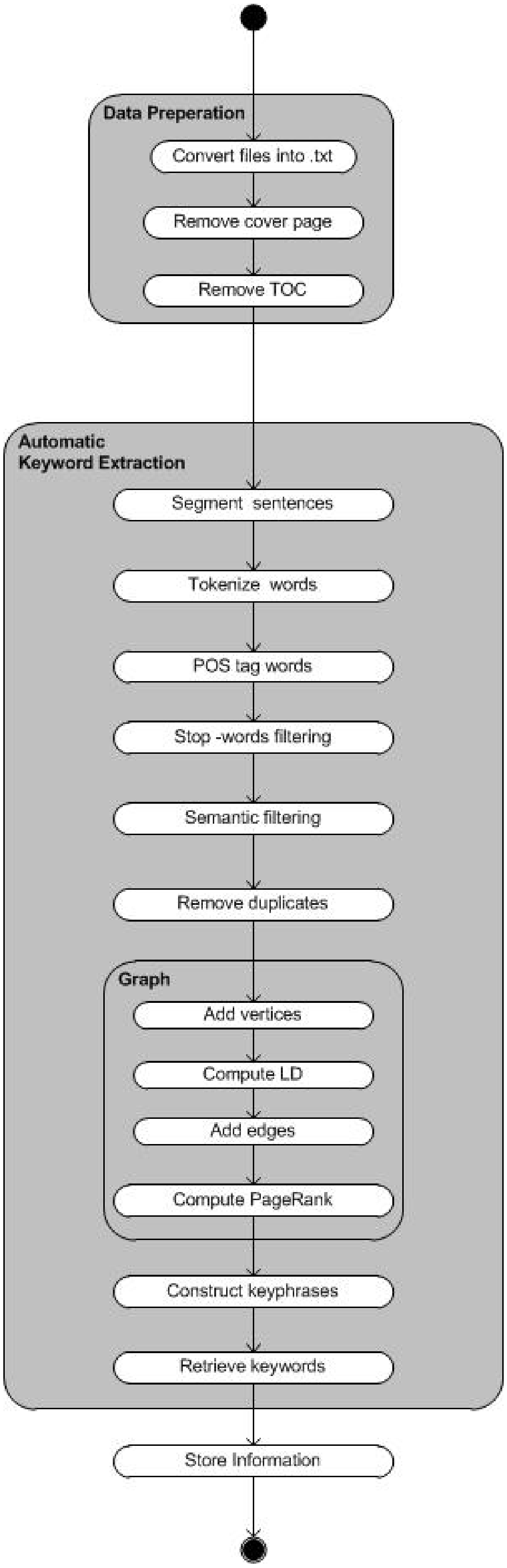


FIGURE 12 - KEYWORD EXTRACTION FRAMEWORK

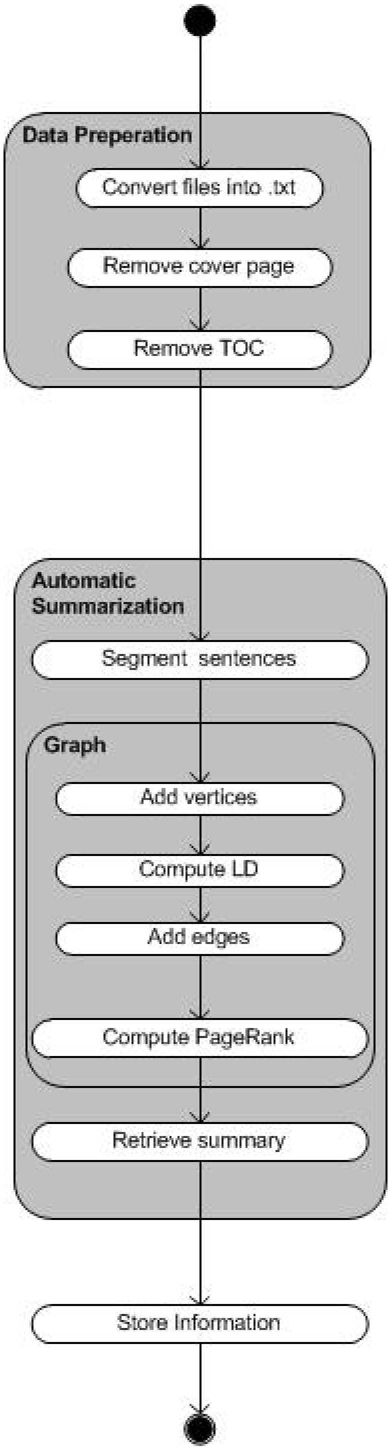


FIGURE 13 - AUTOMATIC SUMMARIZATION FRAMEWORK

Appendix G – Activity Table

Activity	Sub-Activity	Description
Data Preparation	Convert files into .txt	The corpus of document should initially be converted into .txt format, since it is the most appropriate data format for TM.
	Remove Cover Page	The cover pages of the documents provide no significant information regarding the concepts that the policy treats, thus they are removed.
	Remove TOC	Table of content is removed since it provides no significant information regarding the context of the documents.
Information Extraction	Segment Sentences	The entire textual content of the policy is segmented into respective sentences.
	Tokenize Words	The segmented sentences are further tokenized into respective words.
	POS tag words	The tokenized words are annotated with a Part of Speech tag, depending on its role in the sentence.
	Construct Chunking Grammar	Knowing the Part of Speech representation of the relevant words in the text, a chunking grammar is constructed to detect these words based on their POS representation and group them together.
	Parse Chunking Grammar	The constructed chunking grammar is parsed into the entire annotated corpus of words.
	Extract NE Chunks	When parsed into the textual content, Named Entities are detected based on the definition coming from the chunking grammar. These Named Entities are then extracted as information.
Graph	Add vertices	To construct the graph, vertices are added, which are represented by a single word or a sentence.
	Compute LD	The similarity between vertices is computed with the help of LD.
	Add edges	The results from the similarity computation are added as edges between the respective vertices.
	Compute Page Rank	The PageRank algorithm is computed on the entire graph of vertices and edges to determine the importance of each vertex with respect to the entire graph.
Automatic Summarization	Segment Sentences	The entire textual content of the policy is segmented into respective sentences.
	Retrieve Summary	A collection of highest scoring sentence vertices is retrieved as a summary representation of the document.
Automatic Keyword Extraction	Segment sentences	The entire textual content of the policy is segmented into respective sentences.
	Tokenize word	The previously segmented sentences are further tokenized into respective words.
	POS tag words	The tokenized words are annotated with a Part of Speech tag, depending on their role in the sentence.
	Stop-word filtering	Words without any significant meaning, such as stop-words,

		are filtered out from further analysis.
	Semantic filtering	Words with insignificant Part of Speech tag, are filtered out from further analysis. This is determined after a manual inspection is conducted on the words that hold a potential to be keywords.
	Remove Duplicates	The remaining collection words are checked for duplicates and cleansed accordingly.
	Construct key-phrases	From the list of highest scoring words, after the PageRank computation, two word key-phrases are constructed
	Retrieve Keywords	From the entire list of highest scoring keywords and key-phrases,
Store Information		The information that result from the three independent techniques are stored separately for further analysis and review.

TABLE 11- FRAMEWORK ACTIVITY DESCRIPTION