Utrecht University

School of Natural Sciences
Master Business Informatics

Gaining Business Value from Unstructured Data

Master Thesis Project

J.W. de Boer
5930758

First supervisor:
Prof. dr. ir. H.A. Reijers

Second supervisor:
Dr. G.C. van de Weerd

March 14, 2025

Abstract

This research aims to investigate how organizations can improve their business value when handling unstructured data. While they manage their product often in a structured form, unstructured forms are disregarded. The unstructured data is an untapped resource which can be 80% of the data of an organization. It contains a lot of knowledge and is at risk of being forgotten or ignored, requiring organizations to put in effort to investigate and document again and again.

At first, a structured literature review was performed to understand the background. The literature provided insights on organizations and their lack of a standardized way of ensuring the quality of their unstructured data. When focus is put on interpretability, relevancy, and accuracy in an iterative manner, organizations are bound to improve the quality. In the context of this research recommendations are made to apply data curation teams ensuring the quality of metadata to improve accessibility, sharing, and management of data. For managing data, literature suggests to apply domain-specific methods to provide structure.

To determine the impact of processing techniques on data quality, a comparison was made on quality metrics as a result of classifying differently processed datasets. Three methods were investigated, two with a different order of processing techniques - the methods of Barbantan and Lim - and one with a different set of steps altogether - the method of Sanchez-Segura. The results of the comparison show an overall lack of significant differences, indicating that the implemented processing techniques are not the sole reason for differences in quality metrics. Slight improvements in Accuracy and precision were observed with RF and SVM classifiers in the similarly structured methods, but large variations were found for recall and F1-scores for the NB and DT classifiers. Further research is necessary to gain full understanding of the potential impact of different processing techniques.

Keywords: *Business Value, Data Quality, Information Retrieval, Knowledge Retention, Metadata, Method Comparison, Natural Language Processing Techniques, Text Mining Techniques, Unstructured Data*

Acknowledgements

I want to express my thanks to you - the reader - for picking up this conclusion to my journey of the Master Business Informatics. It has been a challenge and an obstacle, looming in the distance, always on the horizon of my time at Utrecht University. It is now coming to an end. The finalization could have only been possible with the help of some individuals whom I would like to extend my gratitude towards.

First of all, I would to thank my primary supervisor - Prof. dr. ir. H.A. (Hajo) Reijers - for his involvement and motivation in this research. When I started the thesis project, I had not been as actively involved with studying as I would have liked. This meant that I needed some help with finding my footing again in the world of research and I am glad that my supervisor was able to help me in this. At first, it was difficult for me to keep a consistent and realistic schedule. Many of my research-related skills had fallen away over the course of the years. Hajo was able to present me the effects of my indecisiveness and showed me the possible consequences. Due to this, I was able to face reality and pick up matters into my own hands, making a better planning and providing consistent updates about my progress. Sincerely, thank you for all the guidance you have provided me, I would not have been able to finalize it without your help.

Second of all, I would like to extend my gratitude towards my family and girlfriend who have supported me on this journey. To Lia and Willem, my mom and dad, who listened to my ramblings and discussed with me topics of interest I found throughout the research. Their love, support, and optimism helped me get through the harder aspects and I would have struggled much more without their involvement. Also to Joost and Denise, my brother and sister, who provided invaluable insights through discussions and who gave me great feedback and pointers from their own experiences. Lastly to Alina, whose endless enthusiasm, love, and silliness infected, motivated, and distracted me at the right times to be able to finalize this research.

Last but not least, I would like to extend my gratitude towards my friends, both fellow-students and old friends from back home. Marian, Willem, Marijn, Lucas, Lisa, Mats, Mette, Sander, Jari, and many others of my old board and from Study Association Sticky; Martijn, Pepijn, Daan, Rick, and others from my friends back home - Thank you all. For your support, insights, ideas, and time that you have given me which helped me shape this research into what it is today. Without all of you, this would not have been possible.

Thanks, from the bottom of my heart!

Contents

Chapter 1

Introduction

1.1 Problem Statement

In the last years, multiple studies have found that the amount of unstructured data that an organization produces is increasing to about 80% and increasing exponentially [14, 17, 18, 24]. Exponential data creation also enables exponential knowledge creation, which is at risk of being drowned out in the noise. How does an organization make sure that their knowledge remains understandable? This is a challenge, as most organizations have a lot of unstructured documents laying around with difficult-to-access knowledge whose content is not immediately understood [15, 21]. Organizations create many documents, some structured and some unstructured. These documents describe knowledge concepts by providing specific perspectives using information. This information is nothing more than data being given context. Misunderstandings from faulty interpretations of unstructured documents can have a major impact on the decision making process and therefore decrease the overall value these documents have on an organization. Due to the nature of unstructured data, the contained information is different per format. While one format can contain straightforward information such as blog posts or wiki-articles, other formats are harder to get the information from such as images or audio logs, but others exist such as e-mails, resumes, reviews, and more. As can be seen from the variety of formats, there is not an easy, one-size-fits-all way to structure this information.

Why one would go for unstructured data over structured data has to do with the aforementioned knowledge - or information - that resides in unstructured data [2]. To gain value from *structured data*, one can investigate a particular knowledge concept by searching for its known attributes. With unstructured data, this is slightly different as the attributes are not directly

known, making it harder to determine if specific knowledge is present. Unstructured documents still contain a wealth of information that is useful for different business cases, such as in decision making [15]. To investigate unstructured data means to investigate how information can be made more useful for organizations. Organizations produce so much of this data, but how do they use it?

There exist multiple ways to handle structured data, often in a straightforward process involving processing steps on specific concepts and transforming them to gain useful information [28]. This is a direct means of using a document, but this can not always be done when the contained information is unstructured. Since structured data occurs mostly in tabular format with rows and columns, it is easy to find specific information as the data is known in advance. However, with a variety of formats and associated attributes in unstructured data, it is more difficult to find specific information. Applying the same methods one would use for structured data to an unstructured data set would not have the same effect. The information is not in the same places, and the information could be redundant or unclear. It is therefore impossible to use the same methods to unstructured data without structuring the data in some way. Different processing steps are required in order to make use of the unstructured data.

The steps after processing the data vary for each domain. For example, structuring data into specific conceptual information for a healthcare domain should not be done when structuring data from user-reviews for a cosmetics website. What these methods do have in common however, is that they are using techniques to process the data in some way *before* their domain-specific use. These techniques are not always the same per domain. While they often have some form of Natural Language Processing involved, different methods apply them in different ways. Some are straightforward and use basic processing techniques, others use techniques from a trusted pipeline-like order, with other methods using different ordering in these steps. The variety of approaches can also be found in methods for structured data, but it is not clear if similar approaches can be applied to unstructured methods too. As processing is at the heart of many methods, it is important to know which of these approaches works best and if there is a difference in results for providing structure to the unstructured data.

1.2 Research Aim and Objectives

With the enormous volume of unstructured data that exists for organizations, it is necessary to structure any information so it can be reasoned over [17]. Using a structure to define and enrich the existing data can therefore lead to an increase of value for the different data assets that an organization has [15]. However, structuring the unstructured data is not done with one overarching method. Rather, different methods exist for different applications of unstructured data, such as methods specific to image processing or methods for textual processing. Due to these different applications, it is difficult to find a common factor as a cause for certain effects on the processed data. Research shows various quality indicators for testing for specific effects within unstructured data, such as accuracy, recall, precision, and F1-scores [16].

It is interesting to determine how these quality metrics are affected by differences in processing techniques. By finding which processing steps improve the data quality, the overall value for a business can be increased by applying these same steps. For a business managing its data, it is crucial to know what information resides in which document. Lower scores for these metrics can indicate a lower quality of the data, which might indicate that the information within the document is not accurate and should therefore not be included for decision making. Besides enhancing the decision making, knowing that the contents of the documents are accurate also improves customer intelligence and provides operational efficiency. Customer intelligence is improved as the accurate knowledge assets also provide more accurate predictions of their needs, which can proactively be addressed. Operational efficiency is reached through streamlining activities and improving processes with improved decision making from the accurate reflections of the knowledge assets.

Processing techniques for unstructured data vastly differs per domain. Therefore, this research aims to highlight the effects of different unstructured data methods over textual data, specifically the effects of transforming the text. In this research, the possible value for an organization is determined by the usefulness of the processed, unstructured data. This usefulness consists of various factors derived from sources in the literature. Focus is put on the different processing approaches that transform the data.

This research aims to fill a knowledge gap in the literature by investigating the effect of different processing techniques on the quality of unstructured data by measuring specific quality metrics. With the knowledge of what leads to the highest quality, practitioners can be assured of the inherent accuracy of their data and therefore improving the value of that data for

an organization. No previous research was found that generalizes, combines, and recommends specific techniques for processing this type of data. By comparing three different approaches from different domains mentioned in the literature on unstructured data, this research aims to find a generalized approach or recommendation regardless of domain. With this knowledge, it might become clearer how an organization can improve their business value through processing for quality.

1.3 Research Questions

To answer the above research problem, multiple research questions were formed. The primary focus of this study is to find out how the business value of an organization can be improved by focusing on the application of processing techniques to unstructured data. With the main topic being on the impact of unstructured data on an organization, the main research question is constructed as:

MRQ: *How can unstructured data be used to improve business value?*

To answer this research question, different aspects of unstructured data need to be analyzed. For example, it is necessary to know what unstructured data looks like, what kind of value it contains, and more. The sub-questions are answered throughout the rest of this research.

SQ 1: What are characteristics of unstructured data?

SQ 2: What value lies within unstructured data for organizations?

SQ 3: Which methods exist to obtain value from unstructured data?

SQ 4: How are methods for obtaining value from unstructured data different?

SQ 5: How is unstructured data affected by differences in processing techniques?

1.4 Approach

To answer the research questions, a systematic literature review was performed to gain insight into what unstructured data exactly is, how it is handled by organizations, and what kind of value it could bring to organizations. In the literature review, a variety of different methods was found

across domains that show different approaches to processing unstructured data. From these methods, three were selected for a comparison on how they process the information, by measuring their performance on specific qualitative metrics. The comparison was made by implementing the described processing techniques and applying them to different datasets. The results were tested on their *accuracy*, *precision-*, *recall-*, and *F1-* scores as suggested by Kiefer [16]. These scores show the results of the processing steps to different types of unstructured data and could therefore provide an indication on domain suitability and possible generalizations. This is further expanded upon in chapter 3.

1.5 Contributions

Practical Contribution
The practical contribution for this thesis is found in the presentation of the results. In current research, some effort has been made to determine the effects of different processing steps, but a generalized recommendation has not yet been given [17]. By comparing the information from different domains, it is possible to determine the effect that different techniques could provide. It also becomes possible to use the results to provide a direction for a generalized approach, recommendable to anyone using unstructured data. This way, practitioners of unstructured data can be assured they are getting the most accurate information and use this in further decision-making.

The results of this study might make it easier for various practitioners to work with unstructured data when using different methods. Researchers and practitioners that use this type of data can be assured of the impact of the differences in processing techniques and can therefore focus more on other effects within their research domain.

Scientific Contribution
The scientific contribution for this paper is found in the implications from the implementation. By comparing the results of the different processing techniques, it can become possible to comment on their relevance in general methods. This knowledge can be applied in research on other methods containing unstructured data, where the recommended steps that are found as a result of this study can be applied too.

Comments made by Halevy, state that there are many differences between structured- and unstructured data [10]. Their comments on accuracy leads one to believe that the definition of what improves accuracy can contribute to bridging the existing gap, useful for future scientific researchers.

1.6 Thesis Structure

The paper continues in chapter 2 with an extensive description of the current literature regarding unstructured data, business value, and how different methods affect the business value. The topics mentioned in this chapter are based on the research questions posed in section 1.3, and are answered in the last section of this chapter. In chapter 3, a description is given on the research method used in this study, describing the approaches to the literature review as well as the implementation with the used datasets. In chapter 4, the actual implementation for this research is described with code examples that have been used in the research. Sections within this chapter describe how some of the default metrics were calculated, as well as how the different processing techniques were constructed within the investigated methods. In chapter 5, the results of the study are presented in a section on the *Accuracy*, as well as a section on the other metrics. The implications of these results are discussed in chapter 6, where limitations to the study as well as recommendations for future work are also discussed. The thesis is finalized in chapter 7, where the research questions are discussed in retrospect, allowing for a final conclusion on how unstructured data affects business value.

Chapter 2

Literature Review

In this chapter, the Systematic Literature Review (SLR) is presented that aims to answer the sub-questions posed in section 1.3. First, a short overview is provided on the process that was used for executing the SLR in section 2.1, resulting in the background necessary for answering the sub-questions. The background information is made explicit in the sections following this, starting with a discussion on unstructured data and its value to organizations in sections 2.2 and 2.3, with an overview of existing methods on processing data in section 2.4. Lastly, the answers to the sub-questions are presented in section 2.5.

2.1 SLR process

The SLR was used to understand the context surrounding unstructured data and to find existing methods of structuring the data. In this process, existing papers, books, studies, and other scholarly material has been investigated relevant to the described problem in section 1.1. From the information that was initially found, answers were found for the questions posed in 1.3, providing a solid overview and direction for this research.

The SLR was conducted according to the protocol defined in Appendix A.1. In this protocol, the main source of information consisted of Google Scholar searches due to the familiarity of the researcher with this platform. Other platforms were briefly consulted for verification of information. Using the protocol, keywords and phrases were formed through which many valuable sources were found and assessed, resulting in a shortlist of references with which the research questions were answered. The main topics of these questions are found in the sections below, in which the key takeaways from these references are summarized.

2.2 Unstructured data

It is important to know that *Structured Data* and *Unstructured Data* are not total opposites, as data is always structured in some way. Sapsford notes that structured data is coded and structured according to specific analytical categories in advance, while unstructured data lacks such structure [23]. Inmon and Nesavich take a different approach to defining the two. They define structured data more literally, describing it as information that is represented by numbers in various tables, divided over rows and columns [14]. They state that this type of data is often disciplined, predictable, and repeatable. Inmon describes data as numeric in nature. This means that the data is easily used for analytical purposes. However, the numerical nature does not mean that structured data can never appear in a textual format. While most data exists numerically, the textual data can be used to identify or describe some form of numeric data. For example, a *true* or *false* value usually indicates a boolean field of 0 or 1, but values for *small*, *medium*, or *large* can indicate the numerical sizes of an item. While the textual representation differs, it still refers to numeric data.

Investigating unstructured data tells a different story. Inmon and Nesavich describe that unstructured data comes in many formats such as image-, color-, audio-, shape-, or textual-data. This variety of formats means that there is no common structure or repeatability of information that can easily be recognized. The authors state that unstructured data occurs almost anywhere and is the cause of many challenges and opportunities for organizations in the decision-making process. These notions are supported by other experts in the field. In his book on *Mining the Talk*, Spangler notes that unstructured data arises when the information is described in ordinary, everyday language, as the information is conveyed without consistent word-choices or sentence-structures [26]. Spangler defines this type of unstructured data as the "*talk*" as it is as the way humans are communicating with one another. Spangler finds that the *talk* is the kind of data that appears most in the world and is potentially the most valuable. He states that the small pieces of information embedded in the data can show actionable intelligence for businesses and potentially present solutions that a business might face.

There are more differences between these types of data than what is only implied by their names. A study performed by Halevy et al. found that there are already differences on a technical level with authoring, querying, information sensitivity, data sharing, and accuracy [10]. They argue that the flexibility of unstructured data on these topics is its greatest strength, as the information can be retrieved with a relevancy-score to similar topics. The

authors argue that the authoring of information can lead to the best benefits, but state that the challenge lies in how to add the information.

From the previous information about structured data, one can conclude that its qualities are aimed at consistency and searchability. Due to the known structures, it is easy to create data and search for the requested values. Unstructured data is more difficult to assess the quality of, which was the main topic of research by Kiefer. In their research, Kiefer states the importance of assessing the quality of the underlying data to be able to make good decisions [16]. In their paper, Kiefer scores the unstructured data on three dimensions: *interpretability*, *relevancy*, and *accuracy*.

- **Interpretability:** This quality concerns the similarity between how one receives- and how one expects the data. The quality decreases when the data that is used is farther from the data that is necessary and therefore expected. For instance, differences in training data can affect the resulting prediction. Kiefer states that the completeness of the data is crucial to interpret the data properly.

- **Relevancy:** For this quality, the similarity between the input data and the optimal data needs to be determined. A measure for this could be the frequency of keywords between the expected text and the produced text, or the amount of domain-specific terms covered in the produced text.

- **Accuracy:** The third quality assesses the similarity between the input data and the data representing the real world. The author mentions the use of precision, recall, and accuracy measurements for the best results, as they are examples of gold standard metrics.

These qualities are not the only ones mentioned in literature. In a book on *Data Collection and Analysis* by Sapsford and Jupp, details are presented on the best way to conduct analysis on unstructured documents. Such a document can be seen as a collection of unstructured data points. The authors identify key characteristics of unstructured data that need to be reflected upon for a proper qualitative analysis. These consist of the data being interpretable, it being an iterative process, and requires a reflexive assessment [23].

- **Interpretability:** Similar to what is mentioned by Kiefer above, the content of the data needs to be checked. Since the data is gathered in different ways, it is possible that it is missing specific information. To increase this quality for unstructured data, it is important that

information needs to be collected as accurate as possible, so that it leaves no room for different interpretations.

- **Iterativity:** The authors describe the qualitative approach for unstructured data as a creative process, where the exploratory nature of investigating possible connections is key. They state the importance of labeling and categorizing the data in an iterative manner to make sure all relevant concepts are accounted for.

- **Reflexivity:** Lastly, a reflexivity assessment is mentioned for improving the quality of data. To ensure that a reasonable amount of possible errors is minimized, the authors suggest to track the information that is added to the unstructured data about its nature. For example, if this describes the origins of the data properly, the quality of data is deemed higher.

Comparing the components mentioned by Kiefer to the ones mentioned by Sapsford, it can be seen that the described qualities are very similar. While the interpretability is an immediately obvious commonality due to its name, the other matches are less so. Iterativity can be matched with Relevancy, because with this quality one can make sure that all relevant data is included for analysis. Furthermore, Reflexivity can be used to improve the Accuracy, as the assessment can show aspects on which the data can possibly be improved. This makes the information more accurate and increases its quality.

Such qualities are also important to know for processing the unstructured data. In a study by Ajah and Nweke, unstructured data is compared to *Big Data*. Big Data is characterized by a high volume, variety, velocity, and veracity and is produced by various sources such as mobile devices, sensors, social media, and more [1]. Since these characteristics overlap for unstructured data, they state that applying similar techniques can be beneficial to process the data as they will become more effective and efficient. A successful implementation of an approach would require different factors, such as having the right technology infrastructure, the use of skilled data scientist teams, and motivation from an underlying data-driven culture. The authors state that with these key factors in mind, unstructured data can be successfully processed for organizations.

All in all, unstructured data contains information on how specific business concepts are communicated and can therefore show insights on what processes can be made more efficient. This type of data is highly flexible but lacks consistency and searchability, making it difficult to assess its inherent quality. Unstructured data can be assessed by testing for different quality

factors, such as interpretability, relevancy, and accuracy. For such tests, quality metrics such as accuracy, recall, precision, and F1-scores can be used by specific data teams. These teams can be used to observe the differences between previous iterations and new ones, providing insights and actionable points for an organization.

2.3 Value in data

Data is always valuable in the world of IT. In their paper on how executives perceive the value of IT, Tallon, Kraemer, and Gurbaxani highlight two findings that show how IT can affect business value [27]. They state that clearly defining and assessing the current goals can show the direction a process can take and which IT systems are useful for achieving them. The authors also find that regular evaluations of the applied strategies can affect how a business is applying its IT. With changing requirements and market needs being able to influence a project greatly, organizations are recommended to improve their responsiveness and flexibility by developing and applying new IT strategies. With the recommendations for timely evaluations, misalignment issues between the goals and possible solutions can be caught early on. Strategies to change part of the IT infrastructure can then be made, increasing efficiency and effectiveness resulting in an improvement to the business value.

While the business can be improved with changes in the IT systems, the associated actions still need to be made with the right context in mind. Knowing about the associated domain knowledge is crucial to choose the right IT solution and the steps that should be taken afterwards. Without this knowledge it is not possible to achieve the same value, so it is crucial that knowledge is retained among workers and remains in the organization. For this purpose, Burmeister and Deller have identified different opportunities for organization to improve knowledge retention in their work [4]. The authors find that organizations can improve not just on increasing the points of interaction and the context in which information can be shared, but also on understanding the innate drivers of workers to share information.

The suggestions by Burmeister can be applied to start the knowledge retention process, but this does not help to document the actual knowledge. According to a study by Levy, a recommended structure to preserve the knowledge is crucial to retain and re-use the otherwise lost knowledge [19]. In their study, Levy finds that it is essential to prioritize a specific aspect of knowledge before trying to transfer and re-use it. Without prioritization, it

is possible that knowledge will be retained that is not necessary for proper knowledge transfer.

Priority is often put on concepts related to the business goal that a project is trying to achieve. While this information can be derived from structured documents quite easily, this is often not the case for unstructured documents. Structured data has its information often in a regular, prescribed location, while unstructured data is saved more irregularly [15]. In their book on *business metadata*, Inmon, O'Neil and Fryman describe that the metadata of both structured and unstructured documents can be used to understand the enterprise data assets, which can help with decision making issues.

It is usually a difficult process to understand the contents of unstructured data. In his work, Inmon recommends that the documents containing this data need to be discovered and classified, so that workers can more easily find the information that is useful for the organization. Inmon suggests using metadata for this process, as this often contains important information about the document itself. He further follows up on the use of metadata in the perspective of knowledge management. While business metadata can not capture the information that is never told, it *can* be used to capture the essential information of a document in such a way that this can be shared and used to improve the organization.

A paper by *Alation* mentions similar use of metadata for unstructured documents as well. They find that metadata is often added to deliver context to a document, such as showing the category a document belongs in, the provenance, ownership, and possible usage [2]. To make actual use of the metadata, the paper suggests using metadata teams to manage the discovery, governing, and utilization of specific data. With a dedicated metadata team, a greater understanding of what kind of data is managed can be achieved. With each team focused on specific data, it can be a challenge to keep the same company-wide structure for how the knowledge assets are recorded. Wilkinson et al. suggest to focus mostly on the key characteristics of the data, so that a common structure can be maintained [30]. They recommend to keep sharing information between the managing parties about the data itself, mainly focusing on the reusable metadata such as provenance and quality.

Using teams to implement and manage metadata is an idea that finds resonance in other work as well. In his book on data stewardship, Plotkin describes the importance of teamwork in achieving full data asset control [21]. In his work, Plotkin describes the full process of data stewardship as the operational side of data governance. Data governance is described as the way *people* interact with data and makes sure that the data is properly under-

stood, while data stewardship focuses on formalizing the accountability for managing information resources. In other words, without proper data stewardship, it can be impossible to find the competitive advantage in the data that an organization might be looking for. All in all, assigning a dedicated team and encouraging information sharing can help to make sure certain knowledge will not be lost and an advantage can be gained.

From the above information, it can be taken that knowledge cannot easily be lost when embedded in the metadata. When teams are used to manage the metadata properly, it will become difficult to lose this information entirely. While this information is not always completley lost, it is not as straightforward to find the data for a specific task. To gain this information, it is best to use steps from data curation according to research by Thirumuruganathan, Tang, Ouzzani, and Doan [28]. They state that the best process to extract real business value from any data is to apply data curation and define it as a process where data discovery, data integration, and data cleaning are essential tasks for any analytical purpose.

- *Data Discovery* means that data needs to be identified and selected according to the requirements of the task. Only the data that is deemed relevant according to specified criteria is kept for the following analytics. To find the correct data, techniques such as table- or keyword searches could be applied.

- *Data Integration* is the process of combining data from various sources - often from across a whole organization - into a single, accessible dataset. Data only needs to be collected when it is relevant to the task and requires the use of techniques such as schema matching and schema mapping to identify similarities between concepts. In this stage it is important to find concepts that describe the same value, as their link might be important to find other information to solve a problem.

- Last, *Data Cleaning* is a straightforward process to identifying errors in the data and fixing them. Without cleaning the data, outliers and general impurity can have a big impact on the final results and could mean that certain context information will be lost.

According to Thirumuruganathan, it is important to use these steps when handling data for analytical tasks.

2.4 Existing methods

The literature provided a wide variety of existing methods for different domains. At their core, many investigated papers recommend the application of high-level structures to unstructured data, as the domains are often too different to make specific recommendations. What works in one domain might not be relevant in another.

One example of a high-level approach can be found in a paper on creating community portals containing unstructured data. In this, DeRose, Shen, Chen, Doan, and Ramakrishnan describe the core steps to retrieve and transform the contained data [8]. They describe an iterative approach with an initial data selection, followed by an extraction- and integration process. The authors note that the resulting information expands over time and therefore becomes more complete with each iteration.

A paper by Chu, Baid, Chen, Doan, and Naughton describe a similar high-level approach. In their paper, they also use a selection method with extraction and integration techniques, but apply different clustering methods to label and group information based on similarity [5]. They state that this can be useful for an easier access to the data, and if this is applied iteratively, can increase understanding of the data.

In a paper on creating a technique for better brand and reputation analysis, Spangler et al. also includes these high-level concepts [25]. In their approach, they provide recommendations for an initial extraction of information, transform it, and apply further techniques for a given domain. In their work, they expand upon a specific technique using snippets to gather context about their data. They use this context-information to identify brands and topics, linking specific information using an orthogonal filtering technique. Using the contextual information in this way, the authors find that knowledge can easily be enriched with complementing information useful for other types of analysis.

From a high-level point of view, methods for unstructured data start with retrieving information, transforming it in some way, then apply domain-specific techniques, as can be seen above. Some authors also describe specific solutions for specific domains, including various forms of *information retrieval*- and *text mining* techniques.

Much value can be gained from unstructured data through such techniques. Several examples can be found in different *Business Intelligence* (BI) steps for managing unstructured data [6, 11]. BI can provide a structured environment for analysis of unstructured documents, with its ability to capture, organize, and convey information.

Some papers describe the usefulness of on orchestration of information retrieval techniques and how they perform together to create a harmonious system such as an *unstructured database management system* in a paper by Doan et al. [9]. In their paper, the authors do not detail specific steps, but find that information retrieval is key to finding the right information from unstructured data. While many information retrieval techniques are key to find the right information, they need to be adjusted to each domain to retrieve the most relevant information [12].

Not only information retrieval techniques need to be adjusted, there are many benefits for different text mining techniques which need to be accounted for [17]. In their work, Kumar, Dabas, and Hooda describe the need for focusing on which text mining technique is required for a specific domain, as each has its own structure and characteristics. Some text mining techniques are required for further analysis, such as term frequency for clustering, or tokenization for classification. Each technique has its strengths and weaknesses, and it is important to find out which works best for improving the knowledge discovery.

There are other methods mentioned that specifically process unstructured information. In a method describing the processing of unstructured medical documents, Barbantan and Potolea dive deeper into making sure the data is as expected for their *MedCIM method* [3]. In their paper, they describe that first, several text processing techniques need to be applied due to the unstructured nature of the data. Without this step, the extraction of the specific concepts would not work sufficiently. After text processing and concept extraction, the resulting data is used to identify specific concepts in the domain to make recommendations.

Another method also details the need for domain-specific adjustments. In the *C4PM method* described by Sanchez-Segura et al., business objectives as well as potential problematic areas need to be identified first to find the unstructured intangible knowledge assets [22]. These assets are explored by cleaning and then using several text mining techniques for classification. With this classification, it is possible to identify the context better, leading to better decision making.

Lastly, a recent method from the field of requirement engineering also discusses unstructured, natural language text and how to process it for defining requirements [20]. In this *Unified Boilerplate method*, Lim et al. mention that the requirements that are created are often ambiguous due to their natural language origins. They state this makes it more difficult to process and it decreases the overall quality for software testing purposes. In their method they propose the use of a Natural Language Processing (NLP) pipeline to-

gether with a unified boilerplate approach after the corrected data has been collected. A boilerplate approach requires the information to be present in a predefined linguistic structure, therefore the data that is used as input also requires more structure. By applying the NLP steps in the same order, they make sure that the data can always be processed further down the line in the same way.

The structure mentioned by Thirumuruganathan describes the three most important steps for improving business value. These steps reappear in most methods described above. However, most of these methods are high-level and use a general variation of these steps. For instance, the *data discovery* can be mapped to variations of finding the data and extracting it, and the *data integration* step is also one that is widely used. However, this is not the case for the *data cleaning* step. This only becomes apparent in the methods of Barbantan, Lim, and Sanchez-Segura. These three methods focus mostly on making sure the data is clean as part of the *data discovery*, after which the data will be integrated for a specific domain. However, they all use slightly different approaches. The Unified Boilerplate approach for example, uses a pipeline of actions with known techniques. The MedCIM method on the other hand, recommends similar processing steps with a slight variation in its order. Lastly, the C4PM method also uses a pipeline of specific techniques, different to the ones in other methods. What they all have in common, is that they apply their techniques before domain-specific ones are applied.

2.5 Conclusion

The first question was created to gain an understanding of what exactly *is* unstructured data and what can be done with it. As described in section 2.2, unstructured data can be seen as information that is not conforming to pre-existing structures and is therefore more difficult to reason over without prior information. Without an additional step, it is not clear what the data looks like and what kind of information can be gained from it. Work by Kiefer and Sapsford and Jupp mention how to understand unstructured data, specifying the need to focus on quality characteristics such as interpretability, relevancy, and accuracy [16, 23].

The second question was posed to find the value in unstructured data. In section 2.3, knowledge retention from unstructured data was found to be one of the key concepts of value for an organization. Organizations have many knowledge assets that provide direct value for their processes. Losing such an asset can therefore impact an organization greatly. Inmon found that

such assets are mainly used in the organizational decision making, focusing on extracting specific information from metadata from unstructured documents [15]. Other sources highlight the need for assigning dedicated teams to manage the extraction of this information, improving reusability and data stewardship [21, 30]. Extracting exact value from data can only occur when the *right* data is examined and it shows structural similarities, but above all the data needs to be processed correctly for specific domain usage [28].

On overlap between features in unstructured data methods was found by investigating a variety of methods, leading to a better understanding of the most important parts of these method. Three main phases were found: a data collection and preparation phase; an extraction and integration phase with specific *text mining* techniques; and an analysis phase to handle the extracted data. Due to differences in the follow-up steps depending on the domain, it is clear that the techniques used in data cleaning need to be reviewed more thoroughly to determine their impact on the business value. To see if the order or combination of different steps affects the accuracy of found concepts, an implementation is made where these effects are tested. This implementation is described in the next chapter.

Chapter 3

Research Method

In this chapter, the research methodology is shortly described. First, the research technique central to this research is described in section 3.1. After its description, it is explained which data is used and why in section 3.2.

3.1 Research Techniques

To reach the objective of this research, a twofold approach was taken where both a systematic literature review and a comparison of methods were performed. The combination was used to find answers to various research questions and finding a gap for this research. The SLR was central for understanding the context and was therefore previously described in section 2.1. Below, the process used for the main focus of this research is presented, describing the steps taken for the method comparison.

3.1.1 Method Comparison

A comparison on the commonalities and the performance of different processing steps from different unstructured data methods was performed as part of the second research technique. As briefly described at the end of section 2.4, three methods were found that focused more on *data cleaning* prior to using the data for their respective domains. The contents of these methods was also an interesting aspect for their inclusion in this research. Two of the methods include similar processing techniques, yet they appear in slightly different ways. The other method includes different steps altogether, including some common processing techniques but also ones not present in the other methods. Barbantan's method and Lim's method for instance, focus more on text mining techniques where they assign various labels and reduce words to their

base forms. The method by Sanchez-Segura (in this method, also sometimes abbreviated to "Sanchez' method") focuses more on the Business Intelligence aspect and less on the text mining techniques. With the focus of the last sub-questions on the effects of different processing steps on the quality of the unstructured data, these methods were selected for implementation and assessing any changes in quality metrics when applied to a dataset.

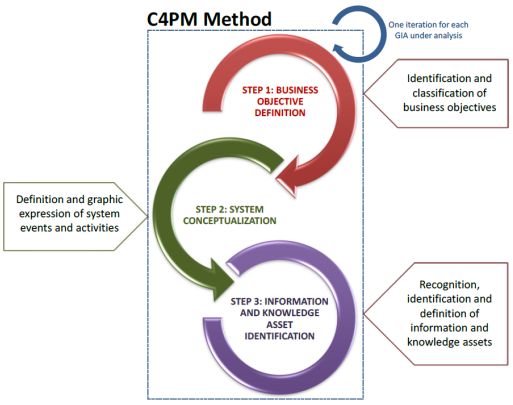

Figure 3.1: *Proposed C4PM method p.10 [22]*

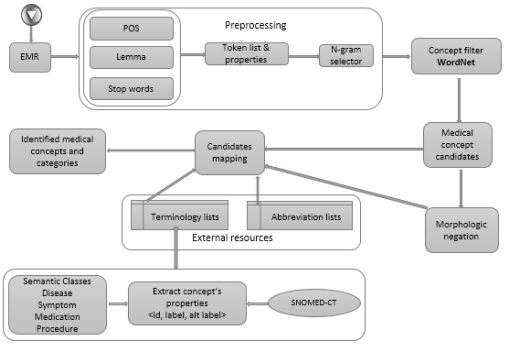

Figure 3.2: *MedCIM method as described by Barbantan & Potolea, p.5 [3]*

Commonalities and differences were found between Sanchez' method, Barbantan's method, and Lim's method. The first method, also known as C4PM, focused on a combination of NLP and business intelligence steps on a minimal level. The author states that over complication might lead to inaccuracy, which they tried to prevent by applying this method [22]. Barbantan's method, known as MedCIM, also focused on NLP steps and expanded it with a concept comparison and an extraction technique due to the importance of

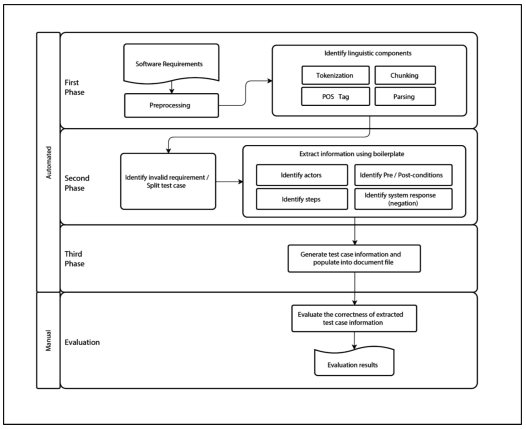

Figure 3.3: *Proposed Boilerplate Approach p.5 [20]*

concept matching for the medical domain [3]. Lim's method consisted of a *Unified Boilerplate Approach (UBA)* and focused on a more rigid NLP pipeline, where systematic transformation was found to be a key factor for positive change, based on earlier work [20, 29]. Their approach involves an identification and separation of the data before a linguistic analysis is performed. Visual representations of the methods as made by the authors are presented throughout Figures 3.1 to 3.3.

To answer the fifth sub-question stated in section 1.3, a specific comparison was made to determine the *quality* of the resulting unstructured data. As seen in section 2.2, measuring the quality of unstructured data requires a good *interpretability*, *relevancy*, and *accuracy*. With these three characteristics in mind, Kiefer notes that focusing on a comparison using the precision and recall scores is the best measure for the quality of the data [16].

For this research, a comparison is therefore made between the aforementioned quality metrics of an unmodified dataset and those of a dataset processed as suggested by one of the three methods. To gain these quality metrics, a specific type of classifier is chosen first. As suggested by Kumar et al., there are many different classification techniques that behave in a different way, with different strengths and weaknesses [17]. By applying the datasets on multiple classification techniques, any bias that could come up from one technique on a specific unstructured datatype could be mitigated. With a classifier selected, a dataset needs to be selected too. After that, the dataset needs to be either processed using the processing steps from the selected method, or not be processed at all and kept as the original dataset. Either way, the classification is applied by training the classifier on a training set and then predicted with a test set. From this prediction, the various quality

metrics could be gathered and are noted down. In Figure 3.4, the process for acquiring these metrics as described above is visualized. Each of the metrics was recorded separately to enable a comparison between the scores of the original dataset and the scores of the processed datasets.

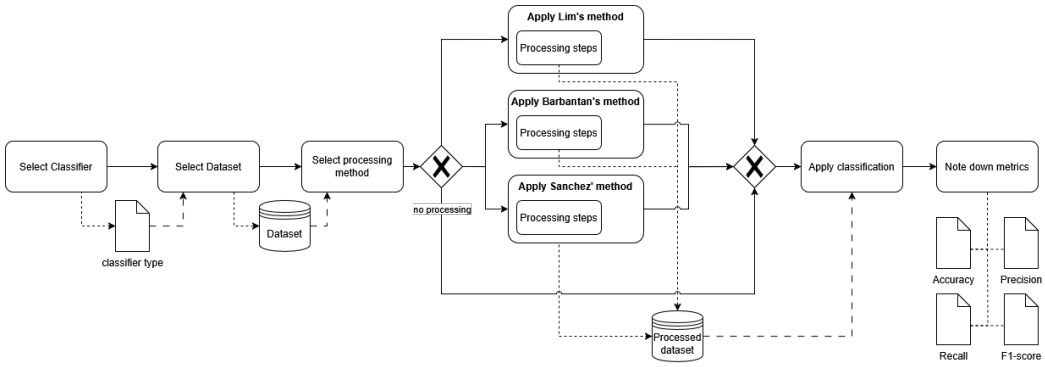

Figure 3.4: Process used in the comparison of methods

The comparison can have two outcomes: 1) No significant increase in accuracy could be measured - indicating that the specific differences in the processing steps have little to no impact on the resulting accuracy; 2) The combination of some of the processing steps provided a measurable increase of accuracy - indicating that this combination of processing the data is recommended for assigning the most accurate labels.

3.2 Data Handling

3.2.1 Data Collection

The processing steps of the methods described in the previous section were applied to various datasets, so that the quality metrics from the prediction of the processed datasets could be compared to those of the original datasets. All three approaches were implemented separately and applied to different datasets for validation of the results. The variety of different types of unstructured data is what was used to combat bias. By using different types of unstructured data in the evaluation of the methods and not only focusing on one type of data, the results will reflect how different information could affect the outcome. For proper analysis and evaluation it was necessary to avoid biases wherever possible. By including different types of unstructured data in the analysis, it was possible to determine if any specific type of data would perform better and allow for generalization of the results. Due to the nature of the investigated methods, only unstructured, textual datasets are

included in this research. To include other types of unstructured data, such as audio logs or images, it would have been necessary to modify or adapt several of the methods, which was not in the scope of this research.

- The first dataset concerns a large set of reviews taken from Amazon with over 870 million entries, containing various user reviews and a rating on Amazon[1] [13]. The set contains reviews from a wide set of categories and includes data as far back as 1996. Due to its large size, only a subset of the data was used. See chapter 4 for the details. Not all columns were used in this research. Only the *text* and *rating* columns were used. The reviews found in this dataset count as a specific type of unstructured data, as users are free to structure and format their review however they want. While they might encounter some total requirements such as maximum length, the data itself is free and unstructured, therefore an excellent inclusion for this research. The total amount of data columns from the dataset are found in table A.1. This dataset shall be referred to as the *Reviews Dataset* further in this document.

- A second dataset that was used in the validation concerns a variety of blog and news articles[2] [7]. The purpose of publishing this set was to facilitate research on news articles and news retrieval techniques. The set includes over 93.000 unique sources from which a total of 1 million articles have been collected. About 25% of the articles are blog posts, while the other 75% are news articles. The dataset contains the fields as explained in table A.3. Due to its size, a subset of this dataset was also taken for this research. How this is done can be found in chapter 4. The news and blog dataset contains examples of unstructured data seen in daily life and is also different from the aforementioned reviews. No restrictions are in place on the contents of these articles, nor are there specific structures for all news and blog articles, allowing for a different perspective on the effects of the processing techniques. This dataset shall be referred to as the *News Dataset* further in this document.

- A third dataset that was used for the validation of the implementation was a dataset with text representations of resumes (Curriculum Vitaes), including those for various positions across different categories[3]. The resumes have specific information replaced, such as CompanyName

[1] The dataset can be found at: https://amazon-reviews-2023.github.io/
[2] See: https://research.signal-ai.com/datasets/signal1m.html
[3] See: https://www.kaggle.com/datasets/pranavvenugo/resume-and-job-description

in place of the actual company, and City for the location, making the data more generalizable. The dataset contains 2.484 entries of different resumes, each of which is an application for one of 24 categories. The dataset contains the fields as described in table A.2. While resumes often are in specific formats and include similar paragraphs, there is not one general structure to them. Every person is free to include or exclude specific headlines or structure it to their liking. This dataset has already flattened the information in the resumes to plain text, allowing processing techniques to be applied directly, therefore providing a great source for the validation. For future references in this document, this dataset shall be referred to as the *Resume Dataset*.

- A fourth dataset used in the validation is an email spam classification dataset[4]. This set contains 5.728 datapoints, of which 5.695 are unique. The dataset consists of a textual representation of the contents, as well as a boolean indication for if it is considered a spam email ('1') or not ('0'). See table A.4 for the contents of the columns for this dataset. As the contents of an email do not have to adhere to a specific structure, it is an excellent additional data source for validation, as it contains distinctly different information than the other investigated datasets. This dataset will be referred to as the *Emails Dataset*.

- The last dataset used to validate the implementation concerns a dataset regarding emotion-detection in social media posts[5]. The dataset is a collection of comments made by users on Facebook, X (formerly known as Twitter), Reddit, and also comments taken from various YouTube videos. The set contains 75.454 unique entries, each of which is assigned one of five primary emotions. These are: happy, surprise, angry, sad, and neutral. The contents of each entry varies greatly, as the length of each message ranges between six and 289 characters long. The dataset only contains two columns, being a "text" column and a "label" column. See also table A.5 for reference. The contents of these comments by users are of similar nature to those present in the Reviews dataset, but also different due to the focus on a rating in the aforementioned set. As only the general emotion associated with the user's comment is recorded, the differences between this set and the Reviews set allow for interesting additional insights. This dataset will be referred to as the *Emotions Dataset* in future references.

[4] See: https://www.kaggle.com/datasets/prernanchan/email-classification
[5] See: https://www.kaggle.com/datasets/gangulyamrita/social-media-emotion-dataset

3.2.2 Data Analysis

As stated before, to find the values for accuracy, recall, precision, and F1-scores as recommended by Kiefer, a prediction needs to be made. For this, multiple classifiers were found and used according to notions by Kumar on their respective strengths and weaknesses [17]. Specific columns for extracting the contents were used, while others were selected for the validation labels. These fields are shown per dataset in table 3.1.

Dataset	**Content**	**Label**
Reviews Dataset	text	rating
News Dataset	content	media-type
Resume Dataset	resume_str	category
Email Dataset	text	spam
Emotion Dataset	text	label

Table 3.1: Contents and Validations labels for the classifier

The implementation of classifiers is made such that a specific label will be assigned based on one of the columns mentioned in table 3.1. All data was left as originally taken from the dataset, as no transformation of labels was deemed necessary for the classifiers to work. The original datasets were only minimally adjusted to determine a baseline score. Steps for this included dropping non-necessary columns and renaming the remaining ones. After the baseline accuracy was determined, the different datasets were processed using the steps as proposed in the processing methods, after which the processed datasets were again classified to determine possible changes. All scores were calculated using the default settings for the classifiers so that the most generalizable results could be given. The results were compared between the classifiers and processed datasets to determine any possible improvement.

As this process was repeated for all datasets, an extensive validation from different perspectives within unstructured data was possible. From the validation, it was possible to learn if the applied processing steps in either one of the comparisons is more accurate and/or more precise. This result allowed us to infer if those steps could be used for other unstructured documents as well and provide indications for which processing techniques improve the quality.

Chapter 4

Implementation

In this chapter, the datasets mentioned throughout section 3.2.1 are applied and implemented to assess the differences in processing techniques on the data quality. This chapter is divided into two sections; one describing which preliminary actions were made to adjust and use each dataset, the second detailing how the classifiers were applied. The other section describes the implementation of each of the processing techniques per method as well as the pipeline for applying them to the datasets. In chapter 5, the results of the implementations are shown for each classifier model, where the resulting accuracy-, F1-, precision-, and recall-scores are reported up on. All code snippets mentioned in this chapter are also included in a github project, aptly named *MasterUnstructuredData*[1]. In this project, one can find the exact code as it is used in this research. With this, one can replicate the results and find more details about the exact implementation.

4.1 Preliminary processing

To predict the class for the test set of the data, various different classifier models can be used. While many different classification models exist, only a selection of classifier models were used within the context of this research.

In the domain of unstructured data classification, Kumar et al. performed a SWOT (*Strength*, *Weaknesses*, *Opportunities*, *Threats*) analysis of various classification models. In this analysis, they found many strengths with regards to speed, accessibility, and interpretability for constituent learning algorithms such as Decision Trees, Naïve Bayes, and SVM models. Other investigated methods include ensemble learning methods such as Random

[1] The project can be found here: https://github.com/Jorisdeboer/MasterUnstructuredData

Forest and Ensemble SVM methods with strengths such as effectiveness, intrinsic simplicity, and robustness [17]. In their research, Kumar et al. found that taking a hybrid approach using multiple algorithms can take advantage of the strengths and cover for weaknesses of individual models, at a cost for the performance time. Because of their scalability, flexibility, and robustness to noise, such ensemble methods often require higher computation time and memory power.

For a thorough test, the four main classifier models described in their research were used to determine the scores for precision, recall, F1, and accuracy. All classification models that were implemented were taken from the Natural Language Toolkit, accessible within Python from the *sklearn* library and include the *Naïve Bayes* classifier[2], a *Decision Tree* classifier[3], a *Random Forest* classifier[4], and the *SVM* classifier[5].

A brief consideration was made to include an additional ensemble method for the SVM classifier. From notes by Kumar, it was clear the SVM model is usually the highest performing model, but it has a downside: the time it takes to run. In table 4.2, one can see that it takes the SVM models significantly longer to run than the other models. For this reason, another ensemble method was considered in the form of an Ensemble SVM method. This method was also implemented in the same manner to test for possible improved accuracy or runtime. As can be seen in table 4.1, the ensemble SVM method is faster on the training sets, but does not always improve on the prediction sets when compared to the normal SVM method. While this is an interesting addition and might be investigated further in future work, it is not considered further in this research.

Dataset: **Ensemble SVM**	Reviews	News	Resume	Emails	Emotions
Runtime					
Training	2309.478	37.111	92.433	6.354	803.948
Prediction	624.221	17.665	22.541	2.499	241.707

Table 4.1: Runtime in seconds of the additional Ensemble SVM method

For the implementation of the datasets, the CSV and JSON files were loaded using standard functions from the *pandas* library of Python. Afterwards, the

[2] https://scikit-learn.org/1.6/modules/generated/sklearn.naive_bayes.MultinomialNB.html
[3] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[4] https://scikit-learn.org/1.6/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[5] https://scikit-learn.org/1.6/modules/generated/sklearn.svm.SVC.html

datasets were cleaned to only contain the columns associated with the text and label necessary for each classifier model.

While a high amount of data points was an advantage for many of the datasets, in practice these large amount of data points were found to be a bottleneck for the implementations as the time for training and testing also increased significantly. This is why a proportional selection was made for the larger datasets based on the labels of the dataset. For a dataset with two labels `"X"` and `"Y"` and a given `[partition]`, the following code was used:

```
# Create separate dataframes for each label
df_label1_count = df[df["label"] == "X"]
df_label2_count = df[df["label"] == "Y"]

# Count instances and take only a [partition] of items.
count_label1 = int(round(len(df_label1_count.index)*
    [partition]))
count_label2 = int(round(len(df_label2_count.index)*
    [partition]))
df_label1_count = df_label1_count.head(n = count_news)
df_label2_count = df_label2_count.head(n = count_blogs)

# Combine frames
df_combined = pd.concat([df_label1_count,
    df_label2_count], ignore_index=True)
```

This code was especially needed for the News and the Reviews datasets, as they both exceeded 1.000.000 data points. The total runtime of the code was collected by assessing the difference between the starting time and the ending time. The runtime can be found in table 4.2, where the total elapsed seconds are shown per original dataset. Because of their long running times, the News dataset was reduced with the above code using `partition = 0.1`, reducing the amount of data points to a proportional 100.000 entries, while the Reviews dataset was reduced using `partition = 0.3` to a proportional 199.248 entries.

In the next step, the data was split into a training- and test-set using the *train_test_split* function from the *sklearn.model_selection* library. The training and test split was made with 80% of the data contained in the training sets, and 20% of the data contained in the test sets. Since each entry contains a large, textual entry that needs to be analyzed, the *TfidfVectorizer* function from the *sklearn.feature_extraction.text* library was used to vectorize the texts. The resulting information was fitted to each of the models, which was then used to predict the labels of the test set. The accuracy score, as well as other scores of this result were then presented using the *accuracy_score* and *classification_report* functions from the *sklearn.metrics* library.

The following function was made to calculate the different scores based on one of the classification models, which could be called upon to quickly get a result. This function works the same for each model-type, as the same steps are used for each of the classification models. The code is also able to keep track of the time spent on each classification.

Dataset: / **Codepiece:**	Reviews	News	Resume	Emails	Emotions
Naïve Bayes					
Training	0.03	0.037	0.032	0.006	0.141
Prediction	0.005	0.003	0.006	0.002	0.006
Decision Trees					
Training	100.506	14.306	2.973	0.823	37.589
Prediction	0.017	0.004	0.001	0.001	0.008
Random Forest					
Training	112.349	4.811	1.174	0.795	99.041
Prediction	0.380	0.044	0.015	0.014	0.248
SVM					
Training	3910.129	69.108	18.796	7.11	2239.071
Prediction	229.971	14.553	2.421	1.562	99.428

Table 4.2: The total time in seconds that elapsed for each codepiece.

```
def classify_data(X_train, y_train, X_test, y_test,
    model_type, print_output):

    # Instantiate model based on type
    if model_type == "rf":
        model = RandomForestClassifier(n_estimators = 25)
    elif model_type == "svc":
        model = SVC(kernel='rbf')
    elif model_type == "dt":
        model = DecisionTreeClassifier(random_state=4)
    elif model_type == "nb":
        model = MultinomialNB()
    else:
        print("not␣a␣valid␣model␣type")
        return

    # Fit the model to the data
    start = time.time()
    model.fit(X_train, y_train)
    end = time.time()

    # print fitting time if true
    if(print_output == True):
        print(f"Done␣training␣dataset␣with␣{model_type}.")
        print("Time␣spent:", end-start, "␣seconds.")

    # Predict classifier from fitted model
    start = time.time()
    prediction = model.predict(X_test)
    end = time.time()

    # print predicting time if true
    if(print_output == True):
        print(f"Done␣predicting␣dataset␣with␣{model_type}.")
        print("Time␣spent:", end-start, "␣seconds.")

    # And print model accuracy reports
    print("============================")
    print(f"Accuracy␣of␣{model_type}:␣", accuracy_score(
        y_test, prediction))
    print(f"====␣Classification␣Report␣====\n",
        classification_report(y_test, prediction,
        target_names=list(map(str,y_test.unique()))))
```

The accuracies calculated from the above code are presented in the chapter on results, in table 5.1, where also the accuracies are presented for the classification on the processed datasets.

4.2 Implementation of processing techniques

The implementation of the different combinations of processing techniques from the unstructured data methods relied on a combination of assessing the descriptions in the original papers, as well as the visualizations presented in those same papers.

For the method of Barbantan, it was found that one needs to start with tokenizing the data, then add Part-of-Speech tags (POS tags), followed by lemmatization [3]. The final step of processing is done by removing the stopwords in the text. In the construction of the original method, the Stanford POS tagger[6] and a WordNet Lemmatizer[7] were used which are therefore included in the implementation of this research as well. In the implemented method, the calculation times could also be requested through the use of a parameter. The code for this can be seen in the snippet below.

```
def barbantan_processing(data, print_output):
    start_method = time.time()
    # lemmatize the words to their base form and extract to a single string
    lemmatizer = WordNetLemmatizer()
    text = str(data)
    doc = nlp(text)
    lemmatized_tokens = [token.lemma_ for token in doc]
    lemmatized_text = ' '.join(lemmatized_tokens)

    # create tokens of the lemmatized text
    doc = nlp(lemmatized_text)
    filtered_words = [token.text for token in doc if not token.is_stop]
    clean_text = ' '.join(filtered_words)
    tokens = word_tokenize(clean_text)

    # create POS tags of the tokenized text and print their class
    pos_tagged = nltk.pos_tag(tokens)
    tagged_words = []
    for word,word_class in pos_tagged:
        concatenated_word = word + "_" + word_class
        tagged_words.append(concatenated_word)

    # transform list result of single postagged words back to a string
    single_postagged = remove_doubles(tagged_words)
    single_postagged = ' '.join(str(s) for s in single_postagged)

    # remove the stopwords
    stop_words = set(stopwords.words('english'))
    word_tokens = word_tokenize(single_postagged)
    filtered_text = [word for word in word_tokens if not word.lower() in stop_words]
    filtered_text = []
    for word in word_tokens:
        if word not in stop_words:
            filtered_text.append(word)

    # make a string of the result
```

[6] See: https://www.linguisticsweb.org/doku.php?id=linguisticsweb:tutorials:linguistics_tutorials:automaticannotation:stanford_pos_tagger_python

[7] See: https://www.nltk.org/api/nltk.stem.wordnet.html

```
    result = ' '.join(str(s) for s in filtered_text)
    end_method = time.time()

    if(print_output == True):
        print("Time spent:", end_method-start_method, " seconds.\n")
    return result
```

Within the method of Sanchez-Seguera, the first step suggests to discard additional columns and any double or missing values [22]. With this initially-cleaned result, the text should be put to lowercase and any stopwords appearing in the text should be eliminated. From here, additional blank spaces and special characters should also be removed. A TFIDF vectorizer[8] should be applied to the resulting text so that classification or other domain-specific steps can be applied. This method also had parameters added to measure the time spent on calculating the result. The function as defined in the below code snippet represents the associated implementation.

```
def sanchez_processing(data, print_output):
    start_method = time.time()

    # first deduplication step happens outside of method, so only lowercasing text
    text = str(data)
    data_lowered = text.lower()

    # filter out stopwords
    stop_words = set(stopwords.words('english'))
    word_tokens = word_tokenize(data_lowered)
    filtered_text = [word for word in word_tokens if not word.lower() in stop_words]
    filtered_text = []
    for word in word_tokens:
        if word not in stop_words:
            filtered_text.append(word)
    filtered_text_string = ' '.join(str(s) for s in filtered_text)

    # remove the blanks
    data_blankless = " ".join(filtered_text_string.split())
    data_result = []
    # and remove the special chars
    sentence = data_blankless.split(" ")
    for word in sentence:
        normal_string = "".join(l for l in word if l.isalnum())
        data_result.append(normal_string)

    # make a string of result
    result = ' '.join(str(s) for s in data_result)
    end_method = time.time()

    if(print_output == True):
        print("Time spent:", end_method-start_method, " seconds.\n")
    return result
```

The last method, defined by Lim, starts with removing all existing stopwords within the to-be-processed text [20]. After these have been taken out, each statement is separated and any additional information is removed. Lastly, tokenization and a POS tagger are used on the resulting text before it receives any other domain-specific steps. The time spent on calculation could again be measured by enabling a parameter in the function. The code snippet below displays the function for the method as described by Lim as it was implemented for this research.

[8] See: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

```
def lim_processing(data, print_output):
    start_method = time.time()

    # remove stop words
    text = str(data)
    stop_words = set(stopwords.words('english'))
    word_tokens = word_tokenize(text)
    filtered_text = [word for word in word_tokens if not word.lower() in stop_words]
    filtered_text = []
    for word in word_tokens:
        if word not in stop_words:
            filtered_text.append(word)
    filtered_text_string = ' '.join(str(s) for s in filtered_text)

    # tokenize for separate words
    doc = nlp(filtered_text_string)
    filtered_words = [token.text for token in doc if not token.is_stop]
    clean_text = ' '.join(filtered_words)

    # use tokenized data for POS tags
    tokens = word_tokenize(clean_text)
    pos_tagged = nltk.pos_tag(tokens)
    tagged_words = []
    for word,word_class in pos_tagged:
        concatenated_word = word + "_" + word_class
        tagged_words.append(concatenated_word)

    # create string result
    result = ' '.join(str(s) for s in tagged_words)
    end_method = time.time()

    if(print_output == True):
        print("Time spent:", end_method-start_method, " seconds.\n")
    return result
```

Using the functions as defined above in this chapter, the implementation was executed. Within the execution of these steps, the initial datasets were first loaded and classified using all four classification models. This resulted in the baseline accuracies needed for the comparison. Afterwards, each dataset was converted to gain a preprocessed unstructured dataset as prescribed by one of the methods. The method was applied using a lambda function on the text column of the dataset with the specific functions of each processing method. A lambda function immediately applies the given argument on each element it is connected to. When applying a lambda of `+1` to a list of numbers for example, it will add `+1` to each item in the list. This function is used to apply each processing method to all elements of the dataset, resulting in a completely converted dataset. Using the resulting sets, another classifier was instantiated to determine the accuracy, precision, recall, and F1-scores. The results of this implementation are described in chapter 5.

Chapter 5

Results

In this chapter, a description of the results from running the code from chapter 4 is presented. In section 5.1, the accuracies from the models are presented in a tabular overview and then shortly described. In the other section in this chapter, section 5.2, the other quality metrics as prescribed by Kiefer for investigating the quality of unstructured textual data are presented.

5.1 Accuracy scores

The code as described in chapter 4 allowed to investigate the accuracy of the classification of the different labels for the different datasets. Below, in table 5.1, an overview of the accuracies of these different models on the different datasets can be found. Each row displays the model and shows the accuracy per dataset, which shows when the original dataset was used to determine the accuracy and when - and which - processed dataset was used.

When exploring the results for the Reviews dataset, one can see that for each classification model the original dataset often has the highest accuracy score listed. According to these *accuracy* results, only the method proposed by Lim showed a slight increase in accuracy, and that only to 0.84 for the SVM classification model, meaning a 0.01 increase. With only a small improvement in one out of four classification models, Lim's method is not showing significant enough results to state that its implementation resulted in this accuracy increase. Since the accuracy as a result from the other methods remained similar, it can be said that the differences in the processing methods are not the sole cause for the differences in accuracy. Differences might have come from outside influences, such as the receptiveness to the processing steps by the contents of the dataset. Some information included in the

Model:	Dataset:	Reviews	News	Resume	Emails	Emotions
Naïve Bayes	*Original*	0.80	0.731	0.51	0.874	0.59
	Barbantan	0.79	0.736	0.48	0.866	0.60
	Sanchez	0.80	0.729	0.49	0.878	0.59
	Lim	0.80	0.716	0.46	0.857	0.59
Decision Trees	*Original*	0.77	0.731	0.56	0.961	0.59
	Barbantan	0.77	0.707	0.58	0.947	0.59
	Sanchez	0.77	0.672	0.62	0.946	0.59
	Lim	0.77	0.702	0.62	0.961	0.59
Random Forest	*Original*	0.82	0.783	0.56	0.964	0.63
	Barbantan	0.82	0.769	0.59	0.978	0.63
	Sanchez	0.82	0.768	0.56	0.963	0.64
	Lim	0.82	0.765	0.53	0.960	0.64
SVM	*Original*	0.83	0.814	0.60	0.994	0.68
	Barbantan	0.83	0.794	0.62	0.989	0.67
	Sanchez	0.83	0.782	0.59	0.991	0.68
	Lim	0.84	0.779	0.59	0.983	0.67

Table 5.1: The accuracies of each model of each (processed) dataset

dataset is a description provided by a user, who might have used language that is filtered out by the processing steps but which might have affected the choice of classified label, increasing one of the quality metrics. Another influence includes possible prior cleaning of data, due to which the chosen processing steps would have had less of an effect.

An exploration of the *accuracy* results in the News dataset shows similar findings. There, it can be seen that only when the dataset was processed using the steps as prescribed by Barbantan and classified by the Naïve Bayes model, the accuracy improves slightly from 0.731 to 0.736, a 0.005 increase. In all other cases, the accuracy was found to be lower than with the unprocessed data. This leads one to believe that differences in the processing methods are not impacting the accuracy for classifying similar information to that of the News dataset.

Third, an investigation of the results for the Resume dataset show different results. For the result with the Naïve Bayes model, the accuracy does not improve by using the processed datasets over the original dataset as the original accuracy of 0.51 remains the highest. On the other hand, the Decision Trees, Random Forest, and SVM models all show slight increases in accuracy with the use of a processed dataset. For example, the accuracy

increases from 0.56 in the original set to 0.59 in Barbantan's set, meaning a 0.03 increase. Interestingly, the biggest improvement can be found in the Decision Trees model, where the original set produces a 0.56 accuracy score and the processed datasets of Sanchez and Lim a 0.62 score, meaning a 0.06 accuracy improvement. Unfortunately, these same processed datasets do not show similar improvements in the Random Forest and SVM models. As this only occurs for one model, it can be said that the type of information from this dataset can benefit from the processing steps for classification purposes with a Decision Tree model. For other types of classification, the processing steps would not impact the accuracy much.

The fourth column, showing the *accuracy* results of the classification applied to the Emails dataset, shows similar results to that of the News and Reviews datasets. As was found for the other datasets, it shows that only one of the processing steps applied to one of the classification models shows slight improvement, while in the majority of the cases the accuracy is not improved at all. In the improvement case, the original dataset shows a score of 0.964 for the Random Forest model, increased to 0.978 when the Barbantan method was used. This means an overall increase of 0.014. All in all, it seems that the differences in the processing steps do not impact the accuracy when using the different classification models for information such as presented in the Emails dataset.

Lastly, the results for the Emotions dataset show no significant differences with regards to the *accuracy* and are comparable to the results of all other columns except those of the Resume dataset. As can be seen in the table, the only improvements of 0.01 in accuracy were found in two cases. The first being when the Barbantan method was used in the Naïve Bayes model, increasing the accuracy from 0.59 to 0.60. The second small improvement was found in the Random Forest model for both Sanchez' and Lim's method, improving from 0.63 to 0.64. The accuracy scores stay the same for the SVM model and the Decision Trees model. With the spread of these results, it cannot be stated if the differences between the processing steps have affected the accuracy when classifying data similar to that present in the Emotions dataset.

5.2 Other quality metrics

Besides the accuracy score, Kiefer also suggested to evaluate the precision, recall, and F1- scores. Due to the large amount of data and limited space, visualizations were created for each of these scores, which can be found in Appendix A.3. In the legend of these visualizations, some of the models are

abbreviated with 'NB' for Naïve Bayes, 'DT' for Decision Trees, and 'RF' for Random Forest. Below, the results are discussed in subsections that split the results per dataset in a small analysis on labels, then focusing one-by-one on the quality metrics.

5.2.1 Reviews Dataset

In this set there were five labels to which the text could be classified. As can be seen when comparing the left and right figure in figure 5.1, the label distribution ratio was kept the same with the code mentioned in chapter 4. The other scores that were kept track of, together with the label-count of these histograms, show additional information about these labels. For instance, the precision for some labels remained similar for the NB model, showing values between 0.73 and 0.79 for label '2' and between 0.79 and 0.8 for label '5'. Interestingly, the precision increased for Barbantan's method for label '1' in DT and in SVM, increasing from 0.18 to 0.44 in DT and from 0.24 to 0.86 in SVM. Unfortunately, similar increases were not found for other labels when this method was applied. Similar one-off increases were found for Sanchez' method as well for DT, where label '4' increased from 0.43 to 0.87. In this case, the more-common label - '5' - decreased from 0.87 to 0.43 leading to believe that the precision might have been affected by the sample taken in these results' training/testing sets.

This trend continues in the results of the recall scores, where the NB model shows consistent results for labels '2' and '5'. Many additional findings show only a one label improvement for a processing method while others decrease. For instance, Sanchez' method again shows for the DT model that label '4' increases from 0.41 to 0.92, showing an 0.51 improvement. However, it also shows a decrease from 0.91 to 0.43 for label '5', a deterioration of 0.48. Similar improvement-deterioration-pairs were found for Barbantan's method for SVM. The RF model only showed a consistency for label '5', ranging between 0.44 and 0.4 but otherwise showing no consistent changes per label. These recall scores are reflected in the F1-scores as well, as this is the resulting average of precision and recall. The consistency of results related to label '5' can be attributed to its high presence in the dataset (see figure 5.1). On the other hand, it is interesting that label '2' has the lowest overall totals, yet still provides high scores.

5.2.2 News Dataset

The this dataset was also subjected to a size reduction and as can be seen in between the left- and right-side of figure 5.2, the distribution of labels was

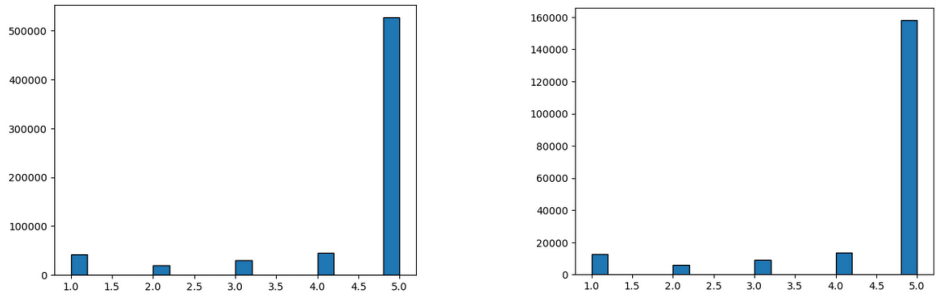

Figure 5.1: Total label counts in full (Left) and combined (Right) for the Reviews dataset

kept the same. The precision score shows overall higher for the '*Blog*' label than for the '*News*' label in the NB model, as its value is often 1.0 instead of between 0.73 and 0.75. Only Sanchez' model deviates from this and has these scores reversed. The results for the DT model show more affection towards the '*News*' label than the '*Blog*' label, with values closer to 0.8 for '*News*' than 0.43 for '*Blog*'. Only when using Lim's method this is reversed. With other classification models, the precision for both labels were roughly similar, with no significant differences between the results. Only slight improvement was found in Sanchez' method for SVM, increasing the precision from 0.81 to 0.84 for the '*Blog*' label yet decreasing from 0.83 to 0.78 for the '*News*' label. Again, it seems that the precision is not consistently increased by applying one specific set of processing techniques.

The results of the recall and F1-scores are different from the precision score, but have a similar distribution. In case of the recall scores, there were many labels identified with a recall score of near 1.0, meaning that most of those labels were correctly identified within the models. For instance, the NB model showed a recall score of 1.0 for the '*News*' and '*Blog*' labels across the processed datasets, but also showed high scores in other models. This includes a score close to 1.0 for the '*Blog*' label in all datasets in the RF model, as well as a score of near 1.0 for the SVM model using Lim's dataset and the original dataset. In the SVM model, the score was also near 1.0 for the '*News*' label for the sets processed with Barbantan's and Sanchez' method. For the DT model, the recall score was close to 80% for the '*News*' label, while around 40% for the '*Blog*' label. These values were the other way around when the dataset was processed using Lim's method. These results show therefore no consistent effect of the applied processing techniques.

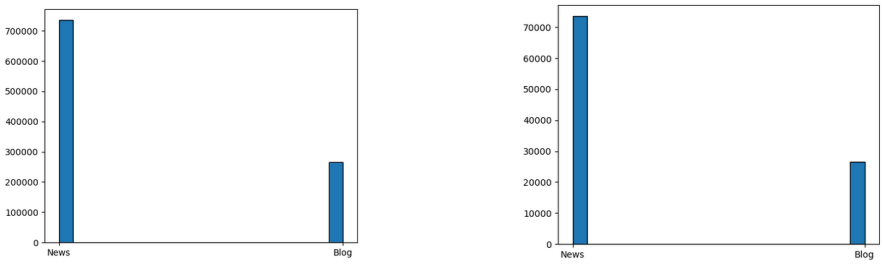

Figure 5.2: Total label counts in full (Left) and combined (Right) for the News dataset

5.2.3 Resumes Dataset

The labels as used in the Resumes dataset are distributed as seen in figure 5.3. There are some labels with only a relative small amount of occurrences, such as '*Automobile*', '*BPO*', and '*Agriculture*'. Other labels occur roughly 100 times or more. Using the information from Appendix A.3, it can be seen that even though the amount of occurrences is relatively lower than the other labels, the precision is still high for the three aforementioned labels in three instances for Lim's method, which will therefore be reported on below. Other methods mainly show decreases in the precision and are therefore not included.

For the DT model, we see an original precision of 0.56 grow to 0.9 for '*Agriculture*', 0.61 grow to 0.94 for '*Automotive*' and 0.27 to 0.81 for '*BPO*'. Other high-occurring labels such as '*Aviation*' and '*Information-Technology*' (IT). While these show a slight improvement in the NB model with 0.45 to 0.54 for '*Aviation*' and 0.17 to 0.33 for '*IT*', they decrease with the DT model application, going from 0.94 to 0.67 for '*Aviation*' and 0.52 to 0.38 for '*IT*', all for Lim's method. Similar increases can be seen for the RF and SVM models for the latter labels, increasing from 0.55 to 0.63 for '*Aviation*' with RF, only increasing from 0.56 to 0.57 with SVM. The '*IT*' label shows a small increase from 0.48 to 0.6 with RF, but lacks significant inclusion in the test set for statements about SVM. The labels with a small amount of occurrences do not see improvements in either RF or SVM model. The fluctuation in precision scores makes it difficult to determine the overall effect, but even a score of almost 1.0 only appears for three out of 24 labels. All in all, this means that there is no consistent effect on the precision for either method.

For the recall scores, the '*Apparel*' label yields the highest overall results for NB, but is not showing consistent improvements with any of the applied processing techniques. Another interesting label to highlight is the '*Designer*' label, which increases with all methods for the RF model. Especially with

Sanchez' method, which increases the recall score from 0.67 to 0.92. Sanchez' method also shows improvement for this label in the DT model (increasing from 0.43 to 0.8), but not in the NB or SVM models. A last label to highlight is the '*Fitness*' label, which also shows better results in the datasets as processed by the Sanchez method than in other datasets, increasing from 0.26 to 0.42 in the DT model or from 0.5 to 0.88 in the SVM model. Unfortunately, other processing techniques mainly show decreases in score for either recall or F1-score. These findings suggests that variations between the different processed datasets and models lead to inconsistency among the results as processed according to any other method than that of Sanchez.

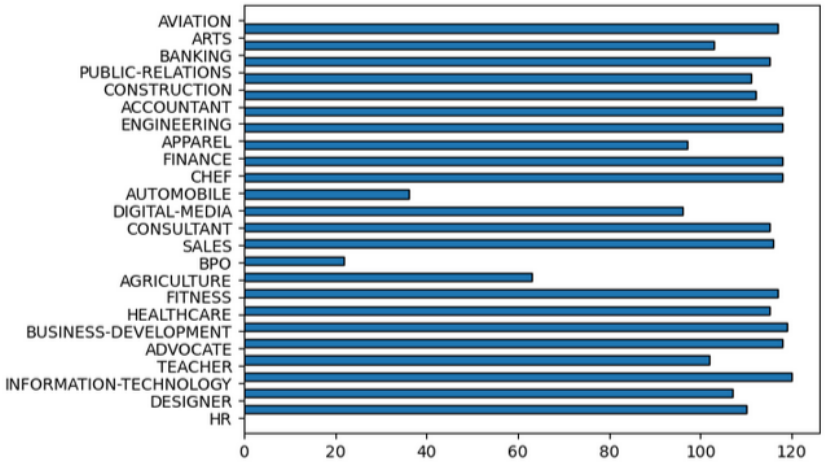

Figure 5.3: Total label counts for the Resumes dataset

5.2.4 Emails Dataset

Next, the precision scores for the Emails dataset show that none of the classification models have difficulty with classifying a data point correctly as spam, as it often equals or nears a score of 1.0. On the other hand, the precision to classify the data point as '0' is only high for DT, RF, and SVM models, as it drops significantly to about 50% for NB when the dataset is processed using Sanchez' method. Small increases can be seen with Barbantan's method in the RF model, where label '0' increases from 0.96 to 0.97, while the other label decreases in value. Other labels remain fairly consistent with the scores produced from the original dataset.

The recall scores for the Emails dataset show a difference between he NB model and other models. This is mainly for label '1', where a maximum recall score between 40% and 50% is achieved. This means that using the

NB classifier, it was more difficult to determine whether the data point was considered spam in comparison to with other models. Interestingly, while the recall score often scored close to 1.0 for label '0', this dropped to 0.9 for the processed dataset using Sanchez's method using RF. The F1-score shows similar results to the distribution of the recall scores, also showing lesser scores for label '1' in the NB models. None of the processing techniques showed an increase of both labels for either of the classification models, the original datasets always scored higher than the processed ones.

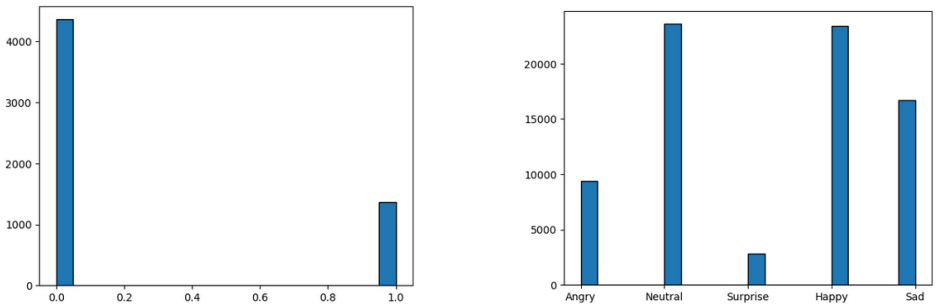

Figure 5.4: Total label counts for the Emails dataset (Left) and the Emotions dataset (Right)

5.2.5 Emotions Dataset

The last dataset to discuss is the Emotions dataset, for which the distribution of labels can be found in figure 5.4. The figure shows a high amount of occurrences for the '*Neutral*' and '*Happy*' labels, a slightly lower amount for the '*Sad*' label, but the lowest amount for the '*Angry*' and '*Surprise*' labels. With this information, together with the scores from Appendix A.3, it was found that the dataset is most accurate for the NB model where the '*Angry*' label improved for both Sanchez' and Lim's method from 0.56 to 0.91 and 0.89 respectively. Unfortunately, this label had only small improvements - if any - for the other classification models. The '*Happy*' label showed improvements from 0.56 to 0.71 with Sanchez' method within NB, as well as an improvement of 0.61 to 0.8 for SVM with Sanchez' method. Other methods and classification models did not effect this label much. The '*Sad*' label showed an improvement within the RF model for Barbantan's method by increasing from 0.57 to 0.77, but otherwise was not significantly affected by other models or processing techniques. The '*Suprise*' label only improved with Sanchez' method, going from 0.68 to 0.74 in the RF model and from 0.31 to 0.56 in the DT model. As can be seen, there are some merits for

calculating the precision with Sanchez' method when facing data similar to that of the Emotions dataset, but results are not very consistent.

The recall score remains high for the '*Angry*' label throughout the different models, except when Sanchez' and Lim's methods are used in the NB model. The recall score is also lower across the models for the '*Surprise*' label, even while the score is between 60% and 70% when the processed dataset with Barbantan is used for the NB model, as well as when the dataset is processed with Sanchez' and Lim's method for the DT model. Similar results are presented in the F1-scores for this dataset, where the '*Sad*' label is consistently around 60%, the '*Angry*' and '*Happy*' labels are between 60% and 70%, and the '*Surprise*' label has varying scores over the different models and processed datasets. As some scores are higher and some are lower throughout the different processed datasets and models, the varying scores cannot give definite conclusions for a better-performing set of processing steps.

Chapter 6

Discussion

In this chapter, the research project and the results are discussed. First, the implications from the results are discussed in section 6.1, after which the limitations are discussed in section 6.2. This chapter concludes by providing some indications and directions for future work in section 6.3.

6.1 Implications

6.1.1 Accuracy scores

The results described in chapter 5 show that the accuracy between the differently processed datasets and across the classifier models is very similar. While some of the data would support claims for a slight improvement in accuracy, the overall results do not support these insights. As described in the previous chapter, some results were found where the accuracy improved slightly using a processed dataset. However, as these results were never showing a consistent improvement, there was not enough evidence to determine whether the processing steps were the sole factor for the accuracy improvement. This implies that the differences encountered in the accuracies of the classifier models have been caused by chance due to the random selection for the training and testing subsets of data, or through other means outside of this research such as processing applied before the data was collected.

While the main implication of the results is that the different order of processing steps do not necessarily influence the accuracy of a model, it seems that the subject of the dataset, as well as the amount of labels and the datatype of the label, can give a slight indication for the resulting accuracy. In chapter 5, a comparison was made between an original, unprocessed

dataset and a processed dataset. First, the Reviews and Emotions dataset both have five labels, yet a different label datatype. While their accuracies stay consistently within 0.01 from the accuracy of the original dataset, the overall accuracy of the Reviews dataset was higher. While both datasets contain informal, user-generated data, the Reviews dataset contains numerical labels and the Emotions dataset contains textual labels. With the data being pulled from reviews from Amazon for the Reviews dataset and from comments and posts from different social media platforms for the Emotions dataset, it could mean that the data sources do not have content that is rich enough in terms of structuring for the processing steps to have a noticeable effect. The informal nature and the difference in datatypes therefore could have affected the results.

This is also supported by the comparison of the News and Emails datasets, where only two labels were used in the classification. Here, the News dataset had textual labels and provided an overall lower accuracy, while the numerical labels of the Emails dataset provided an overall higher accuracy. While the accuracy may have increased slightly for the NB and RF models due to the nature of the unstructured data itself, no conclusive effects of the processing steps could be found for any improvements. The Resume dataset also supports these notions, as the high amount of textual labels provided a high variance in data, leaving more uncertainty and therefore becoming less accurate. It seems that there was often not enough training data to support a good fit for each label. Due to the spread of the results it could not be concluded whether the processing steps themselves were the sole reason for improvement. While the formal nature of a resume could indicate more structure and therefore a higher score, the spread of the data was too big for providing a significant result.

6.1.2 Other quality metrics

The precision varied a lot for the NB and DT models when more than two labels were used in classification, but remained more consistent for the RF and SVM models across the different datasets. When only two labels were used, the precision was sometimes higher for the DT model, but not consistent enough to draw conclusive results. For the results of the NB models, the scores also varied too much for generalizable results.

Comparing the scores for the different processing methods, it could be seen that the precision is not improving consistently enough within one or more methods to determine a generalized improvement. When classifying for more than two labels, the datasets processed with Barbantan's and Lim's methods show more consistency with a subset of the labels in the RF and

SVM models, but the remaining labels otherwise vary too much. Any possible improvements can therefore not solely be attributed to the effect of the processing steps due to this inconsistency. The results also vary a lot for the NB and DT models, where the scores show even less labels with consistent precision. While the results for the datasets processed with Sanchez' method show more consistency with the DT, RF, and SVM models, they are not consistent with the results of the NB model. For only two labels, the results are not conclusive enough as the results for the NB and DT models is too different between the datasets. This means that while the RF and SVM models perform better, the effect of the processing steps is too inconsistent to determine a better-performing processing technique for a generalizable result.

Similar observations were made for the recall and F1-scores. The RF and SVM models both show better scores in general, but when comparing the results across the different processing methods, there is not enough consistency between the results. When there are more than two labels present, all processing methods have at least one label that has up to 50% difference between two classification models. For the NB and DT models, this gap is even bigger, often scoring between 10% and 60% lower between models. With only two labels present in the dataset, the scores for the DT, RF, and SVM models are more similar. However, the scores are also similar when compared among themselves. Due to this similarity, there is not one single processing method providing significantly better results.

All in all, the precision, recall, and F1-scores are showing higher results when more advanced classification methods are used, but are too similar for the different processed datasets to gain generalizable results. While the RF and SVM models provide better results for the datasets with a higher amount of labels, the DT model can also be used as an important asset when only two labels are used in classification. Since the scores are not improving consistently for one of the processed datasets, the results are not conclusive enough to determine if one of the processing methods investigated in this research was the sole cause for any improvement. This would imply that the differences between the processing techniques of the unstructured data methods do not influence the resulting metrics.

These implications show that organizations should focus less on trying to find an 'ultimate set of processing techniques' for a one-size-fits-all solution. While some general processing techniques as found in the fields of *Information Retrieval* and *Text Mining* can be applied to datasets - for instance for the purpose of tokenization - practitioners focused on improving the quality of the data should focus their attention an adjusting their processing tech-

niques to their specific domain. Follow up is necessary to determine if specific processing steps *within domains* might be generalizable. Focus on data quality with the same metrics but different processing steps allow organizations to ensure the value of their knowledge assets. When these knowledge assets become inaccessible and inaccurate due to poor data quality management, the decision-making process is affected greatly. Since these assets are the foundation of the decisions, it is of utmost importance to keep them accessible and accurate through other quality approaches.

6.2 Limitations

6.2.1 Validity Concerns

The credibility of a research can be threatened by various validity concerns. In his paper on validity threats in research, Yu describes different categories that represent a critical reflection on the contents and execution of research [31]. Yu describes threats to the Internal Validity as the parts of the research that could affect the outcome of the treatment as well as the evidence that supports these claims. He describes External Validity as the generalizability of the outcome across different settings or domains. The threats to these areas of validity are discussed below.

Internal Validity
The internal validity can be undermined by different factors. In the context of this study, one such factor concerns the instrumentation and interpretation. For the comparison between the methods, the paper's contents were interpreted as-is, meaning that techniques were noted in the order of occurrence. Due to possible interpretation-errors, it is possible that the exact processing steps and their order is not the same as in the described paper and requires further investigation. To combat this, both the main paper and papers referenced in the studies were investigated to complete the knowledge on the method.

These validity concerns extend into the optimization and execution of the methods to process the data. The researcher has put in his best efforts for interpreting the steps as mentioned in the original works at the base of this research. However, with the possibility that some steps were meant in a specific order, no further optimizations were applied. While this can make the underlying code run inefficient, it could have also slightly affected the resulting datasets. When a specific step was applied in a different way, this could also have resulted in a different data point which could have been clas-

sified differently. By providing the code directly in this research, a measure of re-usability was provided.

These concerns are also reflected in how extensive the results are processed. The code for splitting up the data, processing them according to the different methods, and classifying the labels was ran multiple times and provided similar results each time. Unfortunately, these additional results were not explicitly recorded for each run. However, it was observed that there were only slight changes each time. For every run, it took a long time to process the datasets, sometimes reaching up to 8100 seconds for training a single set. Due to the time spent on processing, and with the previously observed results in mind, it was decided to not repeat the described process to note down the additional results. The average of the resulting multiple data points could strongly support the notions in this research that the processing techniques have no significant impact on the accuracy and other quality metrics. However, due to the associated time concerns and the following inability to finalize the research in time, the results of the last run were provided as the final results in this thesis. These were used to reason over the implications in the discussion.

Another threat concerns the selection of data that is included in the comparison of methods. All methods are situated within a different domain: Requirements Engineering, Medical Prediction, and Knowledge Asset Discovery. If the existing dataset of one of the domains is used for the comparison, it would affect the outcome of the other methods more than the one that already used it. To equalize the field, different datasets are chosen that contain a different type of unstructured data and is applied in a different domain for each dataset.

The concerns around the datasets are also present in the chosen sets for this study, as the research was conducted using datasets curated outside the scope of the researcher. This means that while there was some knowledge on the contents and the origin of the datasets, it was unknown how this information was specifically collected and processed. This information could have been processed in such a way prior to its use in this study so that some processing steps might have had less effect than when no prior processing was done at all. This is especially relevant for the steps used in the method of Sanchez-Segura, as those steps specifically mention to filter out information that could have affected the outcome.

External Validity

There are also threats to the generalizability of the results of this study. For instance, the origin of the different methods used in the comparison

could have affected the generalizability of the results. However, because the comparison is performed with three different methods from different domains, the results could be stated to be generalizable. This is also affected by the included processing steps suggested in these methods. Because of their origin, it is possible that these steps are too closely related to the domain and therefore have no effect, or are not useful to include for the classifier that follows the processing steps. To mitigate this, only the steps that are affecting the text outside of the domain-context are included.

Another threat to the external validity can be found in the type of content of the datasets. In this research, five different types of unstructured data were used, but there are many more different types in existence. There could have been types that are more receptive to the specific steps applied in this research, showing a more significant effect. By not including the full range of types of unstructured data, this research can not fully provide a generalizable result. The researcher has tried to mitigate this by only selecting a few of the unstructured data types he encountered in the literature.

6.2.2 Personal Limitations

Personal limitations were also present in this research. This takes form in the researcher's analysis and interpretation skills. While the results in the aforementioned chapters are shown numerically and visually per dataset, the interpretation of these results were not always straightforward. During the comparison of results related to each processing method per dataset, it was difficult to determine the best way to display the information. To keep information consistent, bar charts related to the labels were used in the visualizations, which are not the best for detailed comparisons. On a similar note, the Resume dataset contained so many labels that it became difficult to directly gain precise results for each processing method. Comparisons were made to the best of the researcher's ability, but some mistakes could still have been present in the interpretation by misjudging certain scores.

Another personal limitation to this research lies in the duration of the thesis project. As this project took over a year to fully complete, some information and aspects could have been forgotten or misremembered over time, leading to inaccurate results. The researcher has tried to mitigate this by staying up to date with the literature and other resources that were at his disposal with each addition to the research.

6.3 Future Work

The research highlighted various areas of opportunity for further research. One of these areas includes to improve the data collection and curation aspect. One way to improve this is to broaden the textual data types that need to be included in the research. In the current research, only emails, news messages, blog posts, reviews, resumes, and social media messages were used. In future work, this can be extended to include other textual types such as invoices, receipts, metafield descriptions, and more. Another way to improve the data collection aspect for future work is to conduct the experiment with self-curated datasets from a use-case within a company and not use datasets from online sources. From a use-case at a company, data can be acquired that is directly relevant for classification which will also provide a better control over specific pre-processing steps that can be applied.

Another aspect interesting for future work is to investigate a wider range of processing methods and finding the overlapping steps between them. This allows to construct a generalized method that could also yield interesting results. In general, future work could benefit from a broader range of unstructured data methods to investigate as well as going over possible optimizations for more streamlined results.

A further area for future work lies in investigating different unstructured data types besides only the textual one. In the current research, only textual unstructured data was investigated, leading to specific processing techniques for textual data and comparing those. It would be an interesting area to investigate different types such as audio, image, or video data, and comparing the techniques used in these with those that can be found in the textual methods.

Lastly, future work can focus itself on building a framework with regards to the findings on data curation teams. It was found that data curation teams are a potential focus for organizations in order to improve the management of their data. In this research, a first area of focus was introduced by pointing at the differences in processing techniques that are found within different domains and understanding what their impact is on the quality. Future work could build on these ideas and focus on finding and creating a conceptual framework that defines the boundaries and way of working of the data curation teams. This way, specific recommendations can be made on a business-level to practitioners, focusing more on the business-side than the data-side.

Chapter 7

Conclusion

In this research, an analysis and investigation is presented on how different ways of handling unstructured data might affect the business value of organizations. The analysis consists of a theoretical investigation and a practical implementation in order to provide an answer to the main research question: "*How can unstructured data be used to improve business value?* "

While the research problem constructed for this research concerned itself with the effects of common factors within processing techniques found in unstructured data methods, these effects are mostly interesting due to the context that was found in the literature in the answers for SQ1 and SQ2. Using these answers, it was found that there are some general recommendations for organizations to apply to improve their business value with unstructured data.

The first of these recommendations suggests that a business should primarily focus on their knowledge assets to improve their business value. This is a direct consequence of information found for SQ2. Organizations always have specific information from various internal sources that they use to steer the organization in a certain direction. It was found that it is important to structure this information in such a way that it can be found and acted upon. Without the accessibility of this information, decision making can be influenced as certain factors can be overlooked. The retainment of these knowledge assets is therefore key to improving the decision making and improving business value in the long term.

Another recommendation concerns itself with the application of data stewards and data curators in data management teams to improve accountability and management of data within an organization. Many references found in the answer on SQ2 support this. In this literature it was stated

that these data stewards and curators should be part of a curation team where they focus on assessing the quality of the data that is produced as well as making the data more accessible. By improving the inherent quality of the data, the aforementioned knowledge assets can be structurally improved. The quality is improved by focusing on the interpretability and relevancy in an iterative process, adding information to the knowledge assets to make them clearer in each iteration. As these knowledge assets help in the general decision making, it is important to keep track of how and why these are constructed. Other findings in the literature suggest to add metadata information to the unstructured documents for this, so that information can be provided on the provenance of data, the usage and application of data, and to add information for accountability.

The above recommendations rest on the assumption that the data used to construct the knowledge assets is accurate. Inaccurate data leads to inaccurate conclusions, which can wrongly be included in associated knowledge assets. To combat inaccuracies, specific processing techniques are often applied when handling data. In the literature, multiple methods were found when investigated SQ3 and SQ4, that handle unstructured data in specific ways for different domains. This research aimed to find generalizable effects within the processing steps from various domains.

The investigation of three such methods for an answer to SQ5 was achieved by implementing their respective processing techniques. The resulting processed datasets were then investigated using metrics suggested in the literature, among which the *Accuracy* and *precision*. While it was found that there were some effects, the results were inconclusive on whether these effects could be attributed to the specific processing techniques or on other factors.

All in all, for an organization to improve their business value using unstructured data, it is recommended to improve quality, accessibility, and accountability. This can be done by focusing on data curation teams with specialized data stewards who can focus on applying their knowledge to give meaning to the data. They can use a variety of methods to achieve this, as the results of the practical implementation show that there is no significant improvement among the investigated metrics. These results of this study could be attributed to the underlying quality of the datasets or on other factors. Therefore, an organization should not focus on a one-size-fits-all solution, but rather on how their data teams could ensure data quality in a different way to ensure the value of their knowledge assets. Organizations need to focus on how they handle the quality aspect of their data, otherwise impacting their decision making negatively.

Bibliography

[1] Ifeyinwa Angela Ajah and Henry Friday Nweke. Big data and business analytics: Trends, platforms, success factors and applications. *Big data and cognitive computing*, 3(2):32, 2019.

[2] Alation. Metadata management methodology, 2024.

[3] Ioana Barbantan and Rodica Potolea. Knowledge extraction and prediction from unstructured medical documents. *ICT innovations*, 2015.

[4] Anne Burmeister and Jürgen Deller. Knowledge retention from older and retiring workers: What do we know, and where do we go from here? *Work, Aging and Retirement*, 2(2):87–104, 2016.

[5] Eric Chu, Akanksha Baid, Ting Chen, AnHai Doan, and Jeffrey Naughton. A relational approach to incrementally extracting and querying structure in unstructured data. In *Proceedings of the 33rd international conference on Very large data bases*, pages 1045–1056, 2007.

[6] William F Cody, Jeffrey T Kreulen, Vikas Krishna, and W Scott Spangler. The integration of business intelligence and knowledge management. *IBM systems journal*, 41(4):697–713, 2002.

[7] David Corney, Dyaa Albakour, Miguel Martinez, and Samir Moussa. What do a million news articles look like? In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.*, pages 42–47, 2016.

[8] Pedro DeRose, Warren Shen, Fei Chen, AnHai Doan, and Raghu Ramakrishnan. Building structured web community portals: A top-down, compositional, and incremental approach. In *Proceedings of the 33rd international conference on Very large data bases*, pages 399–410. Citeseer, 2007.

[9] AnHai Doan, Jeffrey F Naughton, Raghu Ramakrishnan, Akanksha Baid, Xiaoyong Chai, Fei Chen, Ting Chen, Eric Chu, Pedro DeRose, Byron Gao, et al. Information extraction challenges in managing unstructured data. *ACM SIGMOD Record*, 37(4):14–20, 2009.

[10] Alon Y Halevy, Oren Etzioni, AnHai Doan, Zachary G Ives, Jayant Madhavan, Luke K McDowell, and Igor Tatarinov. Crossing the structure chasm. In *CIDR*, 2003.

[11] Joeri Heijnen. Social business intelligence: How and where firms can use social media data for performance measurement, an exploratory study. 2012.

[12] Jerry R Hobbs and Ellen Riloff. Information extraction. *Handbook of natural language processing*, 2, 2010.

[13] Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

[14] William H Inmon and Anthony Nesavich. *Tapping into unstructured data: Integrating unstructured data and textual analytics into business intelligence*. Pearson Education, 2007.

[15] William H Inmon, Bonnie O'Neil, and Lowell Fryman. *Business metadata: Capturing enterprise knowledge*. Morgan Kaufmann, 2010.

[16] Cornelia Kiefer. Assessing the quality of unstructured data: An initial overview. In *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", Potsdam, Germany*, pages 62–73, September 12-14, 2016.

[17] Akshi Kumar, Vikrant Dabas, and Parul Hooda. Text classification algorithms for mining unstructured data: a swot analysis. *International Journal of Information Technology*, 12(4):1159–1169, 2020.

[18] TK Ashwin Kumar, Hong Liu, and Johnson P Thomas. Efficient structuring of data in big data. In *2014 international conference on data science & engineering (ICDSE)*, pages 1–5. IEEE, 2014.

[19] Moria Levy. Knowledge retention: minimizing organizational business loss. *Journal of knowledge management*, 15(4):582–600, 2011.

[20] Jin Wei Lim, Thiam Kian Chiew, Moon Ting Su, Simying Ong, Hema Subramaniam, Mumtaz Begum Mustafa, and Yin Kia Chiam. Test

case information extraction from requirements specifications using nlp-based unified boilerplate approach. *Journal of Systems and Software*, 211:112005, 2024.

[21] David Plotkin. *Data stewardship: An actionable guide to effective data management and data governance*. Academic press, 2020.

[22] Maria-Isabel Sanchez-Segura, Roxana González-Cruz, Fuensanta Medina-Dominguez, and German-Lenin Dugarte-Peña. Valuable business knowledge asset discovery by processing unstructured data. *Sustainability*, 14(20):12971, 2022.

[23] Roger Sapsford and Victor Jupp. *Data collection and analysis*. Sage, 1996.

[24] Rolf Sint, Sebastian Schaffert, Stephanie Stroka, and Roland Ferstl. Combining unstructured, fully structured and semi-structured information in semantic wikis. In *CEUR workshop proceedings*, volume 464, pages 73–87. Citeseer, 2009.

[25] Scott Spangler, Ying Chen, Larry Proctor, Ana Lelescu, Amit Behal, Bin He, Thomas D Griffin, Anna Liu, Brad Wade, and Trevor Davis. Cobra–mining web for corporate brand and reputation analysis. *Web Intelligence and Agent Systems: An International Journal*, 7(3):243–254, 2009.

[26] Scott Spangler and Jeffrey Kreulen. *Mining the talk: Unlocking the business value in unstructured information*. Pearson Education, 2007.

[27] Paul P Tallon, Kenneth L Kraemer, and Vijay Gurbaxani. Executives' perceptions of the business value of information technology: a process-oriented approach. *Journal of management information systems*, 16-4:145–173, 2000.

[28] Saravanan Thirumuruganathan, Nan Tang, Mourad Ouzzani, and An-Hai Doan. Data curation with deep learning. In *EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark*, pages 277–286, 2020.

[29] Saurabh Tiwari, Deepti Ameta, and Asim Banerjee. An approach to identify use case scenarios from textual requirements specification. In *Proceedings of the 12th Innovations in Software Engineering Conference (formerly known as India Software Engineering Conference)*, pages 1–11, 2019.

[30] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[31] Chong-ho Yu and Barbara Ohlund. Threats to validity of research design, 2021. Retrieved from: https://creativewisdom.com/teaching/WBI/threat.shtml, on 17-03-2024.

List of Figures

List of Tables

Appendix A

Appendix

A.1 Protocol for Systematic Literature Review

In this appendix the approach to the literature review is discussed. The steps and considerations are posed below.

A list of search terms and keywords is made based on the topic of interest. A general- (or basic-)search is made with an initial query on the topic. The results of the query shows a list of papers and studies that require more investigation. The investigation is done by reading titles, abstracts, introductions, and conclusions. When related or interesting terms come forward in this approach, they are added to the list of search terms and keywords, and the associated paper is added to a longlist.

From the papers described in the longlist, related references in the bibliography, as well as mentions of the author(s) in other work are collected and investigated. This method is a *backward snowballing* approach, where other work that is possibly relevant is also found and included.

From a longlist of relevant sources, a shortlist of around 25 to 30 items can be created with an *explicit* selection criteria. Per item on the shortlist, a small summary is written and the relevancy to the overall topic is stated. Together, these summaries and the insights on the relevancy of each topic is collected and the main findings from this list is used to answer the associated questions on the topic of interest.

A.2 Dataset Structures

Field	Type	Explanation
rating	float	the rating of the product (1.0 - 5.0)
title	str	the title of the review
text	str	the text body of the review
images	list	a list of images associated with the review
asin	str	the Amazon Standard Identification Number
parent_asin	str	the parent ASIN of the parent product
user_id	str	the ID of the reviewer
timestamp	int	the time of the review in unix time
verified_purchase	bool	a verification of the reviewer's purchase
helpful_vote	int	the amount of helpful votes for the review

Table A.1: Data fields used in the Reviews Dataset from Amazon

Field	Type	Explanation
id	int	a unique integer reference to the resume
resume_str	str	the textual representation of the resume
resume_html	str	the resume including all html tags
category	str	the category of the resume

Table A.2: Data fields used in the Resume Dataset

Field	Type	Explanation
id	str	a unique identifier for the article
content	str	the textual content of the article
title	str	the title of the article
media-type	str	either "News" or "Blog"
source	str	the name of the article source
published	str	the publication date of the article

Table A.3: Data fields used in the News Dataset from Signal

Field	Type	Explanation
text	str	the combination of subject and body of the email
spam	bool	a boolean value marking an email as spam or not

Table A.4: Data fields used in the Email Dataset

Field	Type	Explanation
text	str	the comment as taken from the social media platform
label	str	a label marking the main emotion that is detected

Table A.5: Data fields used in the Emotion Dataset

A.3 Dataset Scores

Below, visualizations are included of the associated precision, recall, and F1-scores for the five datasets that were investigated for this thesis. The first set of scores show the values for the Reviews dataset, the second correspond to the News dataset, the third show set of scores belong to the Resumes dataset, the fourth belong to the Emails dataset, and the last scores are associated with the Emotions dataset.

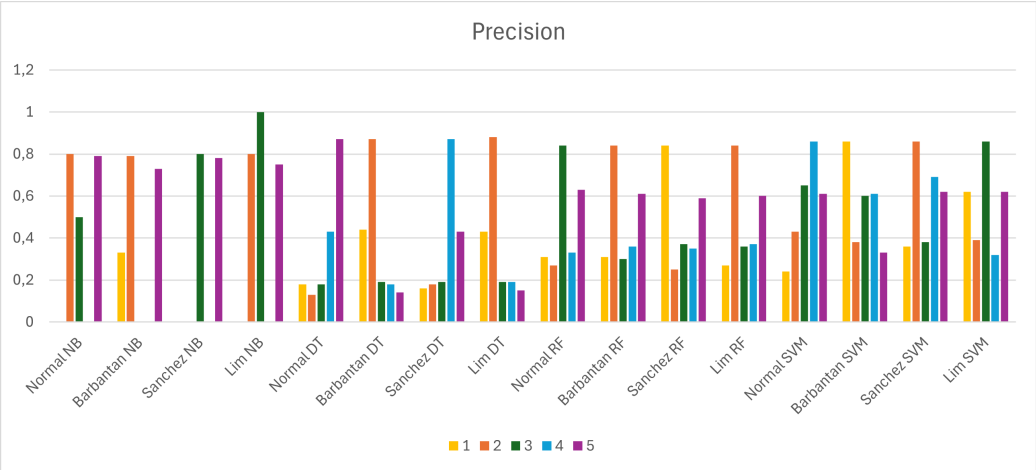

Figure A.1: *Precision scores of Review dataset*

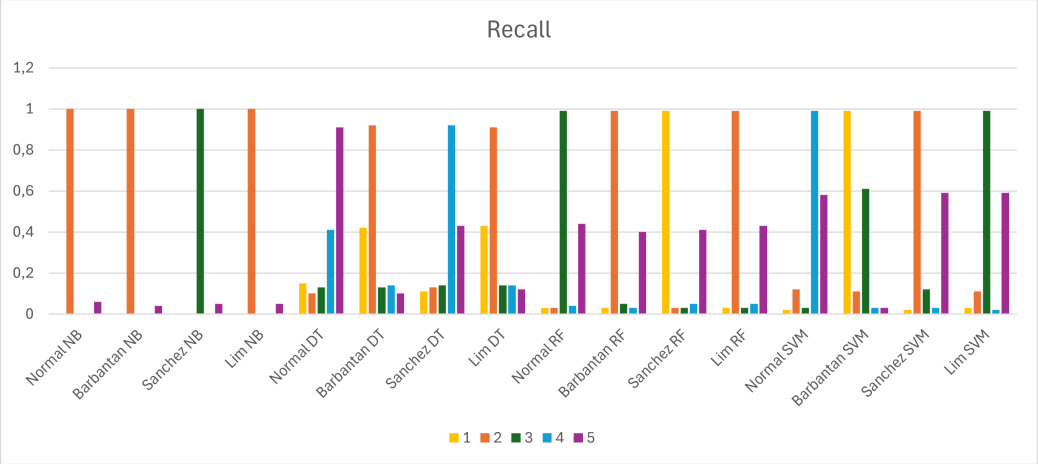

Figure A.2: *Recall scores of Review dataset*

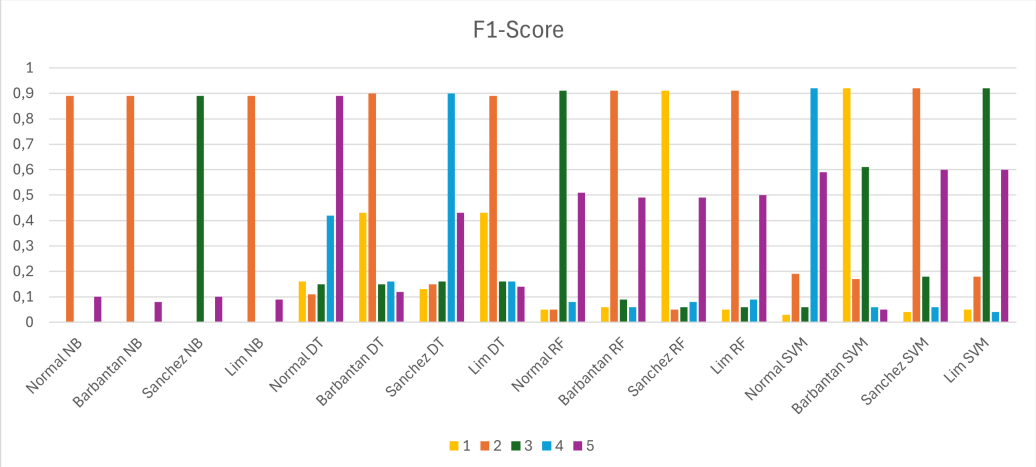

Figure A.3: *F1-scores of Review dataset*

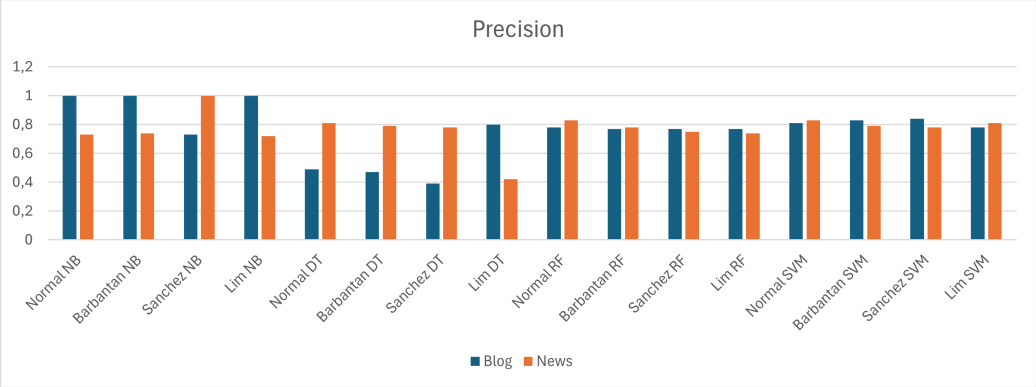

Figure A.4: *Precision scores of News dataset*

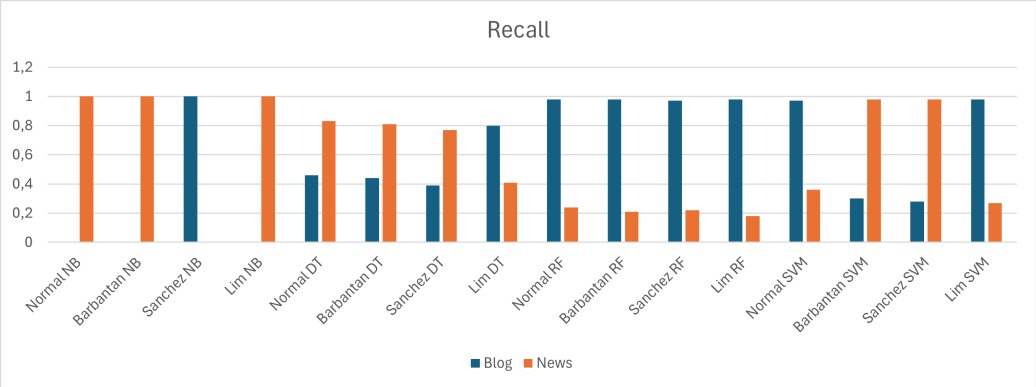

Figure A.5: *Recall scores of News dataset*

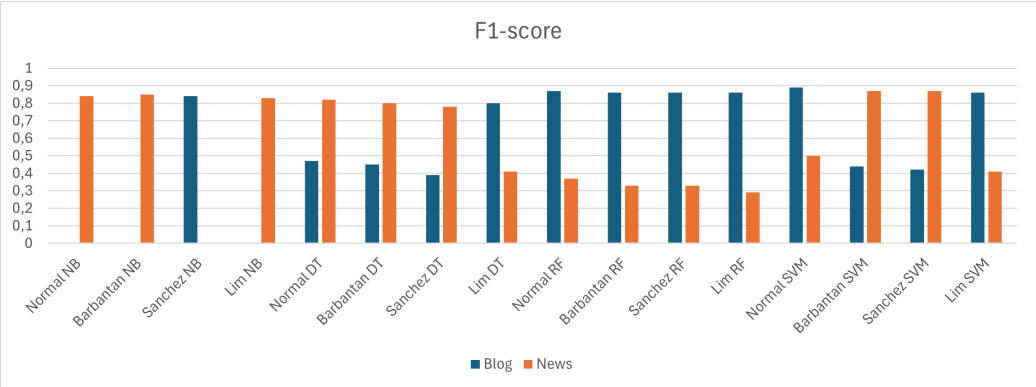

Figure A.6: *F1-scores of News dataset*

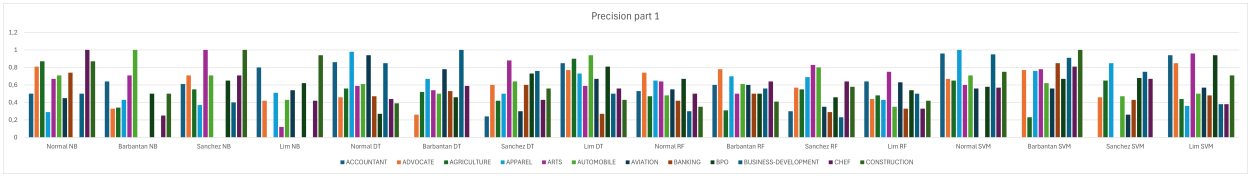

Figure A.7: *Precision scores part 1 of Resumes dataset*

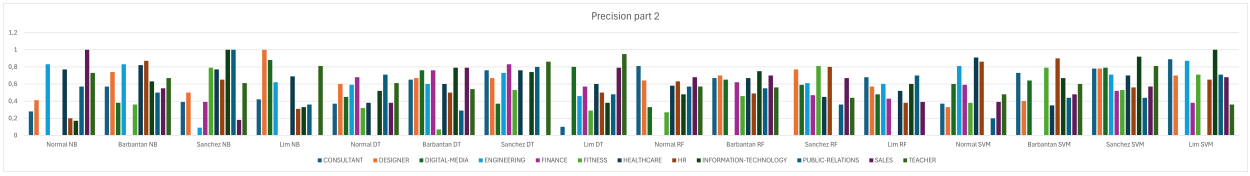

Figure A.8: *Precision scores part 2 of Resumes dataset*

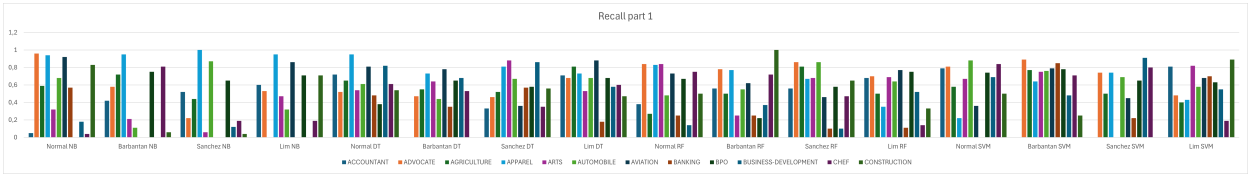

Figure A.9: *Recall scores part 1 of Resumes dataset*

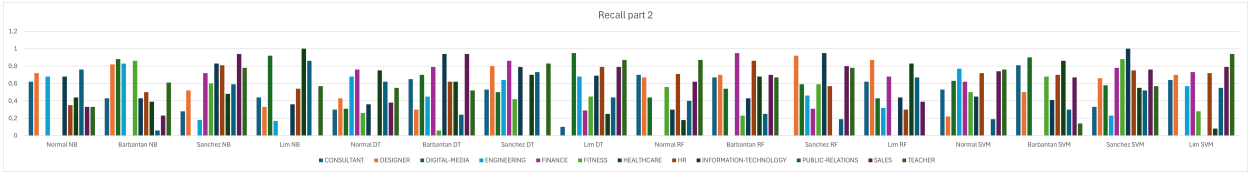

Figure A.10: *Recall scores part 2 of Resumes dataset*

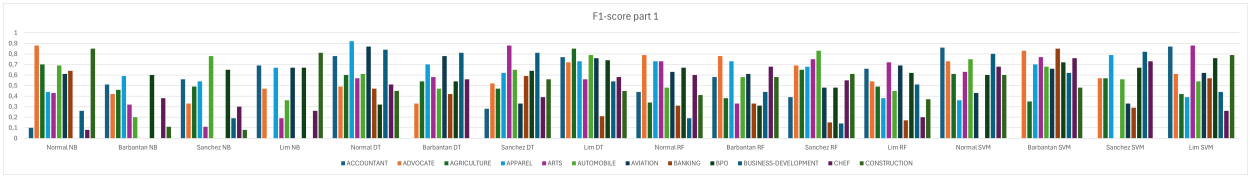

Figure A.11: *F1-scores part 1 of Resumes dataset*

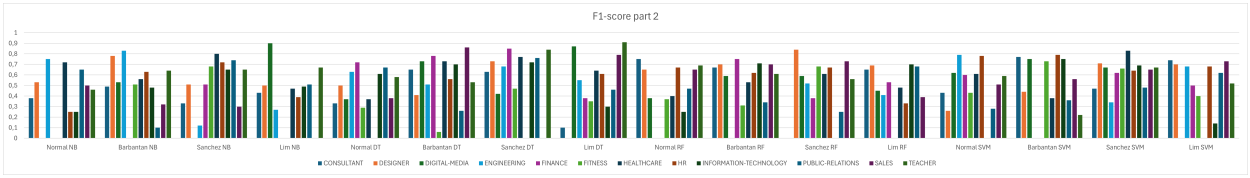

Figure A.12: *F1-scores part 2 of Resumes dataset*

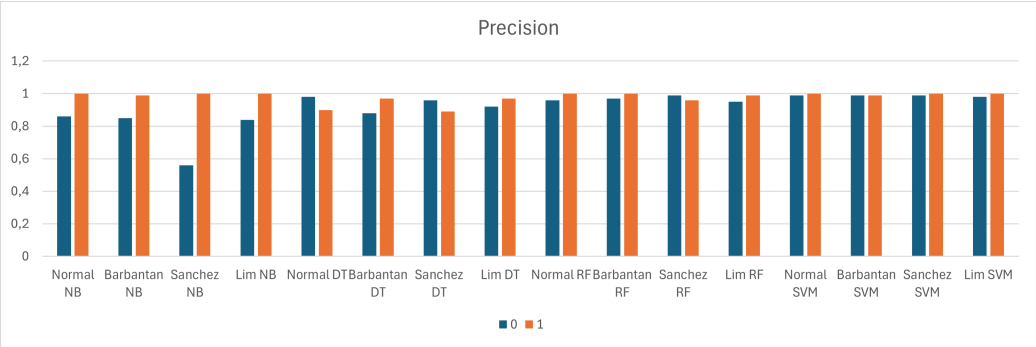

Figure A.13: *Precision scores of Emails dataset*

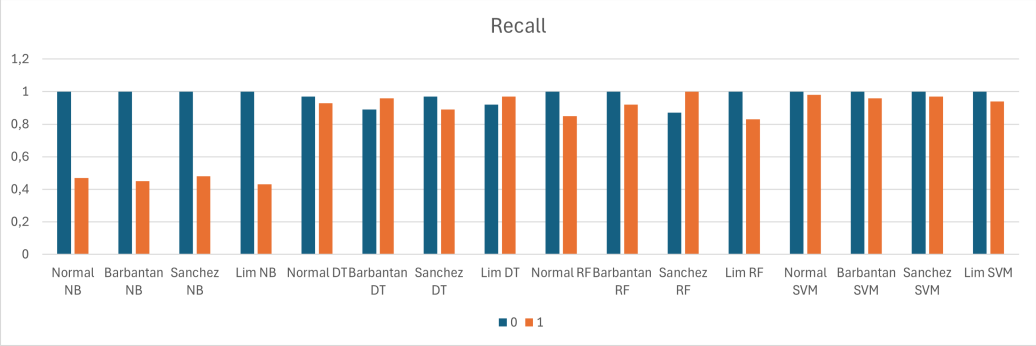

Figure A.14: *Recall scores of Emails dataset*

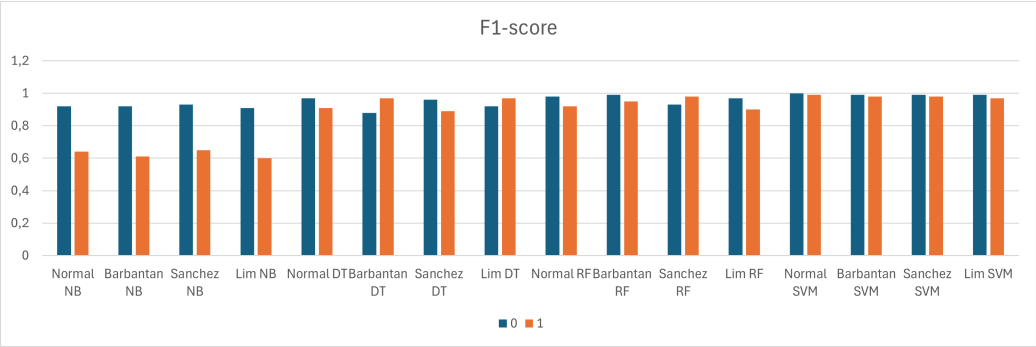

Figure A.15: *F1-scores of Emails dataset*

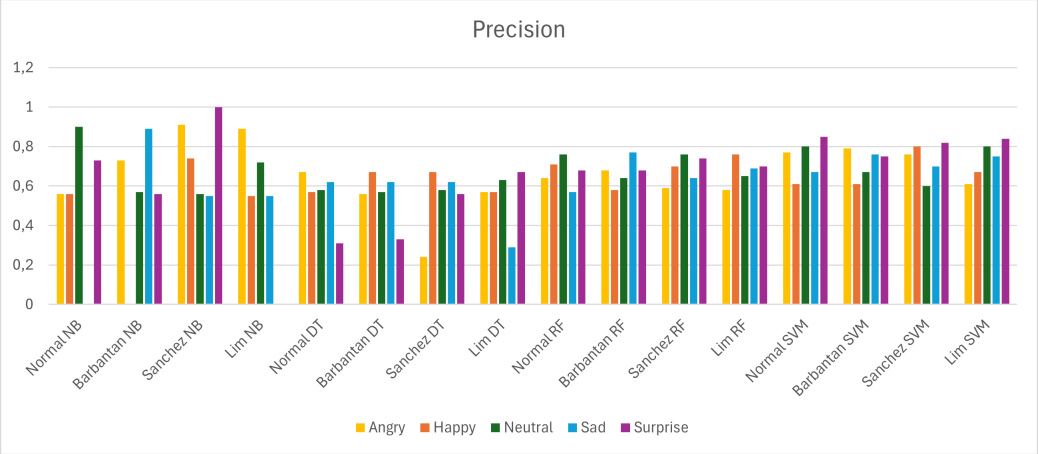

Figure A.16: *Precision scores of Emotions dataset*

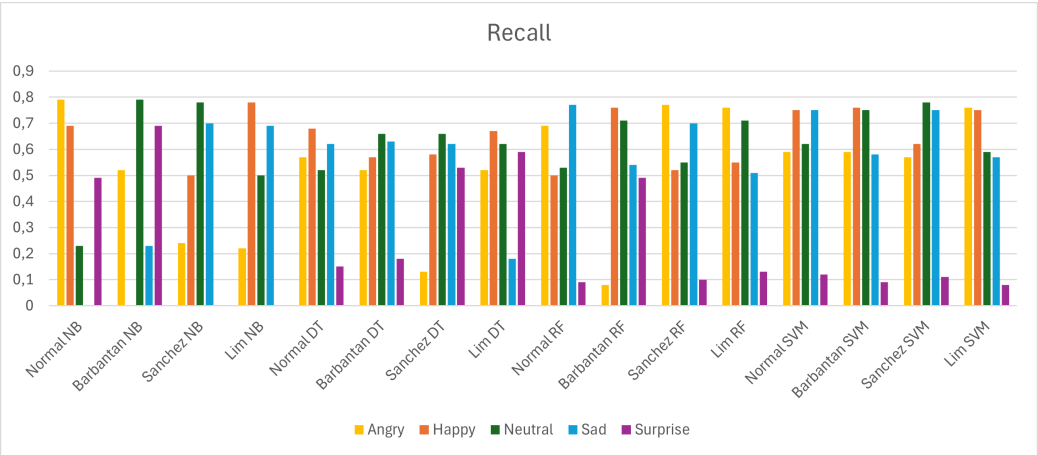

Figure A.17: *Recall scores of Emotions dataset*

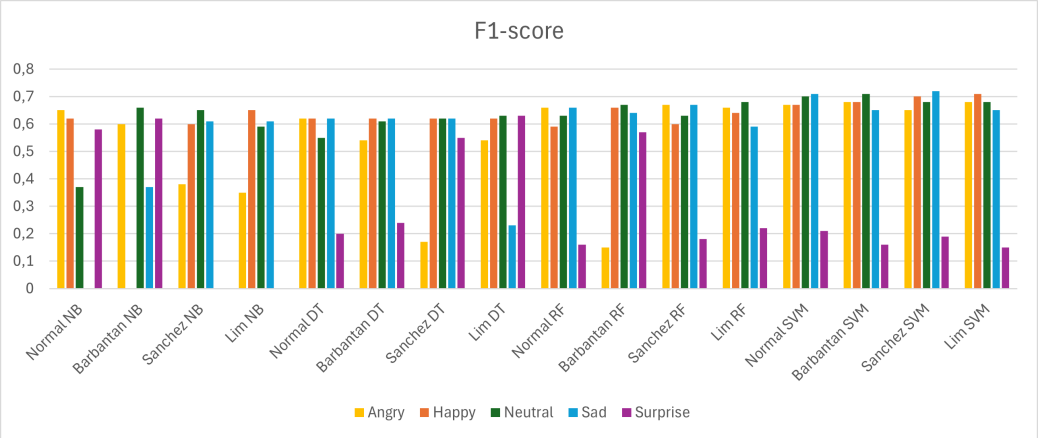

Figure A.18: *F1-scores of Emotions dataset*