Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

Ragnhild Kleiven and Solveig Hergot Langås

Machine Learning for Adverse Event Analysis in Healthcare

Identifying Opportunities and Conducting a Classification Study

Master's thesis in Computer Science
Supervisor: Øystein Nytrø
Co-supervisor: Melissa Yan
June 2023

NTNU
Norwegian University of
Science and Technology

Ragnhild Kleiven and Solveig Hergot Langås

Machine Learning for Adverse Event Analysis in Healthcare

Identifying Opportunities and Conducting a Classification Study

Master's thesis in Computer Science
Supervisor: Øystein Nytrø
Co-supervisor: Melissa Yan
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Computer Science

NTNU
Norwegian University of
Science and Technology

Abstract

Adverse events, which refer to unintended harm caused to patients, serve as a crucial indicator of patient safety within hospitals. While some are unavoidable, research has revealed that many adverse events are preventable. Thus it is crucial for hospitals to have a systematic approach to address adverse events, enabling analysis and insights for preventive measures to enhance patient safety.

This thesis aims to research how applications of machine learning can aid in improving the adverse event reporting and analysis process at St. Olav's Hospital. With access to a dataset of 46,087 adverse event reports spanning from 2015 to 2022, the thesis explores how the textual data in the reports can be utilized with machine learning to enhance the current process.

The thesis has two main objectives: firstly, identifying potential areas within the adverse event reporting and analysis process where machine learning can contribute to improvement, and secondly, selecting a specific area to conduct a machine learning study in order to validate its feasibility and reliability. To gain insights into the current process, interviews were conducted with clinicians and experts. These interviews provided valuable insights into the existing process and helped identify areas that could benefit from enhancements.

One area identified from the interviews where the automatic classification of adverse events according to the National guidelines for classification of patient-related adverse events (NOKUP). This application can potentially improve the current process by addressing issues such as misclassification and inconsistent categorization that hinder accurate analysis and identification of preventive measures. With these benefits in mind and confirming the interest of the clinicians, this application was selected for the machine learning study.

The machine learning study aimed to verify the feasibility of the classification of adverse events into predefined categories. It explored the potential of two distinct classification techniques, Näive Bayes (NB) and Support Vector Machine (SVM). Several experiments were conducted to optimize the performance of both classifiers. The resulting macro F1-score for NB was 0.7182 and 0.7165 for SVM. Although these results could be considered reliable in other domains, further improvement is needed before the models could be implemented in a healthcare context. However, they demonstrate the potential of automatic classification of adverse events.

Thus, this thesis has provided a foundation for the automatic classification of adverse events, demonstrating its significance and potential. Moreover, in collaboration with clinicians, two additional machine learning applications have been identified, providing valuable insights for future research directions.

Sammendrag

Uønskede hendelser er en nøkkelindikator for pasientsikkerhet på sykehus. Slike hendelser innebærer utilsiktede pasientskader, og selv om noen av disse er uunngåelige, viser forskning at mange uønskede hendelser kan forhindres. For å bedre pasientsikkerheten er det viktig å jobbe systematisk med å analysere hendelsene og derav identifisere forebyggende tiltak.

Denne masteroppgaven har som mål å utforske hvordan anvendelse av maskinlæring kan bidra i prosessen med å håndtere uønskede hendelser. Med utgangspunkt i 46,087 avviksrapporter, rapportert mellom 2015 og 2022, utforsker studien hvordan maskinlæring kan anvendes på den tekstlige dataen i disse rapportene for å forbedre dagens avvikshåndtering.

Masteroppgaven har to hovedmål: å identifisere områder innenfor avviksrapportering og -analyse hvor maskinlæring kan bidra til forbedring, og å utføre en maskinlæringsstudie for ett av de identifiserte alternativene. Studien har som mål å undersøke mulighetene og potensialet til det valgte alternativet. For å skaffe innsikt om hvordan avvikshåndteringen foregår i dag, samt identifisere områder for forbedring, har det blitt gjennomført forskningsintervjuer med helsepersonell som er eksperter på området.

Et av de identifiserte alternativene var klassifisering av uønskede hendelser i henhold til de Norsk kodeverk for uønskede pasienthendelser NOKUP. Automatisk klassifisering kan forbedre den nåværende avvikshåndteringen ved å redusere feilklassifisering og inkonsekvent kategorisering, noe som i dag forhindrer nøyaktig analyse og identifisering av forebyggende tiltak. Basert på tilbakemeldinger fra ekspertene ble dette alternativet valgt for maskinlæringsstudien.

Maskinlæringsstudien hadde som mål å verifisere potensialet for klassifisering av avviksrapporter i henhold til NOKUP kategoriene. To ulike klassifiseringsteknikker, Näive Bayes (NB) og Support Vector Machine (SVM), ble implementert og evaluert for denne klassifiseringsoppgaven. Flere eksperimenter ble gjennomført for åoptimalisere klassifiseringmodellene. NB fikk en endelig macro F1-score på 0.7182, og SVM fikk et resultat på 0.7165. Selv om disse resultatene kan bli ansett som pålitelige i andre domener, er det nødvendig med videre forbedring av modellene før integrering i helsesystemer. Derimot viser de potensial for automatisk klassifisering av uønskede hendelser.

Masteroppgaven har lagt et grunnlag for automatisk klassifisering av uønskede hendelser i helsesektoren, og demonstrert potensialet i å anvende maskinlæring på avviksrapporter. I samarbeid med helsepersonell har det også blitt identifisert to andre potensielle bruksområder for maskinlæring innen avvikshåndtering som skaper innledende grunnlag for fremtidig forskning.

Preface

This master's thesis was conducted and written at the Norwegian University of Science and Technology (NTNU) during the spring of 2023. It represents the culmination of our Master of Science in Computer Science and signifies the conclusion of our academic journey as students. This project would not have been possible without the invaluable resources and support provided by the people involved, whom we would like to acknowledge.

We extend our gratitude to our co-supervisor, Melissa Yan, who was been an exceptional source of support throughout the thesis, offering valuable advice and guidance whenever needed. We would also like to express our appreciation to our supervisor, Øystein Nytrø, for providing us with the opportunity to work in such an interesting field of research and for the resources made available for this journey. Additionally, thanks to the research projects Computational Sepsis Mining and Modelling (CoSeM) and Machine Learning for Adverse event Analysis (MLAA) for funding this master's thesis.

We would like to thank all the clinicians and domain experts who have dedicated time in their busy schedules to provide valuable insights. Special thanks are extended to the employees at St. Olav's Hospital and Helse Nord-Trøndelag (HNT) who participated in the research interviews, as their contributions were instrumental in understanding the domain of adverse event reporting and analysis.

Furthermore, we express our gratitude to our fellow computer science students for engaging in discussion and sharing knowledge in stressful times. And finally, we want to thank our family and friends for all their support and encouragement.

Ragnhild Kleiven and Solveig Hergot Langås

Trondheim, June 8, 2023

Contents

List of Figures

List of Tables

List of Acronyms

ANOVA	Analysis of Variance
BOW	Bag-of-words
CHI2	Chi-squared
CNB	Complement Näive Bayes
CoSeM	Computational Sepsis Mining and Modelling
DAGSVM	Directed acyclic graph SVM
DSR	Design science research
EDA	Exploratory Data Analysis
EQS	Extend Quality System
EU	European Union
FN	False Negative
FP	False Positive
GDPR	General Data Protection Regulation
GTT	Global Trigger Tool
HEMIT	Helse Midt-Norge IT
HNT	Helse Nord-Trøndelag
IDF	Inverse Document Frequency
IG	Information Gain
MNB	Multinomal Näive Bayes
MI	Mutual Information

MLAA	Machine Learning for Adverse event Analysis
NB	Näive Bayes
NLP	Natural Language Processing
NOKUP	National guidelines for classification of patient-related adverse events
NTNU	Norwegian University of Science and Technology
REK	Regional Committee for Medical and Health Research Statistics
RBF	Radial Basis Function
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negative
TP	True Positive
WHO	World Health Organization

Chapter 1

Introduction

This thesis aims to investigate the application of machine learning in the reporting and analysis of adverse events at St. Olav's Hospital. Adverse event reporting and analysis play a crucial role in identifying and implementing preventive measures and enhancing patient safety in hospitals. The objective is to explore how machine learning can contribute to this process by potentially streamlining workflows and reducing workload. The thesis has a dual focus: firstly, identifying areas within the adverse event reporting process where the application of machine learning can aid, and secondly, conducting a practical study with regard to one of the identified areas. The machine learning study involves the classification of adverse events using supervised machine learning algorithms. The dataset used for this thesis is unexplored and holds significant potential for improving learning outcomes and enhancing knowledge in this domain.

This chapter serves as an introduction to the thesis, providing an overview of the background, motivation, overall objective, and research questions. Additionally, it outlines the research methodology, contributions, and structure of the thesis.

1.1 Background and Motivation

The field of healthcare is constantly evolving, with a growing emphasis on patient safety and quality of care [Baker et al., 2004]. An important indicator of patient safety at hospitals is adverse events, defined as unintended harm caused to a patient during the process of medical care [Baker et al., 2004]. Adverse events pose significant challenges to healthcare organizations worldwide, with an estimated annual toll that surpasses the combined deaths caused by AIDS and

breast cancer [Donaldson et al., 2000; De Vries et al., 2008].

Even though there has been an increase in research on patient safety and adverse events, the specific application of machine learning in the context of adverse event reporting and analysis seems to be relatively unexplored. The motivation for this project is thus to identify where and how machine learning can contribute to making adverse event processing less resource-demanding, as well as increasing the learning outcome. The term learning outcome refers to the objective of learning from previous incidents to prevent similar incidents in the future.

In Norway, health institutions are obliged to continuously work with improving patient safety, including monitoring and evaluating adverse events [Government of Norway, 2020]. The objective of reporting adverse events is to learn from previous incidents to avoid similar events in the future. This is done by monitoring the amount and type of adverse events in order to detect measures for improvement.

Traditionally, the process of registering and analyzing adverse events at Norwegian hospitals has been a manual and time-consuming process, relying on healthcare professionals to register, categorize and analyze incident reports. However, this approach is prone to human error, inconsistency, and subjectivity. As a part of this thesis, it has been conducted a machine learning experiment with the aim of evaluating if machine learning algorithms can contribute to improving the categorization of adverse events. More specifically, this project examines if machine learning can classify adverse events according to predefined categories.

1.2 Goal and Research Question

The overall goal of this thesis is defined as follows:

Research Goal: *Contribute to the improvement of patient safety in Norwegian hospitals by exploring how machine learning could be utilized to improve the processes of reporting and analyzing adverse events.*

The objective of this research goal, along with the subsequent research questions, encompasses two main aspects: first, the investigation of areas where machine learning can contribute to the adverse event reporting and analysis process, and second, the practical validation of one of these identified areas through a machine learning study. By addressing both theoretical exploration and practical implementation, this thesis aims to provide a comprehensive understanding of the potential applications of machine learning in adverse event reporting and analysis.

Research Question 1*: What are the challenges of the adverse event reporting*

and analysis process, and how can the application of machine learning aid in these challenges?

The first part of this research question emphasizes the need to gain a thorough comprehension of the tasks and the involved clinicians in today's processes. The last part of the research question highlights the need to get the perspectives of clinicians when identifying machine learning applications, ensuring both the utility of and interest in the solutions. For both parts of the research question, it is necessary to establish a dialogue with the clinicians involved in the processes.

***Research Question 2**: To what extent can a selected application of machine learning deliver reliable results for healthcare professionals working with adverse events?*

Based on the findings related to **Research Question 1**, we aim to conduct a practical study involving one of the identified applications of machine learning. The term *reliable* highlights the need to determine the extent to which the given application of machine learning can deliver accurate and trustworthy outcomes in a healthcare context. This implies the need to evaluate the results thoroughly together with clinicians.

1.3 Research method

This section provides an overview of the research methodology employed in this thesis to address the research questions. The research method encompassed interviews with clinicians and experts, as well as a machine learning study focusing on the automatic classification of adverse events, aiming to bridge the gap between theory and practical application.

The interviews conducted with clinicians and experts aimed to gather insights into the current adverse event process at St. Olav's Hospital and identify opportunities for applying machine learning techniques to improve the process. By engaging with healthcare professionals, the research aimed to ensure that the proposed solutions aligned with the practical needs and challenges faced in the real-world healthcare setting.

One of the key areas identified for machine learning applications was the automatic classification of adverse events. To assess the feasibility of this approach, a machine learning study was conducted. The study involved training and evaluating two classification techniques using a dataset containing adverse events recorded from 2015 to 2022. Preprocessing steps, feature extraction, and feature selection methods were also explored to optimize the classification performance. The results obtained from the classification were further evaluated by the clini-

cians involved, providing valuable feedback on the practicality and effectiveness of the proposed approach.

For a more comprehensive explanation of the research methodology, including details on the interview process and the experimental methodology employed for the automatic classification, please refer to Chapter 4.

1.4 Contributions

This thesis makes significant contributions in an area that has received limited attention thus far: the application of machine learning to adverse event processes at Norwegian hospitals. The key contributions of this thesis are as follows:

- Documenting the current workflow of adverse events at St. Olav's Hospital, providing a comprehensive understanding of the processes involved.
- Identifying three potential areas where machine learning could contribute to improving the adverse event reporting and analysis process
- Implementing the application of supervised machine learning for the classification of adverse events, paving the way for further research in an area that has confirmed interest by numerous clinicians.
- Evaluating two classification techniques on Norwegian clinical texts, demonstrating the practicality and effectiveness of these models within this specific context

1.5 Thesis structure

This section outlines the structure of the thesis. The following provides an overview of each chapter and its contents:

Chapter 2 provides a comprehensive overview of the relevant concepts and theories, establishing the necessary background theory for the thesis.

Chapter 3 explores prior research on adverse event reporting, machine learning applications to adverse events, and related work in the field.

Chapter 4 provides an overview of the methodologies used in the thesis, including the thesis timeline, interview methodologies, and the experimental plan for the machine learning study. It also covers the tools and codebase employed in the study.

Chapter 5 presents the dataset and provides descriptions of the relevant content. In addition, the chapter includes the results from EDA performed on the dataset, offering insights into its characteristics and patterns.

Chapter 6 describes the results obtained in the thesis, including domain insights, potential application for machine learning to enhance the adverse event reporting and analysis process, and the experiments conducted for the machine learning study

Chapter 7 evaluates and discusses the methodology and the results from the research interviews and the machine learning study classifying the adverse events into the NOKUP categories.

Finally, **Chapter 8** concludes the thesis by summarizing the work done and providing answers to the research questions presented in Section 1.2. It additionally offers a reflection on the contributions and potential directions for future work.

Chapter 2

Background Theory

The goal of this chapter is to present the theory needed to understand the thesis. This includes domain knowledge about clinical adverse events, Norwegian laws regarding adverse events, as well as a technical section covering the steps of text classification.

2.1 Clinical Adverse Events

This project focuses on patient-related adverse events that occur within a clinical setting. Throughout the project, the term adverse event will specifically refer to in-hospital clinical incidents that affect patients. Adverse events can be defined as unintended injuries or complications caused by medical management, not the patient's underlying disease [De Vries et al., 2008; Brennan et al., 1991]. These events can include medical errors, infections, adverse drug reactions, falls, and other types of harm that result in prolonged hospitalization, disability, or even death. Adverse events can occur due to a variety of factors, such as breakdowns in communication, medication errors, misdiagnosis, and inadequate staffing levels.

It is important to note that not all in-hospital adverse events are preventable. For instance, unanticipated allergic reactions to medicine are an unavoidable consequence of health care [Baker et al., 2004]. However, not all adverse events are unavoidable. According to Baker et al. [2004], studies have shown that between 37% and 51% of adverse events could have been potentially prevented in hindsight. De Vries et al. [2008] has performed a systematic review of in-hospital adverse events and presents that in the 8 studies used, the median of adverse events judged to be preventable was 43.5%. These studies show that even though

some adverse events are unavoidable, a significant amount is preventable.

2.1.1 Adverse Event Reporting and Analysis

Clinical adverse events are an important indicator of patient safety at hospitals [Baker et al., 2004]. The systematic handling of adverse events is essential for advancing patient safety by gaining insights into the prevention of similar incidents. This process of learning extends across various levels, encompassing individuals, groups, organizations, and society as a whole. Establishing a comprehensive system for reporting and analyzing adverse events becomes imperative in order to achieve meaningful learning outcomes beyond the individual level [Donaldson, 2002].

Incident reporting and analyzing is a key mechanism for learning from previous situations, and continuously improving patient safety [World Health Organization et al., 2020]. By collecting data on adverse events, healthcare organizations can analyze trends, identify risk factors, and take steps to prevent similar events from occurring in the future. Reporting adverse events can help identify and address systemic issues in the healthcare system. For instance, consistent reporting of adverse events associated with a specific medical device or drug prompts a closer examination of its safety and efficacy, highlighting the importance of proactive measures in ensuring patient well-being.

2.2 Adverse events in Norway

Hospitals in Norway are required to work systematically to improve patient safety and quality of care under the Act on Specialist Health Services (Lov om Spesialisthelsetjenesten) [Government of Norway, 2023, 2020]. This includes creating preventive measures and learning from adverse events. The Regulation on Leadership and Quality Improvement in Health and Care Services (Forskrift om ledelse og kvalitetsforbedring i helse- og omsorgstjenesten), which is based on the mentioned act, provides more specific guidance on how healthcare providers should establish and operate these internal systems for reporting, analyzing, and monitoring adverse events [Government of Norway, 2020]. In addition to internal reporting systems, it is specified that the organizations should have quality and patient safety committees [Government of Norway, 2023].

The Act on Specialist Health Services also specifies that severe adverse events should immediately be reported to the Norwegian Board of Health Supervision (Statens helsetilsyn) and The Norwegian Board of Health Supervision Investigation Commission (Statens undersøkelseskommisjon for helse- og omsorgstjenesten, Ukom) [Government of Norway, 2023]. A severe adverse event refers to a

situation in which a patient experiences either death or severe injury, with the outcome being unexpected considering the foreseeable risk involved [Government of Norway, 2021]. The Norwegian Board of Health Supervision is a governmental agency responsible for supervising and monitoring healthcare services in Norway. They are tasked with ensuring that healthcare services are provided in accordance with applicable laws and regulations and that patient safety and quality of care are prioritized [Norwegian Board of Health Supervision, 2023]. The Norwegian Board of Health Supervision Investigation Commission is an independent commission appointed by the Ministry of Health and Care Services, with a mandate to investigate serious incidents and events in the healthcare system [Government of Norway, 2021]. They have the responsibility to identify any systemic issues and make recommendations for improvements to prevent similar incidents from occurring in the future.

In a report published by the Norwegian Board of Health Supervision, they received 1849 reports about severe adverse events in 2022 [Norwegian Board of Health Supervision]. These reports were either from health personnel obligated to report or patients and family members that have the right to report deaths or severe events in relation to health services. Among the total of 1849 reports, 915 are from the specialized healthcare service that includes hospitals. The average number of reports from the specialized healthcare service from 2019 to 2022 is 768.

2.2.1 National guidelines for classification of patient-related adverse events (NOKUP)

All adverse events that are reported in Norway should be categorized in accordance with NOKUP. NOKUP is the national guideline for how to classify patient-related clinical adverse events in Norway. NOKUP was developed on behalf of the Norwegian Ministry of Health and Care Services between 2013 and 2015 [Saastad et al., 2015]. The purpose of NOKUP is to identify serious system failures, to pinpoint areas for learning and improvement, and to gain an overview of problem and risk areas. Having national guidelines allows for consistency in reporting and comparison of adverse events across different hospitals and healthcare systems. It also facilitates the collection of data on adverse events at a national level, which can help identify patterns and areas for improvement in patient safety.

NOKUP is based on the World Health Organization (WHO) classification system [World Health Organization et al., 2005] which is a guideline for the development and implementation of adverse event reporting and learning systems created by the World Alliance for Patient Safety. Since NOKUP is built upon the WHO

classification system, it enables the potential to compare adverse events with other nations in the future, promoting a collective effort to improve patient safety on a global scale.

The NOKUP classification system consists of 7 categories: (1) place of the event, (2) type of event, (3) contributing factors, (4) preventability, (5) degree of severity, (6) frequency of the adverse event, and (7) possible consequence in case of repetition.

For this project, the focus will only be on category number two: type of event. When selecting this category, the reporter must consider the nature of the adverse event and identify the issue and the complications that occurred. The categorization involves selecting from ten different event types, each of which may have up to three levels of subcategories. If a subcategory is available, the reporter must select the most specific one. However, for the purpose of this project, our focus will be limited to the top level of the categories and not their subcategories. The types of events are Patient Administration, Diagnostics and Examination, Treatment and Care, Falls and Patient Accidents, Documentation and Information, Infection, Drugs, Medical Equipment, Patient Behavior, and Blood, Cells, Tissue, and Organs. The following is a summary of the types of events described by Saastad et al. in NOKUP. In the cases where St. Olav's Hospital has deviations from the NOKUP categories, it is clearly stated. The differences are also summarized in Table 2.1.

Patient Administration

Patient Administration is defined as administrative matters related to patient care, such as logistics and support functions. Events related to the planning of the patient's stay or course of treatment within the health care system should be classified as Patient Administration. This encompasses events related to appointments, waiting lists, transfer of patient responsibility, admission, and discharge. At St. Olav's Hospital, this type of event is called Patient Administration and Cooperation as it also includes events related to collaboration and interaction with external institutions. This has been added as the hospital has collaboration agreements with other institutions in the municipality.

Diagnostics

Diagnostics is the second type of event and implies events related to diagnostic tests, measurements, samples, and assessments. Events can be classified as Diagnostics if the event is performed at the wrong time, is an examination that is incorrectly or inadequately performed, or is a procedure that is performed incorrectly or at the wrong body part. It should be noted that while laboratory

services are defined as a subcategory of Diagnostics in NOKUP, at St. Olav's Hospital, they are considered a distinct type of event. As the data used in this project is from St. Olav's Hospital, we will treat Laboratory Services as a distinct type of event. Further, at St. Olav's Hospital, they have included Treatment and Care in this type of event.

Treatment and Care

Treatment and Care include incidents related to the choice of treatment, preparation, execution, or treatment delay. It also includes events related to childbirth and pressure ulcers. As mentioned, at St. Olav's Hospital they have merged the NOKUP categories Diagnostics, and Treatment and Care into a type of event called Diagnostics, Treatment, and Care.

Falls and Accidents

Falls and Accidents are used when a patient has fallen or been subject to an accident. It can be during treatment, transportation, or while resting in a bed or chair. At St. Olav's Hospital, this category is identical to the description in NOKUP.

Document and Information

The type of event called Documentation and Information should be used for incidents where documentation or information has been the issue. Examples can be missing or incorrect information in the patient journal, missing or incorrect documentation of the patient's identity, or missing or unclear information given to the patient and their next of kin. It can also include shortcomings related to information security and privacy.

Infection

Infections that were not present before the hospitalization, but have occurred during the stay, should be classified as the type of event named Infection. This can for instance be sepsis, urinary tract infection, respiratory infection, or wound infection.

Drugs

Incidents related to the use of or the preparation and admission of drugs should be classified as Drugs. Even though drugs are used as a part of treatment, the incident should be reported as Drugs, and not Treatment and Care if the adverse event was related to a mistake with the preparation or administration of the drug.

Incidents related to the unavailability of drugs, the distribution of drugs, or side effects are also included in this type of event.

Medical Equipment

If medical equipment has caused an adverse event, the incident should be reported as Medical Equipment. The category includes incidents where the wrong equipment was used, the equipment failed, the labeling, packaging, or user manual was misleading, or contained faults.

Patient Behavior

Adverse events correctly classified as Patient Behavior are events where a patient has caused an unwanted situation affecting the patient themself or another patient. Examples of incidents that should be classified as Patient Behavior are self-injury, suicide, self-medication, escaping or disappearing, violence, threats, and aggression. In the last example, the incident should be reported as two separate incidents, one for the perpetrator and one for the victim.

Blood, Cells, Tissue, and Organs

This category is used for incidents related to the collection, processing, and use of blood, cells, and organs. It can be both related to the donating patient or the receiving patient. It can be related to the choice of donor, side effects in the donor, the storage of the product, the order of the product, the transplantation, side effects in the receiving patient, or lack of traceability. At St. Olav's Hospital, this type of event is referred to as Blood and Blood Products.

NOKUP	**St. Olav's Hospital**	**Comment**
Patient Administration	Patient Administration and Coordination	
Diagnostics	Diagnostics, Treatment, and Care	St. Olav's Hospital have merged Diagnostics with Treatment and Care and removed Laboratory services from Diagnostics to a separate category.
	Laboratory Services	Is part of Diagnostics in NOKUP
Treatment and Care		Is part of Diagnostics, Treatment, and Care at St. Olav's Hospital
Falls and Accidents	Falls and Accidents	
Documentation/ Information	Documentation and Information	
Infection	Infection	
Drugs	Drugs	
Medical/Technical Equipment	Medical Equipment	
Patient Behavior	Patient Behavior	
Blood, Cells, Tissue and Organs	Blood and Blood Products	

Table 2.1: Mapping between the NOKUP categories to the categories utilized at St. Olav's Hospital

2.3 Text Classification

Text classification is the process of sorting text documents into one or more predefined categories [Basu et al., 2003]. It is a supervised machine learning technique where a classifier is trained on a labeled dataset of text documents and then used to predict the category, or class, of new, unseen text documents. The goal is to develop models that can accurately predict the category of a given text document, based on its content.

Text classification usually includes five steps: data collection, text preprocessing, feature extraction and selection, model training, and model evaluation. A diagram of this process is illustrated in Figure 2.1

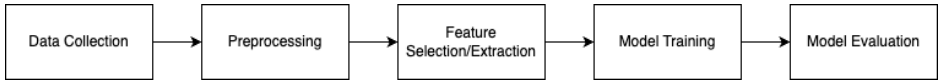


Figure 2.1: Text classification steps

One of the main challenges of text classification is the high dimensionality of the text data [Khan et al., 2010]. Text data typically consists of a large number of words, each treated as a distinct feature and thereby contributing to the overall dimensionality. Furthermore, since each word has the potential to impact the classification of the document it is crucial to analyze and extract the most relevant words. Discriminatory features, which capture the distinctive characteristics or patterns in the text data that differentiate between different classes or categories, play a vital role in text classification. These features could be specific words or word frequencies among others. To address the challenge of high dimensionality and incorporate discriminatory features, various text preprocessing techniques such as removing stop words and stemming, as well as feature selection and extraction methods, have been developed.

This section will give an overview of the text classification steps, including the specific methods chosen for the project.

2.3.1 Preprocessing

Preprocessing is the process of cleaning and transforming the dataset to prepare it for text classification [Haddi et al., 2013]. Effective preprocessing techniques are essential in achieving accurate results in text classification tasks. Cleaning and transforming the data reduces noise and standardizes the text documents, making it easier for the machine learning algorithms to extract relevant features and classify them into pre-defined categories [Khan et al., 2010]. Preprocessing techniques can include a variety of steps, such as tokenization, stop word removal,

and stemming. Each of these steps is crucial in ensuring that the model is fed high-quality data, which in turn helps to improve the accuracy and efficiency of the model.

It has been shown that preprocessing can significantly improve classification accuracy and dimension reduction [Khan et al., 2010]. However, there is not any combination of preprocessing techniques proven to be successful for all use cases and datasets [Uysal and Gunal, 2014]. Therefore, it is important to experiment with different techniques to determine what works best for each specific classification task. This section will define and discuss the preprocessing techniques utilized in the project.

Tokenization

Tokenization is the process of splitting text into meaningful distinct elements called tokens [Kowsari et al., 2019]. A token can be a character, a word, or a sentence. Going forward, each token will represent one word. Tokenization is a crucial part of preprocessing as it transforms the raw data into a format that can be further processed by a text classification algorithm [Ahmed et al., 2022]. An example of the tokenization technique utilized is illustrated below:

"Patient fell while unsupervised." $\rightarrow$ ["Patient", "fell", "while", "unsupervised"]

After applying this technique, each word in the sentence is converted into a token and added to a list.

Removing stop words

Stop words are frequently used words without semantic meaning that serve a grammatical function in natural language [Anandarajan et al., 2019; Ahmed et al., 2022]. Examples of stop words include prepositions such as "of", "on", and "to", pronouns such as "I" and "them", and articles such as "the" and "a". Removing stop words can help reduce noise and word dimensionality, potentially leading to improved accuracy and efficiency for text classification [Silva and Ribeiro, 2003]. Appendix A includes all the Norwegian stop words used for the project.

Stemming

Stemming removes all prefixes and suffixes to obtain the base of a word [Kowsari et al., 2019]. For example, stemming would convert "falling" and "fell" into the stem word "fall". Similarly to removing stop words, stemming may help further reduce the noise and word dimensionality [Anandarajan et al., 2019].

Stemming can have a negative impact on classification performance depending on the use case and dataset since it may not always accurately capture the intended meaning of words, and multiple forms of the same word could contribute positively to the classification outcome. [Yu, 2008]. Another issue with stemming may occur if words with different meanings have the same stem [Anandarajan et al., 2019].

2.3.2 Balancing the dataset

A dataset is imbalanced when the classification categories are unevenly represented, resulting in certain classes being overrepresented while others are underrepresented [Chawla, 2010]. An imbalanced dataset can cause problems for text classification models, as they may be biased towards the majority class and have difficulty accurately predicting the minority class. There are several methods to deal with imbalanced datasets, such as undersampling, oversampling, or a weighted loss function [More, 2016]. This section will describe each of the balancing methods tested for the project.

Random Undersampling

In Random Undersampling, the majority class is reduced by randomly removing instances from that class until a more balanced distribution is achieved between the majority and minority classes [More, 2016]. This involves randomly selecting a subset of instances from the majority class to match the number of instances in the minority class.

The main goal of Random Undersampling is to reduce the dominance of the majority class, allowing the model to pay more attention to the minority class during training. By doing so, it helps prevent the model from being biased towards the majority class and improves its ability to learn from the minority class, which may be more challenging to identify due to its limited representation. In addition, it reduces the training size and therefore the run time of the classification models [Hernandez et al., 2013]

While Random Undersampling can help address class imbalance issues, it also comes with potential drawbacks. Removing instances from the majority class reduces the amount of training data available, which can lead to information loss and potential underfitting [Mohammed et al., 2020]. Additionally, randomly selecting instances for removal may discard informative instances, potentially affecting the model's overall performance.

Random Oversampling

Random Oversampling involves randomly selecting instances from the minority class and duplicating them until a more balanced distribution is achieved between the minority and majority classes [More, 2016]. This approach increases the number of instances in the minority class by creating copies of existing instances. The duplication process is typically performed randomly and independently for each selected instance, resulting in multiple identical or highly similar instances of the minority class.

The main purpose of Random Oversampling is to mitigate class imbalance by providing more training examples for the minority class [Hernandez et al., 2013]. By increasing the representation of the minority class, the model has a better opportunity to learn its distinguishing characteristics and make accurate predictions. This can be particularly useful when the minority class is underrepresented, and its instances are scarce in the original dataset.

Random Oversampling is a relatively simple and straightforward technique to implement. However, it should be used with caution, as it can potentially introduce overfitting and duplicate noisy or irrelevant instances [Mohammed et al., 2020].

Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a popular algorithm used for oversampling to address the class imbalance. Unlike Random Oversampling which duplicates existing minority class instances, SMOTE creates synthetic instances by imploring between neighboring instances from the minority class [Chawla, 2010].

By generating synthetic instances based on the feature relationships of the minority class, SMOTE effectively expands the minority class, providing additional training examples for the model to learn from. This helps in improving the model's ability to generalize and make accurate predictions for the minority class.

Balancing the dataset with class weights

Class weights are a balancing technique where each category receives a weight based on its size [More, 2016]. Equation 2.1 displays how to calculate the weight for category C_y, where N represents the total number of documents, C is the number of categories and N_{c_y} is the number of documents belonging to category C_y.

$$w_y = \frac{N}{C * N_{C_y}} \tag{2.1}$$

Class weights can be used to balance the contribution of each class during the training of a model. They are typically applied in the loss function of the model, where the weights are multiplied by the loss for each instance of that class [More, 2016]. This means that instances of the minority class will have a higher weight, contributing more to the overall loss and gradient update during training.

2.3.3 Feature Selection and Extraction

Feature selection and feature extraction enable the classification algorithm to identify the most relevant and informative aspects of the text that are needed to make accurate classifications [Alsmadi and Gan, 2019; Shah and Patel, 2016]. Feature selection involves choosing a subset of features most relevant to the classification task. This is important as not all features may be equally informative or useful for classification, and using too many irrelevant features can lead to overfitting or poor performance [Deng et al., 2019]. Feature extraction involves identifying and extracting relevant features from the text and transforming them into a numerical representation that can be used by the classification algorithm [Alsmadi and Gan, 2019]. This allows the feature extraction method to also serve as a document representation method.

The feature selection and extraction methods employed are based on the availability of methods in the Sci-kit learn library. This section will introduce and define these methods, including three feature selection methods, Mutual Information (MI), Chi-squared (CHI2), and Analysis of Variance (ANOVA) F-value, and two feature extraction methods, Bag-of-words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF). In addition, N-gram will be presented as it is used in combination with the feature extraction methods to potentially capture additional relevant features. In the description of the methods, the terms "term" and "word" will be used interchangeably to encompass the features that will be assessed for selection and extraction.

Mutual Information (MI)

MI ranks the mutual dependency between term t and category c [Deng et al., 2019]. If term t and category c have the probabilities $p(t)$ and $p(c)$, their MI is defined as:

$$MI(t, c) = \log \frac{p(t, c)}{p(t)p(c)} \tag{2.2}$$

If the MI has the value of natural zero, term t and category c are independent, indicating that the presence or absence of the term does not provide any discriminatory information about the category [El Mrabti et al., 2018].

MI is frequently used in information theory, however, it often performs poorly in text classification [Deng et al., 2019]. This can be contributed due to its sensitivity to marginal probabilities of terms and tendency to prioritize rare terms [Yang and Pedersen, 1997].

Chi-squared (CHI2)

CHI2 represents the dependence between a term t and category c [Yang and Pedersen, 1997]. The equation for the CHI2-statistic for term t and category c can be found in Equation 2.3

$$CHI2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+C)} \tag{2.3}$$

where A represents the documents of category c that contain term t, B represents the documents of other categories that contain term t, C represents the documents of category c that do not contain term t, and D represents documents that do not belong to category c or contain term t. N is the total number of documents in the dataset. If t and c are independent, the score will be 0.

One drawback with CHI2 is that it is known to be unreliable for low-frequency terms [Deng et al., 2019; Yang and Pedersen, 1997].

Analysis of Variance (ANOVA) F-value

ANOVA measures the ratio between the variance of a term between categories and within the categories [El Mrabti et al., 2018]. The equations to compute the F-value are found below where N is the number of documents, S is the number of categories, j_i is the number of terms in category j, $\bar{K}_i$ is the mean of category i, $\bar{K}$ is the overall mean of the data, and K_{ip} is pth term in category i [Pathan et al., 2022]

$$variance\ between\ categories = \frac{\sum_{i=1}^{j} j_i(\bar{K}_i - \bar{K})^2}{(S-1)} \tag{2.4}$$

$$variance\ within\ categories = \frac{\sum_{i=1}^{S}\sum_{p=1}^{j_i}(K_{ip} - \bar{K}_i)^2}{(N-S)} \tag{2.5}$$

$$ANOVA\ F\text{–}value = \frac{variance\ between\ categories}{variance\ within\ categories} \tag{2.6}$$

N-gram

The n-gram method combines n terms in their correct order into a single token [Kowsari et al., 2019]. 1-gram, or unigram, is when each word is transformed into a token by itself. N-grams with n larger than 1 are used to possibly detect or preserve information from the semantic order of terms in a document. An example of a 2-gram, also called a bigram, is:

"Patient suffered allergic reaction to prescribed medication." → ["Patient suffered", "suffered allergic", "allergic reaction", "reaction to", "to prescribed", "prescribed medication"]

N-gram can be used as both a document representation and feature extraction method. In the Sci-kit learn Python library, N-gram functionality seamlessly integrates with feature extraction methods, providing a convenient and straightforward approach for incorporating N-gram features into text analysis tasks [Pedregosa et al., 2011].

Bag-of-words (BOW)

BOW is a frequently used feature extraction method where each document is represented as a bag containing its terms [Deng et al., 2019]. The bag considers the term multiplicity while disregarding grammar or the order of terms. This is a potential drawback if the semantics between the words are important for the classification [Ahmed et al., 2022]. This approach creates a vocabulary with every unique term and represents each document by a vector of word counts with the dimension of the vocabulary.

Table 2.2 gives an example of the BOW-method with the two sentences "Patient fell and sustained a head injury" and "The patient was treated for head injury".

Word	Sentence 1	Sentence 2
patient	1	1
fell	1	0
and	1	0
sustained	1	0
a	1	0
head	1	1
injury	1	1
the	0	1
was	0	1
treated	0	1
for	0	1

Table 2.2: Bag of Words representation of two sentences with lowercase conversion

Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a numerical statistic that aims to reflect the importance of a term in a document relative to its total occurrence across all documents. It considers both the Term Frequency (TF) and the Inverse Document Frequency (IDF) when measuring term importance [Ahmed et al., 2022]. TF assigns a high value to the terms with more frequency in a document and a low value to the less frequent terms [Rani et al., 2022]. The equation for TF is given in Equation 2.7. IDF, on the other hand, considers how often a term is used across all documents by calculating the logarithm of the ratio between all documents and the documents with a specific term included [Rani et al., 2022]. The equation of IDF can be found in Equation 2.8. The basic idea is that a term is more important to a document if it occurs frequently in the document, but less important if it occurs frequently in many other documents. This is found by multiplying the results of the TF and IDF of a term, as seen in Equation 2.9.

$$TF = \frac{Number\ of\ times\ term\ t\ occurs\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d} \tag{2.7}$$

$$IDF = log\left(\frac{Total\ number\ of\ documents\ in\ the\ dataset}{Number\ of\ documents\ with\ term\ t\ included}\right) \tag{2.8}$$

$$TF\text{–}IDF = TF * IDF \tag{2.9}$$

TF-IDF is commonly used as a feature extraction technique in text classification

[Kowsari et al., 2019], where the TF-IDF score for each term is calculated in every document in the dataset and used as a feature for the classification model.

2.3.4 Classification methods

One of the most critical steps in text classification is selecting the best classifier for the use case. Different datasets and use cases will work best with different classification algorithms. Some classifiers may be better suited for handling large amounts of data, while others may be more effective at handling noisy or unstructured text data. The selection process, therefore, requires an understanding of how each algorithm works and its strengths and weaknesses [Kowsari et al., 2019].

This section will introduce and define the two classification methods chosen for this project, NB and Multi-class SVM. NB is a popular text classifier and is often used as a baseline when comparing classification models [Rennie et al., 2003; Kowsari et al., 2019]. The method is also easily implemented and computationally efficient. On the other hand, several studies have found that SVM outperforms other models in text classification tasks [Deng et al., 2019]. However, it is computationally expensive.

Näive Bayes (NB)

For text classification, the NB algorithm has two frequently used variations: Multivariate Bernoulli NB and Multinomal Näive Bayes (MNB) [Singh et al., 2019; McCallum et al., 1998]. To address the issue of imbalanced datasets, a variant of MNB called Complement Näive Bayes (CNB) was adopted. This section will present the basic algorithm of NB and then define and discuss the three NB variations for text classification.

NB is a probabilistic classification algorithm that makes predictions based on Bayes' theorem, which describes the probability of an event based on prior knowledge of conditions that might be related to the event [Kowsari et al., 2019]. The classifier makes the assumption that the features are independent of each other given the class label. This also makes the feature order irrelevant and the presence of one feature doesn't affect the presence or absence of another [Khan et al., 2010]. The definition of NB algorithm for document d and class c can be found in Equation 2.10.

$$P(c|d) = \frac{P(c|d)P(c)}{p(d)} \tag{2.10}$$

Multivariate Bernoulli NB, henceforth called Bernoulli NB, uses independent binary variables as features to indicate the presence or absence of a term in a

document. [Singh et al., 2019]. Since the method works on a binary, it does not consider a term's frequency. The formula for the probability of class c given document d according to Bernoulli NB is found in Equation 2.11, where $x_{t,d}$ is the occurrence of term t in document d, and $\bar{d}$ is all the terms not in document d [Raschka, 2014].

$$P_{bNB}(c|d) = \frac{P(c)\prod_{t\in d} P(t|c)^{x_{t,d}} \prod_{t\in \bar{d}}(1 - P(t|c))^{(1-x_{t,d})}}{P(d)} \tag{2.11}$$

Unlike the Bernoulli NB, MNB captures the term frequency information in the documents [McCallum et al., 1998]. McCallum et al. [1998] also found that MNB uniformly did better than the Bernoulli NB approach. However, it still makes the assumption that the feature's context and position are independent of class. The probability of class c given document d in MNB can be found in Equation 2.12 where $n_{t,d}$ is the frequency of term t in document d. Studies have shown that MNB often does better than other versions of NB for text classification in real-life applications [Deng et al., 2019].

$$P_{mNB}(c|d) = \frac{P(c)\prod_{t\in d} P(t|c)^{n_{t,d}}}{P(d)} \tag{2.12}$$

NB usually performs poorly on imbalanced datasets Kowsari et al. [2019]. CNB is a derivative of MNB created to deal with this issue. Instead of using training data from a single class c as seen with MNB in Equation 2.12, CNB estimates using data from all classes except c [Rennie et al., 2003]. This can create less bias in the weight estimates by using a more even amount of training data per class. The probability of class c with this method is found in Equation 2.13, where $P(t|\bar{c})$ is the conditional probability of term t given all the other classes except c.

$$P_{cNB}(c|d) = \frac{P(c)\prod_{t\in d}(1 - P(t|\bar{c}))^{n_{t,d}}}{P(d)} \tag{2.13}$$

Whereas the strong independence assumption makes NB a highly efficient classification method, it can create limitations in its applications. Conditional independence is often not the case for real-world data and correlations between features are lost. However, it performs surprisingly well in complex real-life situations but is still often outperformed by other discriminatory classification models such as the SVM [Khan et al., 2010].

Multi-class SVM

SVM is a supervised machine learning algorithm commonly used for classification tasks [Khan et al., 2010]. The key concept of SVMs is to find the optimal decision boundary, or hyperplane, that separates different classes of data points in high-dimensional space. The hyperplane maximizes the margin between the closest data point from each class in binary classification problems. [Gunn et al., 1998]. The margin is the distance between the hyperplane and the closest data points from each class. SVMs are particularly useful when the number of features is large compared to the number of observations, or when there is a non-linear relationship between the predictors and the response variable. Non-linear SVMs can be implemented using a kernel function, which maps the input feature vector into a higher-dimensional space where the data may be more separable [Vora and Yang, 2017]. Common kernel functions include the Radial Basis Function (RBF) and polynomial kernel.

As SVM was originally designed for binary classification problems [Kowsari et al., 2019], several strategies to create a multi-class SVM have been suggested, such as one-against-one, one-against-all, and Directed acyclic graph SVM (DAGSVM) [Hsu and Lin, 2002]. One-against-one and one-against-rest are the most frequent methods for mutliclass-SVM, and between these methods, Hsu and Lin [2002] found that one-against-one were more successful in practical cases. Therefore only the one-against-one method will be discussed and implemented going forward.

In 1990, Knerr et al. [1990] introduced the one-against-one algorithm for multiclass classification. This approach involved creating $k(k-1)/2$ binary SVM-classifiers, where k is the number of categories in the dataset [Zhang, 2012]. Each binary classifier is trained to distinguish only between two categories, i and j, with i considered as the positive class and j as the negative class. The training sample used for the binary-SVM consists of the documents that only belong to the two categories being distinguished. The one-against-one algorithm optimizes each classifier to solve the following binary classification problem [Hsu and Lin, 2002]:

$$\begin{aligned} \min_{w^{ij},b^{ij},\xi^{ij}} & \frac{1}{2}(w^{ij})^T w^{ij} + C\sum_t \xi_t^{ij}(w^{ij})^T \\ & (w^{ij})^T\phi(x_t) + b^{ij} \geq 1 - \xi_t^{ij}, \text{ if } y_t = i \\ & (w^{ij})^T\phi(x_t) + b^{ij} \leq -1 + \xi_t^{ij}, \text{ if } y_t = j \end{aligned} \tag{2.14}$$

where w is the weight vector, b is the bias term, ξ is the slack variable allowing for some misclassification, and (x_i, y_i) represents training data points.

During testing, the $k(k-1)/2$ binary-SVM classifiers that were constructed are

used for classification. The final classification decision is made using a voting strategy, where each binary classifier "votes" for the class it has predicted, and the class with the highest number of votes is selected as the classification result [Hsu and Lin, 2002].

Potential downsides with SVM is the memory complexity and lack of transparency in the results because of the high dimensionality. Depending on the chosen kernel, SVM might also be prone to over- or underfitting [Kowsari et al., 2019].

2.3.5 Evaluation metrics

To assess the effectiveness of a text categorization model, it is critical to employ evaluation metrics and methods to understand and optimize its performance [Hossin and Sulaiman, 2015]. This section will define the metrics and methods used to evaluate the classification models in the project.

Confusion Matrix

A confusion matrix is a valuable tool used to evaluate the performance of a classification model. It is a scalable table that presents the true and predicted classifications of a model. Based on the results of the matrix, various performance metrics such as accuracy, precision, recall, and F-score can be calculated [Hossin and Sulaiman, 2015]. Table 2.3 presents a confusion matrix for a binary classification problem [Vujović et al., 2021].

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Table 2.3: A confusion matrix for a binary classification problem

Table 2.4 shows an example of a multiclass confusion matrix with three categories. The cells in the matrix show the number of instances that were classified as each combination of the actual and predicted classes. For instance, there were ten instances that were actually class A and were correctly classified as A, while there were two instances that were actually class A but were misclassified as B.

Accuracy

Accuracy is a simple, intuitive evaluation metric that measures the percentage a model correctly predicts the category or label of a document [Hossin and Su-

Actual/Predicted	**A**	**B**	**C**
A	10	8	1
B	3	8	4
C	2	1	11

Table 2.4: A confusion matrix for a multiclass classification problem

laiman, 2015]. The definition of accuracy is the ratio between the number of correct predictions and the number of total predictions illustrated in Equation 2.15.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.15}$$

When working with unbalanced datasets, accuracy has its limitations. Accuracy does not distinguish between the types of errors it makes [Sokolova et al., 2006]. This can be illustrated with a dataset that has 90% of class A and 10% of class B will still achieve an accuracy of 90% if it only ever predicts the majority class.

Precision

Precision is an evaluation metric that measures the proportion of predicted positive instances that were actually positive [Gu et al., 2009]

$$Precision = \frac{TP}{TP + FP} \tag{2.16}$$

Recall

Recall measures the proportion of actual positive instances that were correctly identified by the model [Gu et al., 2009]. The formula for calculating the recall is found in Equation 2.17 where TP is the instances that were correctly predicted as positive and FN are the instances that were actually positive, but incorrectly classified as negative by the model.

$$Recall = \frac{TP}{TP + FN} \tag{2.17}$$

F-measure

F-measure is a popular evaluation metric when faced with an imbalanced dataset [Gu et al., 2009]. It combines the results of precision and recall depending on the factor β. The formula for calculating the F-measure is given in Equation 2.18

$$F\text{–}measure = \frac{(1+\beta) * Precision * Recall}{\beta * Precision + Recall} \tag{2.18}$$

This project uses $\beta = 1$, which creates a harmonic mean between recall and precision as shown in Equation 2.19 [Gu et al., 2009].

$$F1\text{–}measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.19}$$

In multiclass classification, F1-measure can be calculated using one of three methods: micro, macro, and weighted [Pedregosa et al., 2011]. The micro method computes the F1-measure by counting the total number of TP, FN, and FP. The macro method calculates the F1-score for each category and then takes their average. The weighted method also calculates the F1-measure for each category but adjusts for any class imbalance by finding the weighted average. This means that the F1-score of the most populated category will contribute more to the overall F1-measure. As this project considers all categories equal, the classification models will from now on be optimized using the macro method.

Chapter 3

Related Work

This section explores the prior research conducted on adverse event reporting and analysis before specifically focusing on the application of machine learning techniques in the domain. It also discusses the research on machine learning with Norwegian clinical text. At the beginning of the project, it became evident that there was limited research within the field of machine learning applications within the adverse event process in Norway. Additionally, the utilization of Norwegian clinical text for machine learning purposes had received little attention. Furthermore, the dataset utilized in this project has not previously been explored. This presents both an opportunity and a challenge, as the potential and limitations of the dataset remain unknown. By examining existing literature, including studies conducted in other countries, this section aims to provide a comprehensive overview of the current state of research while identifying the opportunities that this thesis aims to address.

Adverse event reporting plays a crucial role in enhancing patient safety and improving the quality of healthcare. Rafter et al. [2015] highlights the need for a comprehensive and systematic approach to learning from adverse events in healthcare. The study emphasizes the need for a safety culture that fosters a reporting environment, identifies and rectifies system failures, and implements standardized measurement strategies to monitor patient safety trends. To facilitate effective learning and analysis, it is essential for hospitals to not only cultivate a strong reporting culture but also employ an easy-to-use and consistent reporting system.

In line with these findings, Anderson et al. [2013] also found that incident reporting within hospitals should have a well-defined definition of the types of adverse events that should be reported. These studies underscore the necessity for consis-

tent and systematic categorization of adverse events to ensure an effective adverse event reporting and analysis process.

Furthermore, Waring [2004] discusses the variation in adverse event reporting within hospitals and explores the attitudes and participation of different medical professionals. Their study reveals significant differences in how adverse events were reported among various medical specialties, highlighting the need for standardization and consistency in reporting practices across all disciplines.

The Global Trigger Tool (GTT) is a method for retrospectively reviewing patient records using "triggers" to identify adverse events [Adler et al., 2008]. This manual process has been adopted by numerous countries, including Norway [Doupi et al., 2015]. However, due to its manual nature, several studies have been conducted to automate this process. One such study by Stockwell et al. [2013] presents an automated approach where a query is executed on patient records to identify relevant triggers, which are then subjected to manual review by clinicians. Another study by Dolci et al. [2020] developed a successful algorithm that specifically identified triggers associated with falls. Moreover, Doupi et al. [2015] discusses the widespread interest among the Nordic countries in GTT and their efforts towards automating this detection process. One of those efforts in Denmark is detailed in the study by Gerdes and Hardahl [2013]. These studies highlight the universal desire to automate processes related to adverse events; however, none of them have employed machine learning techniques thus far for this purpose.

In Norway, the field of automatic classification of adverse event reports is relatively unexplored. However, in other countries, there have been studies and research conducted on this topic. Ong et al. [2010] conducted a pilot study in Australia using NB and SVM classifiers to categorize clinical incidents based on the Healthcare Incident Types framework. The study achieved promising results, with NB achieving 86% accuracy for one subcategory and SVM demonstrating 98% accuracy for another. In addition, a comparison between reporter-classified and expert-classified incidents revealed potential misclassifications or lack of consistency for the reporter-classified adverse events.

In a similar project, Evans et al. [2020] focused on classifying primary care patient incident reports in the UK using free-text descriptions. SVM yielded the best results, with an AUROC of 0.839 and an F1-score of 0.607. This study demonstrated the potential for automatic classification of incident reports in the UK. Additionally, in the study conducted by McKnight [2012], a semi-supervised classification approach is employed, using both labeled and unlabeled adverse event reports to predict categories for the unlabeled reports. These studies emphasize the potential of various classification methods on clinical text to automate the

categorization of adverse events.

Askar and Züfle conducted two studies exploring how the application of topic modeling and clustering techniques in the analysis of adverse events. In the first study, the focus was on investigating the impact of COVID-19 vaccines across different states in the United States [Askar and Züfle, 2021a]. In the second study, Askar and Züfle [2021b] examined how the potential of these techniques to identify and compare adverse side effects of drugs across countries. Fujita et al. [2012], on the other hand, investigated the use of network analysis and Natural Language Processing (NLP) could be utilized in identifying effective categories for adverse events. These studies showcase diverse methods to gain new insights and enhance the analysis of adverse events.

Røst et al. [2018] focuses on the automatic detection of the use of central venous catheters from Norwegian clinical documentation for quality improvement and surveillance purposes. They experiment with different classification algorithms and feature selection methods. Despite having limited data, they were able to develop sentence classifiers that achieved reasonable results. This study demonstrates the potential of machine learning techniques in analyzing Norwegian clinical text and highlights different approaches that can be employed to address potential challenges.

Prior research underscores the importance of systematic and consistent adverse event reporting for establishing a robust and valuable reporting system in hospitals. Studies conducted in other countries have shown promising results in the field of automatic adverse event reporting. These studies have also served as inspiration for the selection of machine learning techniques employed in this thesis. Furthermore, the study by Røst et al. [2018] demonstrates the potential of utilizing machine learning and classification models specifically on Norwegian clinical texts. These collective findings highlight the significance of advancing automated adverse event reporting practices and exploring the application of machine learning techniques in a Norwegian healthcare setting.

Chapter 4

Methodology

This chapter presents the methodologies employed in the thesis, providing an overall understanding of the research process. It begins by providing an overview of the thesis timeline. Then the chapter focuses on the interview methods utilized to gather insights from the clinicians and experts. Subsequently, the experimental plan for the machine learning study, which focuses on the classification of adverse events into the NOKUP categories, is presented. Finally, the chapter provides a description of the tools utilized in conducting the machine learning study, accompanied by an overview of the resulting codebase.

4.1 Timeline Overview

This section presents an overview of the thesis and the timeline in which the different parts were executed. A visual representation is presented in Figure 4.1.

A part of this thesis has been conducted as a pre-project during the fall of 2022. The primary objectives during this phase were to acquire knowledge about the domain, establish the project's scope, and conduct an experiment on the dataset to become familiar with it and how it could be utilized. The dataset was received after the pre-project phase in mid-January. As a result, we were unable to perform the planned experiment on the dataset during the pre-project phase.

The pre-project phase was primarily spent conducting research interviews to acquire firsthand insights into the adverse event process within hospitals. These interviews served as a valuable source of knowledge and insights, allowing us to identify areas for potential improvement. Due to the unavailability of the dataset

at that stage, our primary focus was on comprehending the intricacies of the adverse event processes and defining various use cases where machine learning could play a role in enhancing the adverse event process.

Based on the three interviews conducted during the pre-project phase and the assumptions about the data we would get access to, the pre-project phase concluded with clustering of adverse events as the most feasible machine learning study to conduct. The experimental plan for this initial machine learning study and the related background theory can be found in Appendix B.

However, after the first extraction of the dataset in mid-January, it became apparent that another data extraction was needed as information mentioned by the clinicians was missing. The final dataset was extracted in February and prompted a shift in focus for the machine learning study to the automatic classification of adverse events. This shift of focus is further discussed in subsection 7.2.2.

After conducting the classification study, the results were evaluated in collaboration with clinicians to gain a better understanding of their implications.

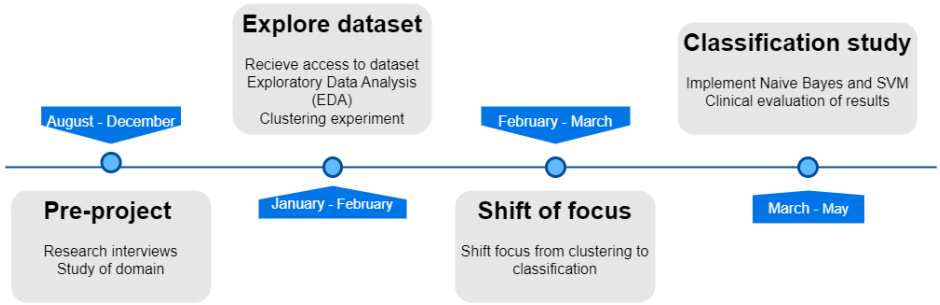


Figure 4.1: Overview of the main milestones and activities during the thesis

4.2 Research Interviews

To acquire insights from users and gather feedback, a series of research interviews were conducted. Interviews and focus groups are the most common techniques for data gathering in a healthcare context [Gill et al., 2008], with semi-structured interviews being the most frequent interview type. The purpose of a research interview is to explore and provide a deeper understanding of specific topics. Our aim was to obtain information about the adverse event reporting process and insights into what applications of machine learning were perceived as valuable by the clinicians and domain experts. A total of five interviews were conducted for

the thesis. This section will describe the interview participants involved in the project, the two interview types used, and the plan for each of the conducted interviews. Summaries of the interviews conducted can be found in Appendix C, while their contributions to the thesis are presented in Section 6.1.

4.2.1 Description of the Interview Participants

Four clinicians or domain experts and one focus group have been interviewed as part of the thesis. This section will describe their role in the adverse event reporting and analysis process.

Department of Corporate Governance (CG-1&2)

The Department of Corporate Governance (Virksomhetsstyring) is responsible, among other things, for ensuring quality management and patient safety at the hospital. As a part of this thesis, two representatives from the Department of Corporate Governance working with adverse events reporting have been interviewed. A part of the representatives' daily tasks consists of analyzing adverse events at the hospital and looking for trends and patterns in the adverse event reports that should be addressed. They are also responsible to handle the most severe adverse events, specifically those reported to the Norwegian Board of Health Supervision and The Norwegian Board of Health Supervision Investigation Commission. One of the employees interviewed is also a part of the group working on improving the NOKUP classification.

The two representatives will henceforth be referenced as **CG-1&2**.

Quality Advisor at St. Olav's Hospital (QA-1)

As mentioned in Section 2.2, hospitals are required to have quality and patient safety committees as a part of the systematic work to improve patient safety and quality of services. All clinics at St. Olav's Hospital have one or more quality advisors that are a part of the mentioned committees. The quality advisor interviewed in the thesis has previously worked with two different clinics. The quality advisor had first-hand knowledge of the NOKUP categories as he was involved in creating the guidelines.

Henceforth, the quality advisor at St. Olav's Hospital will be known as **QA-1**.

Quality Advisor at Helse Nord-Trøndelag (HNT) (QA-2)

Like St. Olav's Hospital, HNT has a team of quality advisors, who are responsible for various tasks, including the coordination of efforts across clinics to achieve

goals related to improving patient safety and quality of care. Each clinic within HNT has one or more quality advisors. The interviewed advisor works closely with other quality advisors to provide training and assistance with the analytic tool they use to extract information on adverse events that occur at their respective clinics. Additionally, the interviewed advisor analyzes adverse events reported across the entire HNT to gain a comprehensive understanding of the situation.

Henceforth, the quality advisor at HNT will be known as **QA-2**.

Clinicians at a Sepsis Seminar

A diverse group of clinicians, as well as some computer scientists, were gathered at a seminar discussing current research done in the field of sepsis in Norway. The seminar was arranged by the Gemini Center for Sepsis Research group. The clinicians attending represented different health professions, as well as several health institutions in the mid-Norway. They thus represented a diverse group of perspectives on the processes regarding adverse events in Norway.

4.2.2 Semi-structured Interview

Semi-structured interviews consist of several key questions to define which areas should be explored during the interview, but also allow to explore more into different areas or ideas brought up during the interview [Gill et al., 2008]. The flexibility also allows for the discovery or elaboration of information that is important to the participants but has not been thought of by the interviewers.

During the project, four semi-structured interviews were conducted with different users. Two interviews were with CG-1&2, one with QA-1, and one with QA-2. This section will outline the plan for each of the interviews.

Interview 1: Department of Corporate Governance

On the 28^{th} of September 2022, an interview was conducted with the two representatives from Corporate Governance at St. Olav's Hospital, CG-1&2, to gain insights into their current tasks and workflow and to identify potential areas for improvement. The interview lasted for one hour. The primary objective of this interview was to obtain a deeper understanding of the adverse event reporting process at St. Olav's Hospital, which included an examination of the tools employed, the personnel involved, and the procedures used to process an adverse event.

The questions used during the interview:

- **Q1:** Can you provide an overview of the adverse event reporting process at St. Olav's Hospital? What are the key steps involved?
- **Q2:** What tools or systems are utilized for reporting adverse events at the hospital? How are these tools used in the reporting process?
- **Q3:** Can you describe the procedures followed when processing an adverse event report? How is the information analyzed and reviewed?

Interview 2 & 3: QA-1 and QA-2

On the 10^{th} of November 2022, an interview was conducted with QA-1, followed by an interview with QA-2 on the 6^{th} of December 2022. Both interviews lasted for one hour. The primary objective of both interviews was to gain additional insights into the adverse event process from new perspectives and uncover any discrepancies in the workflow between the two healthcare institutions. Furthermore, we aimed to identify areas where the quality advisors perceived issues in the adverse event reporting process and explore the potential applications of machine learning in these areas.

The following questions were defined in the preparation for the interviews:

> **Q1:** Can you provide an overview of your role as a quality advisor in the adverse event reporting process at your healthcare institution?
>
> **Q2:** Are there any specific tools or systems that you utilize in the adverse event reporting process? How do you use these tools?
>
> **Q3:** What are the challenges or issues that you observe in the current adverse event reporting process?
>
> **Q4:** What are your thoughts on the potential applications of machine learning in the adverse event reporting process? Do you see any specific areas where machine learning could be beneficial?

Interview 4: Department of Corporate Governance

On the 1^{st} of March 2023, the second interview with CG-1&2 was conducted to discuss their opinions of the potential directions the thesis could take. The goal of the meeting was to discuss the potential value they saw in the different applications of machine learning we had identified based on the previous interviews, and based on their feedback decide which of the alternatives we should use for the machine learning study. For this meeting, it was no predefined questions specified. However, there was created an agenda for the meeting which included

a presentation and a short discussion of each potential application focusing on how they could contribute to the adverse event analysis.

4.2.3 Focus Group

On the 6^{th} of March 2023, we were given the opportunity to present the thesis and get feedback from clinicians from several institutions in Norway. During the presentation, we shared our current insights into the adverse event reporting processes and outlined three potential applications of machine learning that were identified through previous interviews. Additionally, we presented our proposal for the classification of adverse events using the NOKUP categories. The primary objective of this presentation was to gather valuable feedback on our initial findings and research direction. To facilitate active engagement, we structured parts of the presentation as open-ended questions to encourage audience participation and stimulate insightful discussions.

4.3 Experimental Plan for Classification

This section will describe the steps of the experimental plan for classifying the adverse events into the NOKUP categories, defined by the type of event, based on the titles and the description of the adverse event. To obtain the final models, some iterations excluded or refined steps of the experimental plan to potentially optimize the classification performance. This is further elaborated in step 8 in the experimental plan, and the results are outlined in Section 6.2.

1. **Data collection**. The dataset was extracted by Helse Midt-Norge IT (HEMIT). Since the dataset included health information about patients, ethical approval was needed before we could attain access to the data. The data was extracted multiple times, with new data being added each time. The final dataset had a total of 234 columns and 46,087 rows. As not all the columns were needed for the classification task, each needed to be analyzed to find the ones with relevant information. The relevant columns were then extracted for further analysis and preprocessing. More information about the ethical approval process and the dataset can be found in Chapter 5.

2. **Data cleaning and text preprocessing**. Data cleaning and text processing are crucial steps in text classification to ensure data consistency and improve classification performance. Rows with invalid values, such as *Null* values, were removed. The two free-text columns, Title of Report and Description of Event, were converted to lowercase. Special characters, such as punctuation or HTML tags, and numbers were removed from the adverse event title and description. Then the adverse event titles and descriptions

were combined to be one document for classification. Stop word removal and stemming were experimented with to find the optimal combination.

3. **EDA**. EDA is the process of examining and analyzing the data by using visual tools to reveal important information and characteristics about the data set and its variables [Hartwig and Dearing, 1979]. The goal was to better understand the data and identify patterns, trends, and relationships that may exist among the variables in a dataset. The results of the EDA are described in Section 5.3.

4. **Balance the dataset**. Given the highly unbalanced nature of the dataset, addressing this issue through balancing techniques is crucial. Various balancing techniques were employed, with the selection based on the best performance for each specific model. By balancing the dataset, the models can learn from representative samples of all classes, mitigating the impact of class imbalance and enabling more reliable and unbiased predictions.

5. **Feature selection and extraction**. To enable the classification models to process the textual data, the text data underwent feature extraction to convert it into numeric data. Following this, feature selection was performed to identify the 70% of the most relevant features for the classification task. This process aimed to increase the accuracy and effectiveness of the model by selecting only the most significant features.

6. **Classification**. The classification was performed with two classification methods, NB and SVM, to find the best-performing models for the dataset. Hold-out validation was performed to evaluate the performance, with 75% of the dataset used for training and 25% used for validation. The categories, used as the classification labels, were distributed equally between the training and validation set.

7. **Evaluation and visualization**. All the classification methods were optimized for the macro F1-measure to provide an overall view of the model's performance across the categories. Even though the models were optimized for the macro F1 metric, accuracy was still calculated to provide more information about the classification performance. A multiclass confusion matrix was used to compare predicted categories with the actual categories. This visualization allows us to further investigate which categories performed well and which ones were difficult for the model to predict. A table was created to visualize precision and recall for each category, with total accuracy and macro F1-measure included as well.

8. **Refinement and experiments for optimization**. Several experiments were conducted to optimize the performance of both models. In the first

experiment, as outlined in subsection 6.2.1, four different balancing techniques were tested for each classification technique. The aim was to identify the optimal solution for each technique by evaluating the performance of the models with each balancing technique. As there were three possible types of NB for text classification, each type was tested on the dataset with minimal text processing and with both feature extraction methods. The highest-performing method was then selected for further experiments, while the other two were discarded. The results of this can be found in Section 6.2.2. As there is no definitive approach for selecting the best combination of preprocessing steps, feature selection, and feature extraction methods, all possible combinations were tested to find the optimal approach for each model. This process would allow us to maximize the macro F1-measure by exploring different combinations of techniques. The final combinations were found in subsection 6.2.3

9. **Deployment of the best-performing models**. After identifying the best combination of preprocessing steps, feature selection, and feature extraction methods for each model, these final models were deployed for evaluation. The results of these models can be found in subsection 6.2.4.

10. **Clinical evaluation of results**. The evaluation and discussion of the results from the classification experiment were conducted in collaboration with CG-1&2 to assess the clinical relevance. The evaluation process began by gathering feedback on the clinicians' current experience with manual categorization. Specifically, we explored the categories that are frequently confused and those that are easily distinguishable. Next, we examined the results from the experiment, starting with the categories with the highest and lowest F1-scores to gain further insights. Finally, we analyzed the patterns in the errors made by the machine learning models and engaged in discussions surrounding these findings. This comprehensive evaluation allowed us to obtain valuable insights into the clinical implications of the results.

4.4 Tools

All experiments were conducted using the **Python** programming language in a **Jupyter Notebook** environment, employing various Python libraries. The **Pandas** library, which employs the data structure *Dataframe*, was used to extract and modify the database, including removing *Null* rows and gathering information for the EDA. Stemming was performed using the **Snowball** tool and a list of Norwegian stop words was extracted using the **Natural Language Toolkit** [Bird et al., 2009]. The **Scikit-learn** library [Pedregosa et al., 2011], designed for

machine learning and data mining, was used to preprocess the text, extract and select features, create and train the classification models, and calculate the evaluation metrics. To visualize the classification models' performance, the **Seaborn** library [Waskom et al., 2017] was employed, which is a data visualization library built upon **Matplotlib** [Hunter, 2007].

4.5 Codebase

The code for the machine learning study classifying the adverse event into the NOKUP categories and the EDA are publicly available on GitHub[1].

[1] https://github.com/RagnhildK/Adverse-Event-Classification

Chapter 5

Dataset

This chapter provides an overview of the environments in which the data has been processed, and the necessary authorizations and agreements obtained to access the data. The chapter also provides a presentation of the dataset, as well as the EDA. Due to the extensive nature of the dataset, only the most relevant columns will be presented in this chapter.

5.1 Prerequisites

As the dataset contains health information, both ethical approval and secure environments are needed to access the data. This section shortly presents the environments used to access and work with the data during this research, the ethical approval for the project, and the needed data agreements.

5.1.1 Environments

The data used in this research is stored in **HUNT Cloud**, which is a scientific computing environment owned by Norwegian University of Science and Technology (NTNU). HUNT Cloud provides researchers with tools to analyze and work with sensitive data. It ensures the privacy of data donors as well as provides a simple data exploration for researchers [NTNU]. All data-related activities, including management, viewing, extraction, processing, visualization, analysis, and statistical calculations, are exclusively performed within the secure HUNT Cloud environment. The raw data is securely contained within this environment at all times and is not transferred or shared outside of the environment. Connection to

HUNT Cloud is made through **OpenVPN** for Microsoft and **Tunnelbrick** for Mac OS, and access to the lab in HUNT Cloud is done using an SSH connection.

In HUNT Cloud one can request access to a workbench which provides smooth access to modern data science tools such as Jupyter Notebooks and Python. In this thesis, the workbench has been used for downloading Python packages, data preprocessing, EDA, data visualization, classification, and result analysis.

5.1.2 Ethical Project Approval

In accordance with Norwegian regulations, all research projects involving health information must receive approval from the Regional Committee for Medical and Health Research Statistics (REK) before commencing the project [The national research ethics committees, 2014]. This project received the necessary approval from REK as a part of the approval number 26814 (REK approval no. 26814; 2018/1201/REKmidt).

5.1.3 Data Agreements

Prior to accessing the data, it was necessary to sign a non-disclosure agreement. This agreement was implemented due to the presence of sensitive health information in the dataset. Additionally, a user agreement for HUNT Cloud, the digital project laboratory, had to be signed to obtain access.

Access to the HUNT Cloud lab and the dataset was granted on the 15^{th} of January 2023. The complete dataset used in the experiments was first available at the end of February.

5.2 Presentation of the Dataset

In this section, we present the dataset that serves as the foundation for this master's thesis. The dataset consists of adverse event reports that have been reported at St. Olav's Hospital from 2015 to 2022. The reports mostly contain clinical text written in Norwegian and contain abbreviations, medical terms, and incorrect sentence structure as is expected for clinical text [Dalianis and Dalianis, 2018]. The dataset contains minimal metadata explaining the different columns, and the explanations below are thus based on the knowledge gained through dialogue with clinicians.

The dataset provided comprises a total of 234 columns and 46,087 rows, providing a comprehensive collection of information for analysis. Given the scope of our project, we narrowed our focus to specific columns that were directly relevant to

our research objectives. The columns used in the experiments are Title of Report, Description of Event, and Type of Event. Some additional columns have been used in the EDA in order to better understand the dataset. For the purpose of readability, the titles of the columns used in the experiments are translated into English.

After removing reports containing *Null* values in any of the three columns used, the dataset contains a total of 3 columns and 42,825 rows.

5.2.1 Type of Event

An adverse event report should be associated with one, and only one, type of event, also referred to as category. However, in the dataset, there are two columns that both represent the category of the given adverse event. This is likely attributed to the data creation process, which follows the workflow outlined in Section 6.1.1. It is presumed that one column contains information provided by the event reporter, originally named Hendelsetyper, while the other column contains data provided by the reporter's manager, originally named 2. HENDELSESTYPE in the dataset. It is the latter we will use as the classification label for our machine learning models, as this is the one currently used at the hospitals to generate statistics and analysis.

During EDA we found that among the 46,087 rows, there are 3224 that have *Null* values in the column representing the type of event. These rows are not usable in our model as they lack a label, and were thereby removed from the dataset.

The entries in the column Type of Event have 13 different values as seen in Table 5.1. However, "ICT systems", "Other - Patient Accidents" and "No" is not supposed to be a category for patient-related adverse events as explained by CG-1&2, so these rows were deleted from our dataset. The remaining 10 categories are as described in 2.2.1 quite similar, but not equal to the NOKUP categories due to local adjustments in the system used for reporting at St. Olav's Hospital.

Type of Event	Count
Diagnostics, Treatment, and Care	10231
Patient Administration and Coordination	7782
Laboratory Services	6608
Falls and Patient Accidents	6291
Drugs	5348
Documentation and Information	3282
Patient Behaviour	1651
Medical Equipment	913
Infection	456
Blood and Blood Products	268
ICT Systems	3
Other - Patient Accidents	1
No	1

Table 5.1: The number of adverse events for each Type of Event

5.2.2 Title of Report

This column represents the title given to the report at registration. In the dataset, this column's original name is AvviksTittel. In interviews, CG-1&2 explained the titles to be quite descriptive. The average title consists of 31 characters and 4 words. Figure 5.1 and Figure 5.2 display the average title length and word count for each Type of Event.

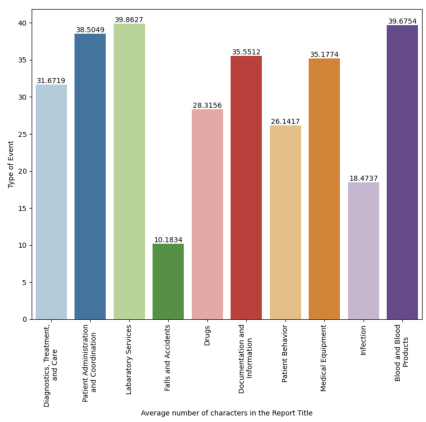


Figure 5.1: The average length of Title of Report for each Type of Event

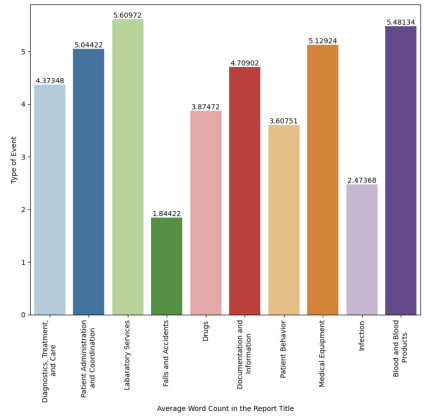


Figure 5.2: The average number of words in Title of Report for each Type of Event

Figure 5.2 shows that the number of words in the title of the reports varies between 1.8 and 5.6. The categories of Falls and Accidents, and Infections stand out with an average of 1.8 and 2.5 words in the title.

5.2.3 Description of Event

The column Description of Event, originally named Hendelsesbeskrivelse in the dataset, contains the reporter's free-text description of the event. The average description consists of 306 characters and 50 words. Figure 5.3 and Figure 5.4 display the average length and word count for the Description of Event for each Type of Event.

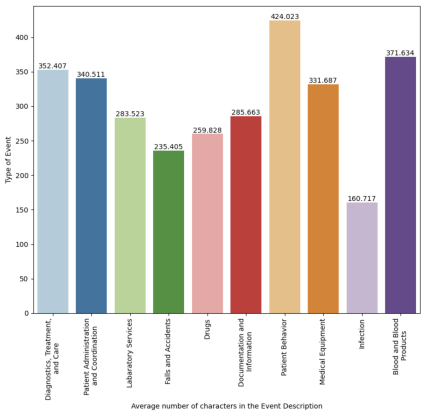


Figure 5.3: The average length of the Description of Event for each Type of Event

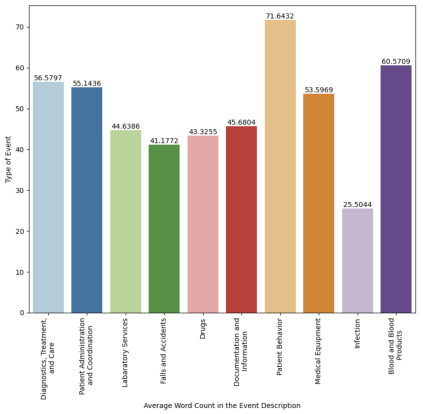


Figure 5.4: The average number of words in the Description of Event for each Type of Event

Figure 5.3 and Figure 5.4 show that the descriptions between the categories vary with an average of 160.7 and 402.0 characters and 25.5 and 71.6 words. In addition to having one of the smaller average title lengths, the category Infection also has the shortest average description length.

5.3 Exploratory Data Analysis (EDA)

This section presents the EDA conducted on the dataset. With this analysis, we aim to gain a deeper understanding of the dataset, uncover patterns and extract valuable insights.

5.3.1 Pre-analysis Statistics of the Dataset

Before delving into the detailed exploration of the dataset, it is essential to examine some pre-analysis statistics. These statistics provide an initial understanding of the dataset's characteristics, such as its size, composition, and basic properties.

The dataset consists of a vast collection of 234 columns, each representing different values in the reporting processes. However, for the specific focus of our research, we only concentrate on the three key columns previously presented in Section 5.2. Nevertheless, during the EDA, we extend our analysis to include the column Severity Level, original name Klassifisering av alvorlighetsgrad, to provide complementary insights into our dataset.

Table 5.2 presents an overview of the data types of the columns used in this analysis, along with the count of columns that do not contain *Null* values. In the dataset, all the data types are represented as the *object* data type, indicating that the columns consist of strings or a mixture of different data types.

Column	Non-Null Count	Datatype
Type of Event	42835	*object*
Description of Event	46080	*object*
Title of Report	46085	*object*
Severity Level	42775	*object*

Table 5.2: Column count and datatype

Prior to conducting further analysis on the dataset, we converted Title of Report and Description of Event into lowercase and removed the rows that contained *Null* values for columns Type of Event, Title of Report, or Description of Event. After these removals, we were left with 42,825 rows for analysis. This data-cleaning process helped to ensure the integrity and quality of the dataset for subsequent analyses.

5.3.2 Distributions of Adverse Events

The distribution of the adverse event types is visualized in Figure 5.5. It is evident from this figure that the dataset exhibits significant class imbalance, with the largest category comprising 10,228 adverse events and the smallest category containing only 268. This considerable class imbalance may pose challenges for the classification task, as it can impact the performance and accuracy of machine learning algorithms. Addressing this class imbalance will be crucial to ensure fair and reliable results in our research.

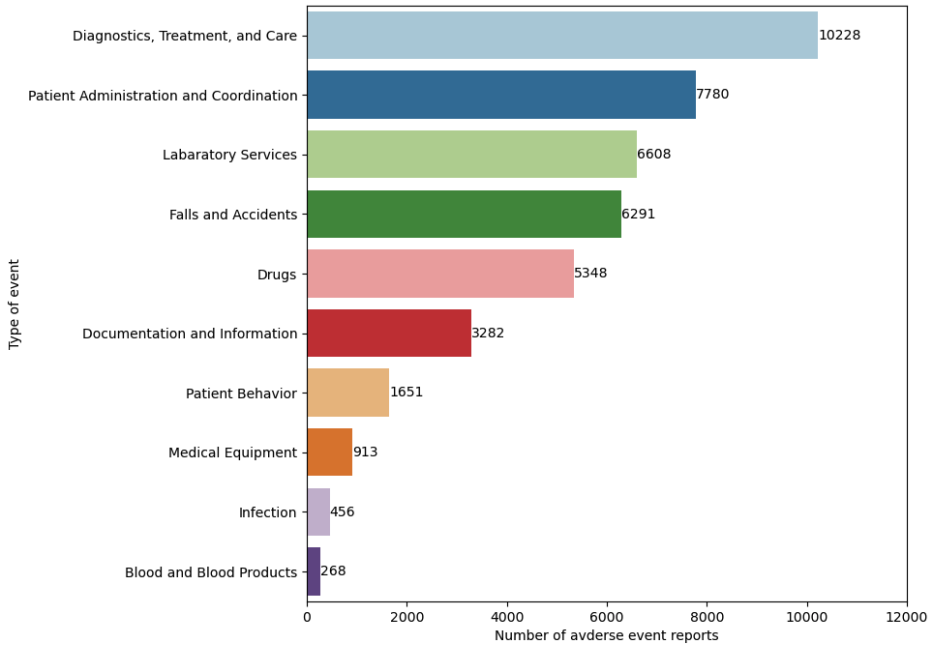


Figure 5.5: Distribution of adverse events for each Type of Event

The distribution of the severity of the adverse events is illustrated in Figure 5.6. Additionally, for a more detailed analysis, the distribution of severity levels for each type of event can be examined in Figure 5.7. These figures reveal that the most frequently reported adverse events at the hospital are those with no to minor consequences.

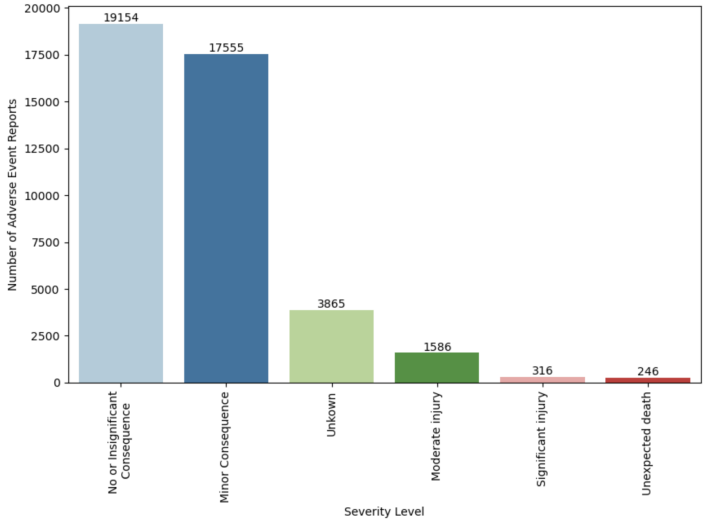


Figure 5.6: The distribution of the reported security of adverse events

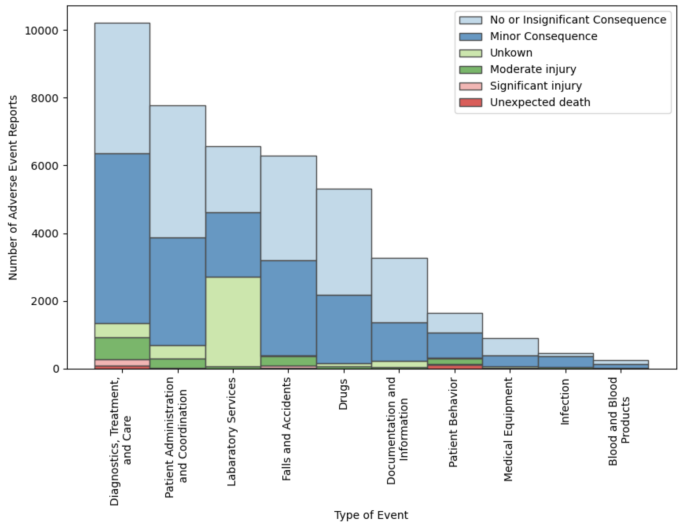


Figure 5.7: The distribution of the reported severity of adverse events per type of event

5.3.3 Exploration of the Adverse Event Descriptions and Titles

To gain a deeper understanding of the textual content associated with the reported adverse events, we further explored the Title of Report and Description of Event columns. The most recurring titles for the adverse event reports can be found in Table 5.3, while the most frequent descriptions are available in Table 5.4.

Title of Report	Count
fall	3889
feilmedisinering	339
brudd på samarbeidsavtale med underliggende retningslinjer	315
trykksår	254
flebitt	227
subcutan infusjon	102
pasientfall	98
medikamentavvik	88
avvik	84
pasientskade	83

Table 5.3: Top ten most recurring adverse event report titles

The most frequently recurring title, surpassing the second-place title by a factor of 10, is "fall" ("fall"). This observation suggests that incidents related to falls may be easily recognizable and distinguishable by the classification models, as the titles share similarities and are straightforward to identify. This assumption is further explored in subsection 7.4.2. In addition, the rest of the titles in Table 5.3 are concise and descriptive, indicating that they can serve as valuable supplementary information to the event descriptions, potentially enhancing the overall understanding and aiding in the classification process.

Furthermore, an interesting finding in the analysis is the presence of titles prefixed with "ESA" and "Elements" followed by a numeric code. These specific adverse events occurred in 1278 and 249 instances, respectively. These prefixes were also found in the adverse event descriptions, with a combined total of 683 instances. The inclusion of these titles in the dataset raises considerations regarding their potential impact on the classification task. They may introduce additional noise to the model or serve as valuable discriminatory features, warranting further investigation.

Table 5.4 reveals interesting patterns and findings. The most common description used by reporters is "see attachment" ("se vedlegg"), indicating the presence of additional information that is not available to the classification models as our

Description of Event	Count
se vedlegg	365
annet	77
2. manglende epikrise	71
3. mangler ved varsling	51
henvisning ikke vurdert av lege innen frist.	42
4. mangler ved medisinliste	28
manglende epikrise	26
mangler ved varsling	19
manglende/feil i legemiddeldokumentasjon	18
2. manglende epikrise¡br /4.mangler ved medisinliste	17

Table 5.4: Top ten most recurring adverse event descriptions

approach relies solely on the report title and event description. This finding highlights a potential limitation of our approach. The second most recurring description, "other" ("annet") is not descriptive and can potentially act as noise for the classification models.

Furthermore, it appears that some adverse event reports incorporate a list format within their descriptions. This can be observed through the presence of numbered sentences used either individually or in combination with one another, suggesting that they originate from the same list.

The effect of these findings on the results is discussed in subsection 7.4.4.

Chapter 6

Experiments and Results

This chapter presents the results obtained during the thesis and is divided based on the two research questions. Section 6.1 focuses on the results gained from the study of the domain and the research interviews conducted. It aims to answer the first research question of how machine learning can contribute to the process of adverse event reporting and analysis. Section 6.2 presents the experiments conducted in order to answer the second research question. The choice of which machine learning study to perform is based on the answer to research question 1.

6.1 Identifying Areas for Improvement in the Adverse Event Process

This section presents the results from the research interviews and the study of the domain. We have chosen to not include the interviews themselves in the thesis, but summaries of each interview can be found in Appendix C. The section starts by presenting today's process of adverse event reporting and analysis before presenting the identified areas where machine learning can contribute to increasing the learning outcome and making the adverse event process less resource-demanding.

6.1.1 The Process of Adverse Event Reporting and Analysis

The first step in answering **Research Question 1** is to understand today's process when handling adverse events. This is information that is not publicly

available, but has been acquired through interviews with healthcare professionals and internal documents from health institutions we have been given access to. The information is presented here as an important result of the research interviews and domain study, as it serves as a foundation for future work in this domain. An understanding of the workflow has been essential to achieving the needed domain knowledge to participate in discussions with domain experts on how machine learning could contribute to the learning outcomes of adverse event reports. It has also been crucial in order to understand the information contained in the dataset. A visual representation of the workflow can be found in Figure 6.1.

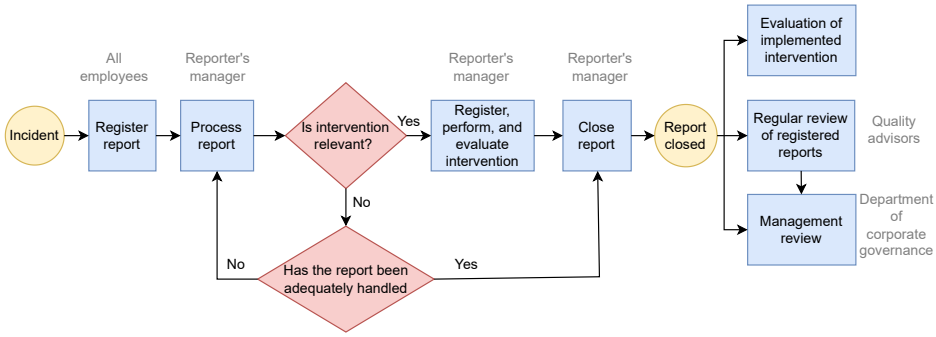


Figure 6.1: The workflow for managing adverse events

When a patient-related adverse event occurs at St. Olav's Hospital, it should be registered by the individual observing it. This individual will from now on be referred to as the **reporter**. The registration is done by filling out a form in the hospital's risk and quality management system, Extend Quality System (EQS). EQS is a web-based risk- and management system developed by the Norwegian software company Extend AS [Extend AS]. In the registration form, the reporter needs to add a title for the report, the time and place where the incident happened, the patient identifier of the patient involved, and a description of the event. Further, the reporter has to define the type of event using the categories presented in subsection 2.2.1 and estimate how often adverse events of a similar type happen. Due to the diverse clinical backgrounds of the reporters and the complexity of the NOKUP categories, this step is prone to misclassification and inconsistent categorization. The reporter also has the opportunity to write the immediate consequences and suggestions for actions to be taken. Lastly, there is also an option for adding attachments.

When registering an adverse event report, a responsible person is also assigned to evaluate the report. This is usually the reporter's immediate manager. Hence-

forth, we will refer to this individual as the **reporter's manager** or the **manager**. After the adverse event report is registered, the manager is notified and receives the adverse event report in EQS. The manager then has to evaluate the report and make a decision on whether the report concerns an adverse event that the hospital is obliged to report to external institutions or authorities, as described in Section 2.2. As a part of the evaluation, the manager should change the assigned NOKUP category when a misclassification has occurred. However, this happens rarely in practice. The manager is also responsible for determining whether preventive measures should be implemented to avoid similar events. If so, the measure is registered and a person is assigned responsibility for its implementation. If an evaluation of the implemented measure is deemed necessary, a person is designated responsible for conducting the evaluation. When the measure is implemented and evaluated the report can be closed. In general, the report should be closed within 30 days from registration.

At each clinic, there are quality advisors that are responsible to monitor the status of adverse events happening at the clinic. They convene regular meetings to discuss adverse events and strategize methods to prevent similar incidents in the future. These discussions go beyond examining individual adverse events, as they also involve identifying trends and comprehending the underlying causes. For instance, one interviewee highlighted an analysis of adverse events revealing a recurrent pattern of falls from chairs in the absence of supervision. Further examination and discussion at meetings revealed that these incidents predominantly occurred when nurses were simultaneously occupied with writing journal notes, which compromised patient safety. To address this issue, the hospital implemented a system where nurses take turns engaging in writing notes, ensuring that not all nurses are occupied with this task simultaneously. Additionally, they made the decision to refrain from purchasing chairs with slippery fabric, reducing the risk of falls. This serves as an example of how the statistics from the adverse event reports together with domain knowledge and experience are used to find the appropriate measures to prevent reoccurring adverse events.

While the quality advisors at each clinic are responsible for the monitoring of adverse events at each clinic, the representatives from the Department of Corporate Governance, described in Section 4.2.1, play a similar role, but at the hospital level. Their responsibility is to monitor the status of adverse events within the hospital and identify patterns that may span across various clinics and departments. By closely observing the reported incidents, they detect recurring themes or emerging trends, allowing them to raise awareness and initiate discussions among relevant stakeholders. During the interviews, it was revealed that they manually review all adverse event reports, with particular emphasis on the free-text fields, as they contain the most details. They especially pay close attention

to the reporter's description and the manager's assessment of the incident.

6.1.2 Potential Applications of Machine Learning

Based on the research interviews and the understanding of the current workflow as described above, three areas where machine learning can potentially enhance the adverse event process were identified. These potential applications of machine learning will be further evaluated and discussed in Section 7.2.

The identification of potential areas was done simultaneously with obtaining information about the current workflow, revealing different applications as more details were discovered. Initially, the interviews were focused on the analysis of adverse events, leading to the identification of clustering and summarization as potential applications for improvement. As the interviews progressed, and access to all the available data was obtained, the automatic categorization of adverse events emerged as another opportunity. The following results will provide a description of these applications, presented in the order in which they were identified.

Clustering of Adverse Events for Pattern Recognition

Through the research interviews, we learned that the process of recognizing patterns and analyzing adverse events demands a lot of resources across the hospitals. The adverse events are analyzed at several levels, first by the reporter's manager, then by the quality advisor, and lastly by the Department of Corporate Governance. QA-1 and CG-1&2 both confirmed in their interviews that they manually read the adverse events and creates diagrams to identify patterns and recognize trends. This manual approach can be challenging and resource-consuming due to the number of adverse event reports that need to be reviewed and the complexity involved in identifying meaningful patterns.

Clustering, as an unsupervised machine learning approach, enables the grouping of similar events based on their inherent characteristics, thereby facilitating the discovery of new ways to group, compare and analyze adverse events. The discovery of new patterns can provide valuable insights into emerging trends, potential systemic issues, or previously unrecognized relationships between adverse events. In the interview with QA-1, summarized in Appendix C.2, there was expressed skepticism regarding the predefined NOKUP categories, and an interest in exploring how a clustering model would group the adverse events compared with today's solution with the NOKUP categories.

Summarization of Adverse Event Reports

At St. Olav's Hospital, all adverse event reports are manually reviewed by representatives in the Department of Corporate Governance, however, not all healthcare institutions have the resources to read all adverse event reports as thoroughly. To address this challenge and enhance learning outcomes, machine learning algorithms can be employed to extract and present valuable insights from large volumes of reports, enabling more efficient analysis.

Machine learning, specifically NLP, can contribute to improving the analysis of adverse event reports by extracting and summarizing important information from the reports. By automatically identifying and condensing key details, machine learning can support decision-making processes and facilitate proactive monitoring of adverse events. This approach has the potential to improve patient safety and quality of care by enabling healthcare organizations to take appropriate actions in a timely manner.

By leveraging machine learning for adverse event report analysis, institutions with limited resources can benefit from enhanced proactive monitoring and improved patient safety. This technology-driven approach not only optimizes the utilization of available data but also supports the continuous efforts to enhance patient safety and quality of care across the healthcare industry.

Automatic Classification of Adverse Events into NOKUP Categories

Currently, adverse event reports are manually reviewed by clinicians who categorize them based on their understanding and expertise. This process can be both time-consuming and subjective, as different individuals may interpret and categorize events differently. The inconsistent utilization of categories leads to statistical inaccuracies, thereby distorting the representation of the distribution of adverse event types. Additionally, during the interviews, all participants highlighted the issue of misclassification in the manual categorization process. However, it is essential to note that the hospitals are required to use the NOKUP categories as it is a national system. This requirement further highlights the importance of ensuring consistent and accurate classification in the adverse event reporting process.

A possible solution to this challenge is to apply classification algorithms that could standardize the process of categorizing adverse events according to the national guidelines presented in NOKUP. The classification algorithms could be trained to analyze the content of adverse event reports and assign them to predefined categories. By employing such classification algorithms, the workload on human resources can be reduced while ensuring a standardized utilization of

categories.

Based on the feedback from CG-1&2 during the fourth research interview, we selected this application as the focus of the machine learning study. The study would then use the Type of Event column in the dataset as the classification label and examine the feasibility of automatic classification. This decision is further evaluated in subsection 7.2.2. The summary of this last interview can be found in Appendix C.4.

6.2 Classification of Adverse Events

This section presents an overview of the results obtained from the machine learning study, which aimed to assess the feasibility of classifying adverse events into predefined categories using their free-text titles and descriptions. To achieve an optimal performance of each of the classification techniques, a series of experiments were conducted to aid the decisions to create the final models for NB and SVM. The following subsections present each of these experiments and their outcomes. Initially, we started with the three variations of NB for text classification, described in Section 2.3.4, and one SVM model.

The first experiment focused on identifying the best balancing techniques for each of the initial models. In the next experiment, the variations of NB were compared to select the most effective variation for the subsequent experiments. The third experiment explored the combination of different preprocessing, feature extraction, and feature selection techniques to determine the optimal combination for each model's performance. Finally, the last experiment presented the results obtained using the final optimized model for NB and SVM.

6.2.1 Experiment 1: Selecting the Balancing Method

The objective of this experiment was to identify the most effective balancing method for each model to address the class imbalance in our dataset. Rather than determining the single overall best method, our aim was to select the most suitable balancing method for each classifier. To conduct this experiment, we evaluated various balancing methods including Random Undersampling, Random Oversampling, and SMOTE. For SVM we also evaluated the use of class weights in the loss function. However, this is not evaluated for the NB models, as NB assumes that the features are conditionally independent given the class, and does not directly incorporate class weights during training.

Experimental Setup

Each balancing method was applied individually to the dataset, and the performance of the classifiers was evaluated using the macro F1-measure. Furthermore, we present the classification performance without employing any balancing methods for the purpose of comparison. This allows us to assess the impact of the balancing techniques on the classifiers' performance and evaluate their effectiveness in addressing the class imbalance.

The balancing methods were tested using minimal preprocessing, which included special character and number removal, and lowercase conversion. Additionally, no feature selection methods were utilized to maintain the integrity of the experimental evaluation and to solely focus on the effectiveness of the balancing methods in addressing the class imbalance. Since feature extraction also served as the chosen document representation technique in this project, the balancing methods were tested for both BOW and TF-IDF to provide an overall view of the optimal method. The evaluation included Bernoulli NB, MNB, CNB, and SVM.

Experimental Results

The experimental results, as shown in Table 6.1 and Table 6.2, provide a comparison of balancing methods for the different classification techniques. The best score for each of the classifiers is highlighted in bold. Table 6.1 showcases the experimental results using the BOW method, while Table 6.2 presents the results utilizing the TF-IDF method.

Balancing Method	**BNB**	**MNB**	**CNB**	**SVM**
No Balancing	0.4259	0.5663	0.6557	0.6039
Random Oversampling	**0.6252**	0.6958	0.6277	**0.6896**
SMOTE	0.5690	**0.7057**	**0.6732**	0.6198
Random Undersampling	0.4727	0.5494	0.5587	0.5038
Class Weights	-	-	-	0.6704

Table 6.1: Comparison of balancing methods for the different classification techniques with BOW as feature extraction method

The results for Bernoulli NB remained consistent regardless of whether the BOW or TF-IDF method was employed. The results showed that all the tested balancing methods had a positive impact on the classifier's performance. Among these methods, Random Oversampling achieved the highest performance. Consequently, Random Oversampling was selected as the preferred balancing method for Bernoulli NB.

Balancing Method	**BNB**	**MNB**	**CNB**	**SVM**
No Balancing	0.4259	0.3967	**0.6422**	0.6782
Random Oversampling	**0.6252**	0.6805	0.6201	0.7005
SMOTE	0.5690	**0.6837**	0.6229	0.6903
Random Undersampling	0.4727	0.5568	0.5579	0.6180
Class Weights	-	-	-	**0.7090**

Table 6.2: Comparison of balancing methods for the different classification techniques with TF-IDF as feature extraction method

MNB achieved the highest performance when the SMOTE balancing method was utilized. When combined with the TF-IDF method, SMOTE resulted in a significant improvement in performance, with the F1-score doubling from 0.3967 without any balancing method to 0.6837 with SMOTE. This significant enhancement in performance led us to select SMOTE as the preferred balancing method for MNB.

CNB achieved the highest performance with different balancing techniques depending on which feature extraction method has been used. The best-performing balancing technique with the BOW method is SMOTE. However, the best performance with the TF-IDF method is achieved when no balancing method is utilized. This can be attributed to the fact that CNB is specifically designed for imbalanced datasets. Considering the overall best performance of CNB is achived with SMOTE and BOW, SMOTE is selected as the perfered balancing method for CNB.

SVM also performed best with different balancing methods when applied with the two feature extraction methods. Random Oversampling and class weights did the best for BOW and TF-IDF, respectively. As the overall best performance was with class weights and the significantly longer runtime associated with oversampling techniques, class weights were selected as the preferred balancing method for SVM.

Classification Model	**Selected Balancing Method**
Bernoulli NB	Random Oversampling
MNB	SMOTE
CNB	SMOTE
SVM	Class Weights

Table 6.3: The selected balancing methods for each of the classification models

This experiment served as a crucial step in addressing class imbalance and cre-

ating a solid foundation for the following experiments. The findings guided our selection of the most appropriate balancing method for each classification technique, contributing to improved model performance and reliable classification results. The decision made due to this experiment is summarized in Table 6.3.

6.2.2 Experiment 2: Selecting a NB Classification Type

In Experiment 2, our objective was to determine the most suitable variant of the NB classifier for our classification task. The best-performing method will be exclusively employed for the subsequent experiments.

Experimental Setup

To conduct the evaluation, the dataset was utilized without any preprocessing or feature selection techniques. This decision was made to focus solely on assessing the performance of the NB classification types, without introducing additional factors that could influence the results. However, it is important to note that the evaluation incorporated the best balancing method for each model, as determined in subsection 6.2.1. This inclusion ensured a more comprehensive evaluation by considering the impact of balancing techniques on the performance of the NB classifiers.

Like the previous experiment, the comparison was performed with both feature extraction methods.

Experimental Results

The experimental results are presented in Table 6.4, showcasing the classification performance of the NB classifiers with their respective optimal balancing methods. The table includes the performance achieved using both the BOW and TF-IDF feature extraction methods.

NB Type	BOW	TF-IDF
BNB	0.6252	0.6252
MNB	0.7057	0.6836
CNB	0.6732	0.6423

Table 6.4: Classification performance of the NB classifiers

Based on the results, the Bernoulli NB classifier achieved an F1-score of 0.6252 for both the BOW and TF-IDF feature extraction methods. The MNB classifier demonstrated improved performance with the BOW feature extraction method, achieving an F1-score of 0.7057, while obtaining an F1-score of 0.6836 with the

TF-IDF method. The CNB classifier yielded an F1-score of 0.6732 with the BOW feature extraction method and 0.6423 with the TF-IDF method.

Considering these findings, we can conclude that the MNB classifier demonstrated the highest performance across both feature extraction methods, with an F1-score of 0.7057 with BOW and 0.6836 with TF-IDF. Therefore, the MNB classifier will be selected as the NB classification type for the subsequent experiments.

6.2.3 Experiment 3: Determining Preprocessing, Feature Extraction, and Feature Selection Method

For this experiment, our objective was to determine the optimal combination of preprocessing techniques, feature extraction methods, and feature selection methods for our classification task for both MNB and SVM. This experiment aimed to identify the configuration that would yield the highest performance and provide insights into the impact of these factors on the overall classification performance.

Experimental Setup

This experiment was only conducted for the best-performing NB variant, MNB, as found in Section 6.2.2, and SVM. Before performing the evaluation, we employed the balancing method best suited for the classification techniques as found in subsection 6.2.1.

We evaluated the impact of two common preprocessing techniques: stop word removal and stemming. Additionally, two feature extraction methods, BOW and TF-IDF, were explored with both unigram and a combination of unigram and bigram to examine if word order and proximity added value to the classification task. Furthermore, three feature selection methods, namely MI, CHI2, and ANOVA F-value were considered to identify the most relevant features.

MI is not evaluated with TF-IDF because the MI implementation in the Sci-kit learn library requires discrete values, whereas TF-IDF provides continuous values. Converting the TF-IDF values to discrete values, such as binary or thresholded representations, would result in a substantial loss of information as the frequency and rarity of the features in the corpus would be discarded.

The evaluation was performed using the macro F1-measure, which provides an overall assessment of the models' performance. The combination of preprocessing techniques, feature extraction methods, and feature selection methods yielding the highest macro F1-score will be utilized for final experiments and analysis.

Experimental Results

The experimental results, presented in Appendix D, showcase the classification performance of our models based on different combinations of preprocessing, feature extraction, and feature selection techniques. The combination of techniques that yielded the highest F1-score is highlighted in bold and with a yellow background. The cells with a red background indicate the combinations with both MI and TF-IDF, which, as previously mentioned, will not be evaluated.

The experimental results demonstrated that the inclusion of stop word removal and stemming techniques had a negligible effect on the classification performance for both classification techniques. Surprisingly, in some cases, their incorporation even led to a decline in performance rather than improvement. These findings suggest that for the specific classification task at hand, the application of stop word removal and stemming did not provide significant benefits and may not be necessary preprocessing steps.

Furthermore, the results revealed that the choice of feature extraction and n-gram influenced the performance of the classification techniques differently. The best performance for MNB was consistently achieved using the BOW method with unigrams, while the worst performance was observed when using both unigram and bigram BOW. A similar pattern was observed for SVM, where the highest performance was achieved with unigram TF-IDF and the lowest with unigram and bigram TF-IDF.

In terms of feature selection, all methods generally provided only marginal enhancements in performance for MNB. The improvement was at most 1.5% compared to no feature selection, with most methods enhancing performance by less than 1%.

The results for SVM did not show a clear trend regarding the impact of feature selection methods on performance. Different feature selection methods, including no feature selection, performed best for different combinations. The highest performance was achieved without applying any feature selection method at all.

The highest performing combination for the two classification techniques and their corresponding macro F-score are presented in Table 6.5. These optimal models will be utilized for the remaining experiment.

6.2.4 Experiment 4: Classification of Adverse Event

The objective of this last experiment is to train the final models of NB and SVM using the selected techniques determined from the previous experiments. Finally, the performance of both models in classifying the adverse events into predefined

Technique	**MNB**	**SVM**
Stopword Removal	No	No
Stemming	No	Yes
Feature Extraction	BOW	TF-IDF
Ngram	Unigram	Unigram
Feautre Selection	MI	ANOVA F-value
macro F1	**0.7182**	**0.7175**

Table 6.5: Best Combinations for MNB and SVM with macro F1

categories will be presented.

Experimental Setup

For this experiment, MNB and SVM was trained and evaluated with a dataset balanced with their best corresponding balancing method found in subsection 6.2.1 and with the preprocessing, feature extraction, and feature selection method found in subsection 6.2.3. Lowercase conversion and the removal of special characters and numbers were also performed as preprocessing steps.

Experimental Results

The experimental results provide an overview of the precision, recall, and macro F1-scores for both MNB and SVM. The performance evaluation of these models offers valuable insights into their effectiveness in classifying adverse events. The final macro F1-score for MNB was 0.7182 and for SVM it was 0.7165.

Figure 6.2 and Figure 6.3 showcase the precision, recall, and F1-scores for the MNB and SVM models, along with the final accuracy. These figures also display the precision, recall, and F1-score for each category, offering a comprehensive summary of the models' classification performance. The precision measures the accuracy of positive predictions, while recall evaluates the model's ability to correctly identify positive instances. The macro F1-score provides a balanced measure of precision and recall.

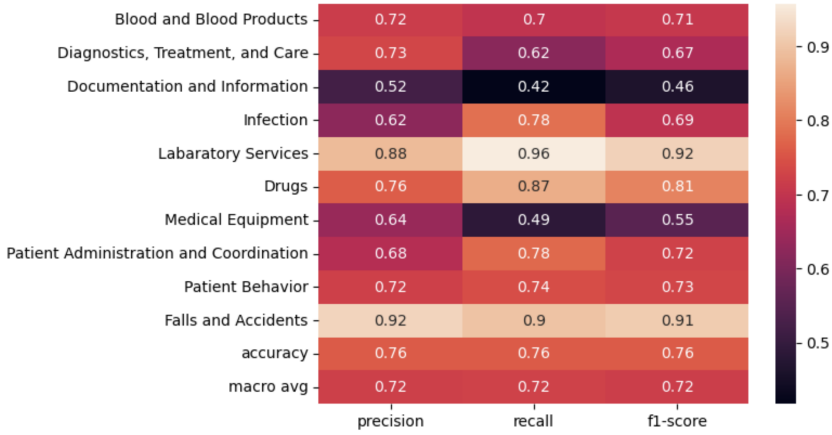


Figure 6.2: Overview of the results for MNB

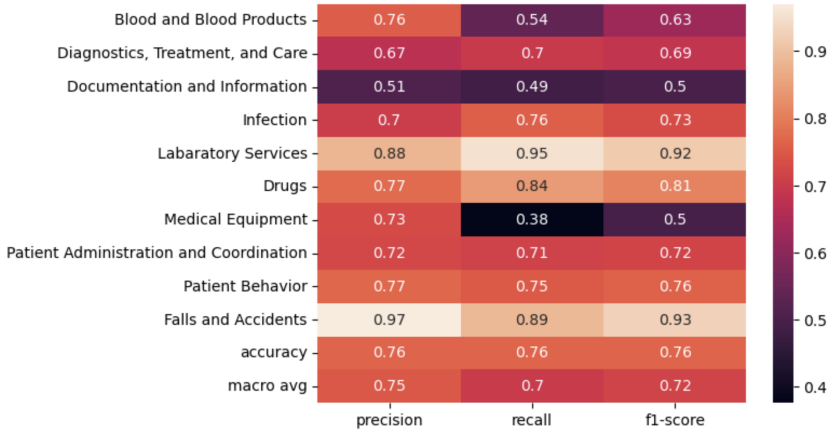


Figure 6.3: Overview of the results for SVM

In addition to the performance metrics, three confusion matrices have been generated for each model to gain further insights into the models' classification behavior and how they performed on each category. These confusion matrices will be presented in subsection 7.3.3 as they serve as an overview of the subsequent discussion and evaluation of the models on the different categories.

These figures offer a comprehensive analysis of the classification performance of both the MNB and SVM models. The precision, recall, and F1-scores provide a quantitative assessment, while the confusion matrices provide a visual representation of the models' classification behavior. The combination of these evaluations helps to understand the strengths and weaknesses of the models in classifying adverse events accurately.

Chapter 7

Evaluation and Discussion

This chapter will evaluate and discuss the results produced in this thesis. The chapter begins by evaluating the methods employed to gain domain insights and define the applications of machine learning in the adverse event reporting and analysis process. Furthermore, the three identified applications are examined, discussing their potential opportunities and challenges. It also examines the rationale behind selecting the classification of adverse events as the focus of the machine learning study. Additionally, the chapter dives into the machine learning study and evaluates the methodology, the final models, and their performance. It then proceeds to discuss the obtained results and their implications within a healthcare context.

7.1 Evaluation of the Identification of Potential Applications of Machine Learning

This section will assess the methodology employed to gather domain insights and identify potential applications of machine learning. To obtain information about the adverse events reporting and analysis process and potential machine learning applications, we have conducted research interviews with clinicians and experts.

The main method used in meetings with the clinicians was semi-structured interviews. This approach proved to be effective in gaining the necessary understanding of the domain and identifying key challenges and disadvantages of the current system. Considering our limited domain knowledge, conducting open-ended interviews has been particularly valuable. These interviews have provided the

opportunity to explore various aspects of the domain in a flexible and adaptable manner, allowing us to gather insights and perspectives that may have otherwise been overlooked.

One notable insight from these interviews was the importance of the NOKUP categories in the analysis of adverse events. During three of the four interviews, significant attention was directed toward the categories, prompting us to thoroughly study and comprehend their significance in the context of adverse event reporting. This led to the discovery of an application of machine learning that may not have been found in a more structured interview approach. The flexible nature of the interviews allowed for a more dynamic exploration of the domain, resulting in valuable insights and a more comprehensive understanding of the challenges and opportunities in adverse event reporting.

Engaging in discussions with various domain experts to gain diverse perspectives has proven invaluable in broadening our understanding of the subject matter. These interactions allowed us to explore different viewpoints, insights, and experiences related to adverse event reporting. However, in retrospect, it may have been beneficial to interview a wider range of domain experts, particularly individuals who have firsthand experience as the reporter's manager. A description of the reporter's manager in the process of adverse event handling can be found in subsection 6.1.1. Their perspectives could potentially have led us to other interesting applications of machine learning.

The unavailability of the dataset during the initial interviews significantly influenced the formulation of questions and ultimately impacted the results obtained. Without knowledge of the dataset's opportunities and limitations, it was challenging to tailor the questions specifically to how the data could be utilized. This limited understanding of the dataset's contents and structure hindered the depth and specificity of the discussions. Consequently, certain aspects and insights related to the dataset that could have been explored and addressed in the interviews may have been overlooked. The lack of access to the dataset underscores the importance of having comprehensive information about the data when conducting interviews, as it can enable a more informed and targeted discussion.

7.2 Discussion of the Identification of Potential Applications of Machine Learning

This section aims to discuss the three identified applications of machine learning that have the potential to enhance the adverse event reporting and analysis process. Each of the applications will be discussed in terms of their respective opportunities and challenges, examining their feasibility and effectiveness. Fur-

thermore, the shift from clustering to classification as the primary focus of the machine learning study will be explained and discussed. By examining this shift and its underlying reasons, we aim to provide an understanding of the decision-making process and its implications for the study's outcomes.

7.2.1 Potential Applications of Machine Learning

The potential applications of machine learning presented in subsection 6.1.2 will now be discussed individually. This discussion will explore how each application can contribute to enhancing the process of adverse event analysis, as well as any potential challenges that may arise.

Clustering of Adverse Events for Pattern Recognition

Clustering of adverse events could contribute to adverse event analysis by providing new insights and detecting patterns and trends more efficiently than humans. By grouping similar adverse events based on their inherent characteristics, clustering can reveal new knowledge and facilitate a more efficient detection of emerging issues. This, in turn, can aid in the formulation and implementation of targeted measures to prevent adverse events from occurring in the future. The clustering of adverse events could contribute to the process of adverse event analysis by potentially generating new insights and patterns in the adverse events occurring.

However, there are challenges associated with clustering adverse events. The unstructured nature of clinical text requires careful preprocessing, feature extraction, and feature selection to capture relevant information for effective clustering. Additionally, the brevity of adverse event reports may limit the amount of meaningful information available for clustering analysis. Selecting appropriate clustering algorithms and determining the optimal number of clusters also pose challenges. Interpretability is another consideration when using clustering techniques for adverse event analysis. The generated clusters may not always have clear interpretations, making it challenging for healthcare professionals to understand and utilize the results effectively.

Finally, it is important to highlight that the categorization using NOKUP remains essential to comply with the relevant laws and regulations outlined in Section 2.2. The use of clustering as a complementary tool does not replace the manual categorization currently employed. Instead, clustering can augment the analysis process by providing additional insights that may not be captured by the predefined NOKUP categories, thereby enriching the overall understanding of adverse events.

Summarization of Adverse Event Reports

Manually reviewing adverse events can be time-consuming and resource-intensive, especially for healthcare institutions with limited resources. To overcome this challenge, machine learning techniques such as NLP can be employed to automatically summarize adverse event reports, extracting key information and presenting it in a concise manner.

The text summarization of adverse events offers several advantages. Firstly, it can contribute to a more efficient analysis by extracting important details from large volumes of reports, saving valuable time for healthcare professionals. This efficiency allows for faster identification of trends and patterns, facilitating the monitoring of adverse events to improve patient safety and quality of care. Additionally, the standardization achieved through summarization algorithms ensures consistent extraction of relevant information, reducing variability in interpretation and analysis.

However, it is important to consider the limitations associated with text summarization of adverse event reports. One major limitation is the potential loss of detail in the summarized version, which may miss the important nuances and context from the original reports. This concern was emphasized by CG-1&2, who expressed a preference for a thorough reading of all reports rather than relying on a summarization tool. Additionally, the algorithms may struggle to comprehend the broader context and implications of adverse events, missing critical connections that human reviewers might capture.

Although the application of text summarization for adverse event reports may be relevant for institutions with limited resources, the clinicians we interviewed did not express a specific need for this capability. However, they recognized the potential value it could offer to other institutions. To validate the potential of the application for further research, additional interviews and investigations at those institutions would be necessary. Therefore, considering the limited demand expressed by the clinicians we consulted, this application was not prioritized as a focus for this thesis.

Automatic Classification of Adverse Events into NOKUP Categories

The classification of adverse events using machine learning has the potential to bring several benefits to the categorization process. By automating the classification process, a more consistent and objective approach can be achieved, reducing subjective variations in categorization and creating more accurate and reliable information for analysis and decision-making.

Automating the manual categorization process of adverse events offers several

benefits. It can reduce the inconsistencies and misclassifications seen in the current process, leading to more accurate classifications. This further helps the analysis of adverse events, as consistent classification provides a reliable and comprehensive picture of the adverse events occurring at the hospital. Additionally, it relieves the clinicians from performing this task, allowing them to focus their efforts on other responsibilities.

However, it is essential to acknowledge the challenges associated with implementing supervised machine learning models for classification. Acquiring high-quality labeled data for training can be a resource-intensive process, and ensuring consistent and accurate labeling can be challenging due to the subjective nature of categorizing adverse events. Additionally, the NOKUP categories may change over time due to updates in regulations or emerging healthcare trends. Machine learning models trained on a specific set of categories may struggle to adapt to new or modified categories without retraining. Continuous monitoring and updating of the model would be necessary to ensure it aligns with the latest categorization requirements.

7.2.2 Shift in Focus and Selection of Machine Learning Application

The goal of the thesis was not only to identify areas in which machine learning could contribute to the adverse event reporting and analysis process but also to conduct a machine learning study to explore the feasibility of the application of machine learning in one of these areas. In making this decision, the preferences of the clinicians played a pivotal role. Their domain knowledge and expertise were invaluable in the selection of the application that would have the greatest impact within the given scope and time constraints.

During the pre-project phase and prior to obtaining the dataset, the primary focus was on applying clustering techniques for adverse event analysis. This selection was based on the benefits of this application revealed in the interviews and assumptions about the dataset. However, upon receiving access to the initial dataset in January and engaging in new discussions with clinicians, it became evident that data was missing. As a result, a request for a new extraction of the dataset was made to ensure the inclusion of all available data. In February, we obtained the final version of the dataset, which included the NOKUP categories. This discovery prompted us to reassess the preferred approach in collaboration with the clinicians.

In order to decide which application of machine learning to use in the practical study, the final interview with CG-1&2 was crucial. As described in the summary of the interview presented in Appendix C.4, it became clear that the clinicians

were more interested in the classification of adverse events rather than unsupervised clustering. Additionally, the classification of adverse events was presented to a group of clinicians who provided positive feedback and expressed their agreement on its potential value. This group feedback further reinforced the selection of the application for the machine learning study. By prioritizing these crucial interactions with clinicians, we can place significant emphasis on their preferences and conduct a machine learning study that aligns with their identified areas of interest.

However, it is important to acknowledge that there may be other potential applications that were not explored in this thesis. Each hospital department or individual may have different perspectives and preferences regarding the applications they find most valuable. The exploration of other applications could be considered as future research to further enhance the adverse event reporting and analysis process using machine learning techniques and to cater to a broader range of perspectives within the hospital setting.

This shift in project focus underscored the importance of continuous and frequent communication with the clinicians throughout the research process. It highlighted the necessity of remaining open to new insights and perspectives throughout the research process and underscored the significance of constant communication and feedback with the clinicians. Through this collaborative approach, the automatic classification of adverse events was identified as a potential opportunity within the dataset and the thesis.

7.3 Evaluation of the Classification of Adverse Events

This section evaluates the machine learning study conducted to explore the application of classifying adverse events into the NOKUP categories. First, the methodology employed is investigated, followed by an examination of the two final classification models. Finally, a comprehensive evaluation of the models' performance in each category is presented.

7.3.1 Methodology

The aim of this thesis was to conduct a study to investigate the potential benefits of employing machine learning approaches in adverse event analysis. The ultimate objective of this machine learning study was to assess the feasibility of classifying adverse events into the NOKUP categories based on their free-text title and description. In order to ensure effective scope management, a number

of choices were made, and this section will outline these decisions along with their underlying reasoning.

Firstly, the classification focused only on the top-level categories and did not extend to their subcategories, despite NOKUP guidelines mandating the use of the most specific subcategory for event classification. The rationale behind this approach was that if it proved challenging to classify the top-level categories, classifying the subcategories would likely pose similar difficulties. After analyzing the data, it also seemed insufficient for the classification of subcategories as most of them were not present in the dataset.

Our intention was to explore the possibility of classification rather than seeking the most optimal approach. Consequently, the feature extraction and selection methods tested were those readily available through the Sci-kit learn Python library. In addition, only two different classification techniques were evaluated.

For future efforts, a comprehensive comparison of classification techniques could be explored to identify the optimal classification model for the task. In addition, the dataset should be revised to incorporate all the relevant subcategories so that the classification aligns with the guidelines presented in NOKUP. By undertaking these measures, we can enhance the effectiveness of the model and ensure its applicability to the real-life scenario of adverse event reporting in Norwegian hospitals.

7.3.2 The Classification Models

This section provides an evaluation of the performance and final model of the two classification techniques employed in this thesis. The models were optimized with the techniques that demonstrated the best performance in the experiments conducted in subsections 6.2.1, 6.2.2, and 6.2.3.

Multinomal Näive Bayes (MNB)

The final MNB model performed the best when no stopword removal or no stemming was employed, using the unigram BOW method for feature extraction and document representation, and mutual information for feature selection. The model has reasonable performance with a macro F1-score of 0.7182, providing a reliable measure over all the categories. It also achieved an accuracy of 0.7610, suggesting that the model predicts the categories correctly in 76.1% of the instances.

The results presented in Appendix D also revealed that the feature selection method did not have a significant effect on the MNB model, with a change in feature selection only lessening the performance by 0.02-1.30 in macro F1-score.

In addition, employing stop-word removal only lessened the performance by 0.28-0.90. These findings suggest that alternative MNB models utilizing different feature selection methods or stop word removal could be good candidates for the classification task as well.

On the other hand, certain processing techniques had a notable detrimental effect on the model's performance. The worst-performing models emerged from combinations that involved stemming and the use of both unigram and bigram tokens. Therefore, careful consideration should be given to these preprocessing techniques to avoid compromising the model's performance.

It is important to note that in subsection 6.2.2, MNB was identified as the best-performing NB variation for this specific classification task. It was selected as the starting point to conduct further experiments aiming to find the optimal combination of various preprocessing, feature extraction, and feature selection techniques. The purpose of these subsequent experiments was to refine the model and improve its performance. However, it is worth considering that there is a possibility that the other two variations could have demonstrated better performance in certain combinations.

Support Vector Machine (SVM)

The final SVM model demonstrated the best performance when stemming, but no stopword removal was utilized. It employed unigram TF-IDF as a feature extraction and document representation method. However, it performed best without any feature selection method. The model achieved similar performance to the final MNB model, with a macro F1-score of 0.7165 and an accuracy of 0.7645. The lower F1-score, but higher accuracy compared to the MNB model can be contributed to SVM's higher accuracy in the majority category, Diagnostics, Treatment, and Care.

Similar to MNB, SVM demonstrates a marginal difference in performance in the top-performing combinations of preprocessing, feature selection, and feature extraction techniques, as seen in Appendix D. The selection of feature selection and stopword removal have a marginal impact on the performance. Between the six top-performing models, the macro F1-score only varies by approximately 0.009. SVM is not as affected by different preprocessing choices as MNB, but the classification technique still displays weaker performance when using both unigram and bigram tokens.

The experiments only included an RBF-kernel-based SVM model, as it was chosen for its well-established performance in various applications, including text classification tasks. However, for future efforts, it might be interesting to explore the performance of alternative kernel functions in the classification of adverse

events. For instance, the prior research by Ong et al. [2010] found that SVM with linear and RBF kernel produced comparable results.

7.3.3 Results by Category

This section provides an in-depth evaluation of both models' performance across the different categories. Instead of focusing on the overall performance, it delves into each category to gain a comprehensive understanding of the results. For further discussion of the results, refer to Section 7.4. The precision, recall, and F1-score for each category can be found in Figure 6.2 for MNB and in Figure 6.3 for SVM.

To aid the following evaluation, three confusion matrices have been generated for both classification models. The raw confusion matrices are found in Appendix E as they are not directly referred to in the evaluation. The normalized confusion matrices provide a more accurate representation of the classification performance by accounting for class imbalance, and will therefore be the focus of the evaluation. However, the unnormalized confusion matrices can offer supplementary information and are thus included in the appendix for reference.

Figure 7.1 and Figure 7.2 present the confusion matrices normalized over the true labels. This normalization highlights the models' classification performance relative to the actual distribution of classes. The diagonal represents the recall for each class. Conversely, Figure 7.3 and Figure 7.4 show a confusion matrix normalized over the predicted labels, providing insights into how the models' predictions align with the predicted class distribution. In these confusion matrices, the diagonal represents the precision for each class.

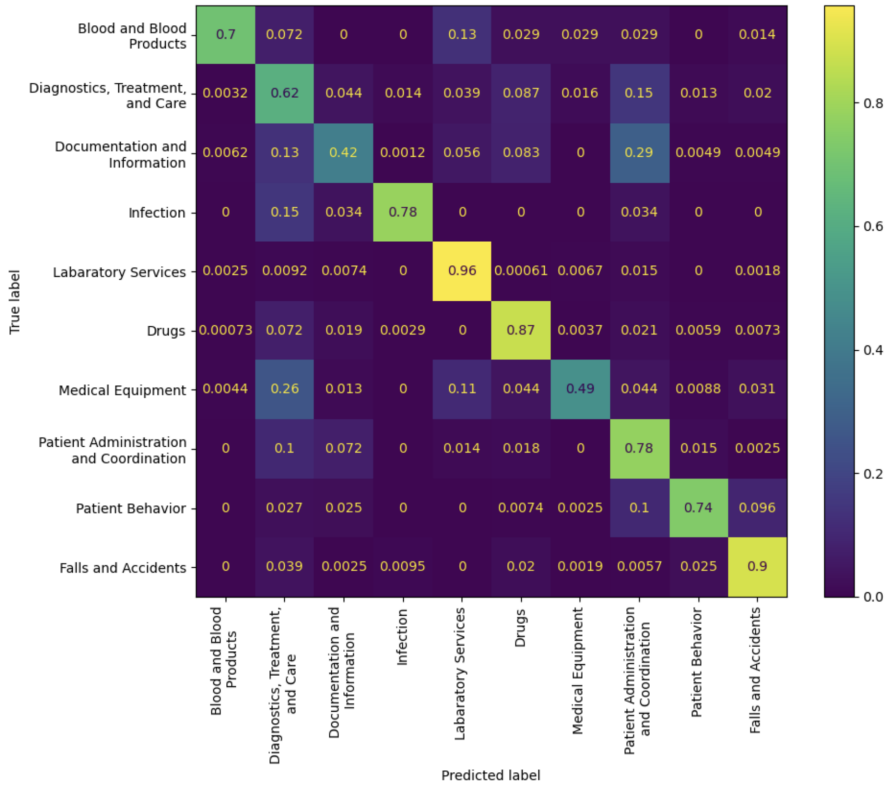


Figure 7.1: Confusion matrix for MNB normalized over the true labels. The diagonal represents the **recall** for each class.

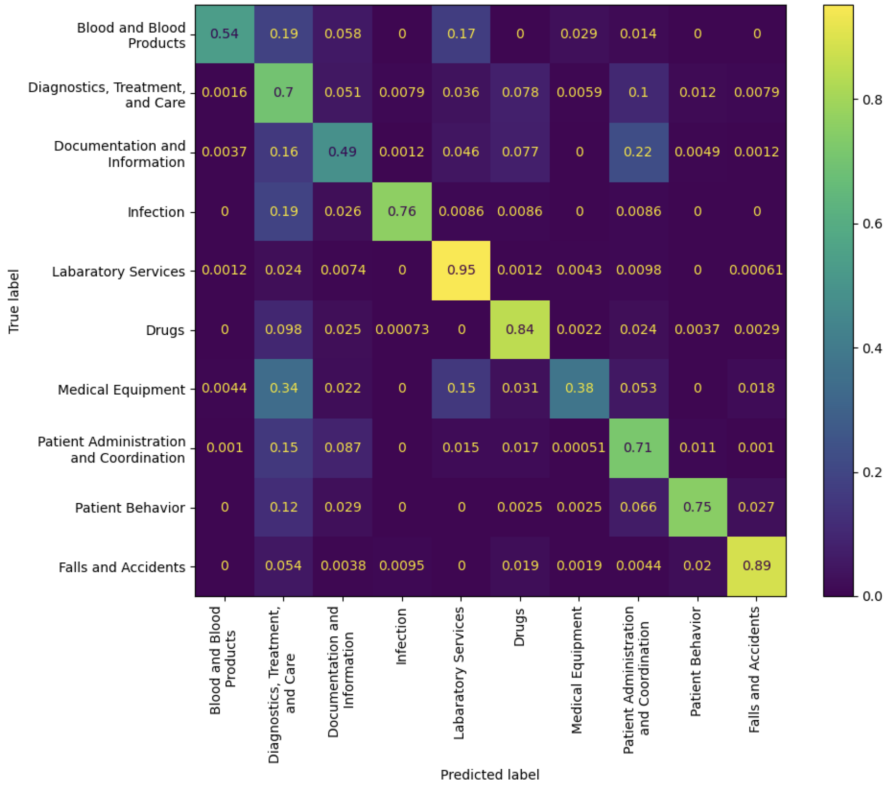


Figure 7.2: Confusion matrix for SVM normalized over the true labels. The diagonal represents the **recall** for each class.

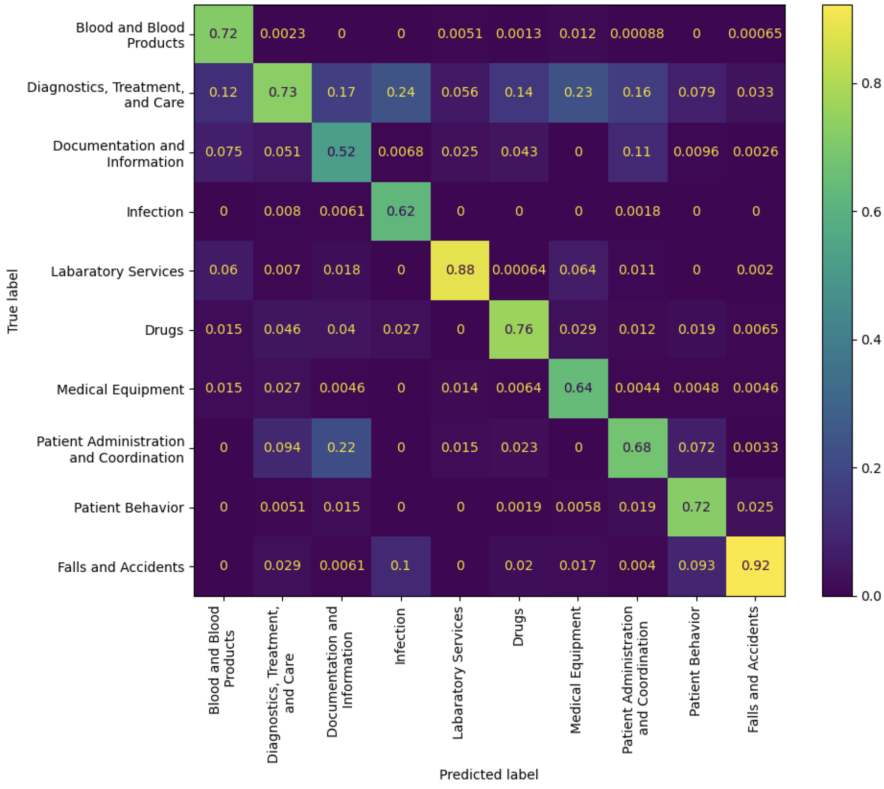


Figure 7.3: Confusion matrix for MNB normalized over the predicted values. The diagonal represents the **precision** for each class

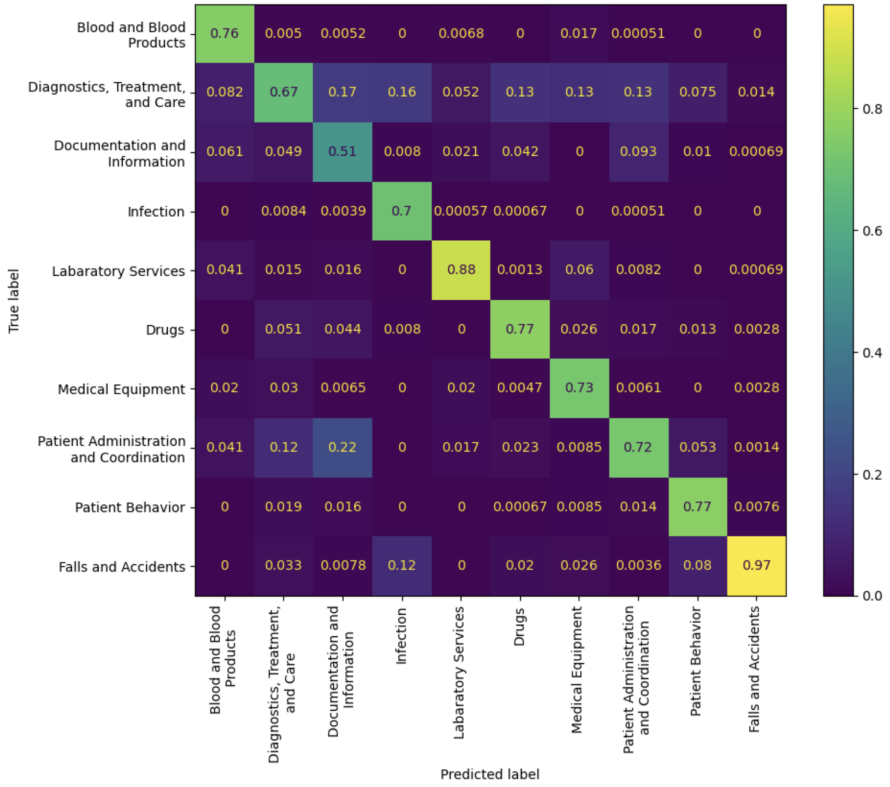


Figure 7.4: Confusion matrix for SVM normalized over the predicted values. The diagonal represents the **precision** for each class

Blood and Blood Products

The models demonstrate varied performance in classifying Blood and Blood Products with MNB achieving an F1-score of 71%, whereas SVM archives an F1-score of 63%. This discrepancy can be contributed to SVM low recall score as it often predicts adverse events that belong to the Blood and Blood Products category as Diagnostics, Treatment, and Care (19%) or Laboratory Services (17%). MNB also has this tendency, but to a lesser extent, with predicting Blood and Blood Product as Diagnostics, Treatment, and Care in 7.2% of the instances and Laboratory Services in 1.3%. On the other hand, the precision score is affected by the same categories as well as the Document and Information category. Diagnostics, Treatment, and Care have the greatest impact with MNB misclassifying this category as Blood and Blood Products in 12% of its predictions.

Diagnostics, Treatment, and Care

Both MNB and SVM achieve moderate F1 scores of 67% and 69% respectively for Diagnostics, Treatment, and Care. Even though the F1-scores are relatively similar, it's worth noting the differences between the models in the precision and recall score for this category. SVM achieves a precision score of 67% and a recall score of 70% while the MNB model achieves a precision score of 73% and a recall score of 62%. This means that the SVM model is better at capturing a higher proportion of adverse events labeled as Diagnostics, Treatment, and Care, while the MNB model is better at only predicting adverse events that are labeled as Diagnostics, Treatment, and Care.

The most notable confusion arises with the category Patient Administration and Cooperation. In the case of MNB, 15% of instances that should have been classified as Diagnostics, Treatment, and Care are misclassified as Patient Administration and Cooperation. Similarly, SVM misclassifies 10% of such instances. On the other hand, when the models predict Diagnostics, Treatment, and Care, in 9% of the instances the true category is actually Patient Administration and Cooperation for MNB, and 12% for SVM.

A challenge that this category presents for the models, is the high number of false positives and false negatives it generates for the other categories. This can be observed in the category's Predicted column in Figure 7.1 and Figure 7.2, and the corresponding True row in Figure 7.3 and Figure 7.4. The potential reasons behind this confusion are further discussed in subsection 7.4.3.

Documentation and Information

The category Documentation and Information demonstrates a poor F1-score, with a score of 46% and 50% for MNB and SVM respectively. This makes it the worst-performing category for MNB, and one of the worst for SVM. The notable challenge lies in the confusion between this category and Diagnostics, Care, and Treatment, as well as Patient Administration with a significant number of both false negatives and false positives.

Infection

Both models achieve reasonable F1-score for the Infection category, with 69% for MNB and 74% for SVM. The only notable confusion for the recall score occurs with Diagnostics, Treatment, and Care, where instances of Infection are falsely predicted in 15% of cases for MNB and 19% of cases for SVM. However, when predicting Infection, the model confuses both Diagnostics, Treatment, and Care and Falls and Accidents for this category. This has an impact on the precision score, with the models predicting this category in 24% and 10% of instances instead of Diagnostic, Treatment, and Care and Falls and Accidents for MNB, and in 16% and 12% of instances for SVM.

Laboratory Services

The Laboratory Services category is the best-performing category for MNB and the second-best for SVM with an F1-score of 92% for both models. However, it is important to note that this category also exhibits a high number of false positives, with instances being predicted as Laboratory Services when they do not belong to this category. This is particularly evident when comparing the models' precision scores with their recall scores.

Drugs

Both MNB and SVM achieve a relatively high F1-score for the Drugs category, with a score of 81% for both models. However, there are a significant number of misclassification, where the models predict Drug instead of Diagnostics, Treatment, and Care in 14% and 13% of instances for MNB and SVM, respectively, as evident in Figure 7.3 and Figure 7.4. Additionally, the models mistake Drugs adverse events as Diagnostics, Treatment, and Care in 7.8% and 9.2% of instances.

Medical Equipment

The Medical Equipment category presents challenges for both MNB and SVM, with F1-scores of 55% and 50% respectively, representing one of the worst-performing categories for SVM. It exhibits a significant number of false positives in the Diagnostics, Treatment, and Care, and Laboratory Services categories. The impact is evident in the recall score of the SVM which stands at 38%. This indicates that SVM frequently fails to identify adverse events within the Medical Equipment category, misclassifying them as other categories.

Patient Administration and Coordination

Patent Administration and Coordination achieves a reasonable F1-score of 72% for both models. This category displays confusion with both Diagnostics, Treatment, and Care, and Documentation and Information, which has the most significant impact on the category's recall and precision score. This indicates that the model finds it hard to separate these categories and classifies them interchangeably.

Patient Behaviour

The Patient Behavior category performs reasonably for both models, with F1-scores of 73% and 76%. However, both models confuse adverse events in the Patient Behavior category with Patient Administration and Coordination with false negative rates of 10% and 6.6% for MNB and SVM respectively. Additionally, MNB confuses this category with Falls and Accidents (9.6%), while SVM confuses it the most with Diagnostics, Treatment, and Care (12%). However, the precision score is impacted by all three categories for both models.

Falls and Accidents

Falls and Accidents is the second-best performing category for MNB with a F1-score of 91% and the best-performing for SVM with 93%. The most notable characteristic for this category is seen with the SVM model, with the variation between the recall score of 89% and the precision score of 97%. This indicates that SVM has an almost perfect score when it predicts Falls and Accidents, however, it does not capture all the instances within the category.

7.3.4 Clinical Evaluation of the Results

During the discussion of the models' performance across the different categories with CG-1&2, it became evident that there were resemblances between the models' classification and the classification that happens at the hospital. The cate-

gories in which the models excelled, namely Falls and Accidents and Laboratory Services, coincided with CG-1&2 finding these categories to be the clearest and easiest to classify during the reporting process. In addition, the categories in which the models encountered confusion or challenges were unsurprising to the clinicians, as the results aligned with their own experiences.

The prevalence of false positives in the Diagnostics, Treatment, and Care category, where the model frequently misclassifies adverse events that don't actually belong to this category, was understandable to CG-1&2. They acknowledged that numerous events could be construed as treatments or procedures, as adverse events occurring in other categories could also impact patient care and treatment. This intricate relationship between adverse events and their effects on treatment and care often leads to misclassifications, stemming from the interpretation of the adverse event's impact.

An illustrative example highlighting the challenges in accurately classifying adverse events can be observed in the underperformance of Medical Equipment, often leading to misclassification as Diagnostics, Treatment, and Care. During the clinical evaluation of the results, CG-1&2 brought attention to a scenario where an adverse event could be attributed to a malfunctioning medical device. However, due to the disruption or alteration of a procedure caused by the equipment malfunction, the adverse event was mistakenly classified under Diagnostics, Treatment, and Care. This discrepancy demonstrates the challenges faced in accurately categorizing adverse events, especially when multiple factors are involved, such as equipment malfunction impacting treatment processes.

The correlation between the performance of the classification models and the clinicians' experiences with the misclassification of adverse events in the hospital highlights a significant issue within the dataset: these misclassifications are a part of the current classification labels. These misclassifications adversely affect and limit the potential effectiveness of the classification models. The implications of this problem will be explored and expanded upon in the subsequent section.

7.4 Discussion of the Classification of Adverse Events

This section encompasses the examination of how the dataset's limitations impact the classification performance, a discussion of the classification results, and an exploration of the potential implications derived from these outcomes.

7.4.1 Performance Analysis and Variations in Adverse Event Classification

The final classification models achieved a reasonable overall performance with an F1-score of 0.7182 for MNB and 0.7165 for SVM. These results indicate that the models demonstrated a reasonable ability to classify adverse events. However, a closer analysis of the performance across different categories reveals significant variations, underscoring the intricate nature of adverse event classification.

When evaluating the results, in subsection 7.3.3, certain categories exhibited outstanding classification outcomes, while others presented distinct challenges. Notably, the categories Fall and Patient Accidents and Laboratory services demonstrated a strong performance for both MNB and SVM, indicating that the models were effective in identifying and classifying adverse events within these categories.

On the other hand, the categories Medical Equipment and Documentation and Information posed challenges for the models, as evidenced by the lower F1-scores in these categories. These results indicate the potential subjectivity surrounding adverse events related to these categories. Upon clinical evaluation of the results, as described in subsection 7.3.4, it was observed that adverse events stemming from medical equipment could sometimes be misclassified as Diagnostics, Treatment, and Care if the adverse event's impact affected a patient's treatment or care process. Similarly, the Documentation and Information category encountered challenges where adverse events belonging to this category shared similarities with other categories.

Overall, while the classification models achieved promising results in a broader context, the varying performance across different categories underscores the need for continued research and development in this field. By identifying the specific areas of strength and weakness, future work can focus on refining the models and addressing the specific challenges posed by different adverse event categories. It is also crucial to consider how these models would perform in a healthcare context, which is further discussed in subsection 7.4.6.

7.4.2 The Potential Misclassification in the Classification Labels

The user interviews highlighted issues related to misclassification and inconsistencies in the classification of the adverse events within the current workflow. Considering that the dataset includes the adverse events reported from 2015 to 2022, it is reasonable to assume that some of the adverse events within the dataset may have been misclassified. Unfortunately, due to time constraints and the vast size of the provided dataset, it was not possible to review all the adverse events

and the correctness of their label. As a result, there is a possibility that the dataset contains inconsistent and misclassified adverse events.

During the clinical evaluation of the results with CG-1&2, it became evident that the impact of misclassifications in the labels was significant. The confusion displayed by the models in categorizing adverse events mirrored the challenges often encountered by human reporters when classifying such events. This observation suggests that there are limitations to the models' performances with the current dataset, as the presence of misclassifications introduces a substantial level of confusion.

A notable example of misclassification was observed during the analysis of adverse event reports with the title "fall." This particular title was identified as the most common among all adverse events, as indicated in the EDA conducted in Section 5.3. When examining the Event Type of these adverse event reports, we observed that some of these events were not classified as Falls and Accidents category, as shown in Table 7.1. Further investigation of the descriptions associated with these events, it became evident that they should indeed be classified as Falls and Accidents. This serves as a clear illustration of the potential misclassifications present within the dataset.

Type of Event	**Count**
Falls and Accidents	3869
Diagnostics, Treatment, and Care	14
Patient Behavior	4
Patient Administration and Cooperation	2

Table 7.1: The distribution of Event Type for adverse event reports with the title "fall"

These discrepancies can compromise the reliability and consistency of the classification labels, making it more challenging for the model to accurately classify future adverse events. Therefore, it is imperative for future efforts to address these issues by conducting comprehensive data validation and ensuring the accuracy of the classification labels.

However, conducting a thorough review and analysis of the current labels to ensure their correctness presents challenges due to the complexity of adverse event classification. Adverse event classification relies on subjective judgment, varying interpretations among human reporters, and contextual factors that can influence how an event is categorized. Addressing these challenges requires a collaborative effort among domain experts, clinicians, and machine learning practitioners to establish clear guidelines, standardize the classification process, and provide

comprehensive training and support. This process could enhance the overall reliability and performance of the models in classifying adverse events effectively.

7.4.3 Insights into the Diagnostics, Treatment, and Care Category

The Diagnostics, Treatment, and Care category presents unique challenges in the classification of adverse events. Additionally to being the majority class, this category is characterized by a high level of confusion, resulting in a significant number of false negatives and false positives, as summarized in Table 7.2 and Table 7.3. In the tables, the label "positive" refers to Diagnostics, Treatment, and Care while "negative" encompasses the rest of the categories.

		Predicted (MNB)	
		Positive	Negative
Actual	Positive	1562	966
	Negative	476	7703

Table 7.2: Confusion Matrix (One vs Rest) for Diagnostics, Treatment, and Care in the MNB model

		Predicted (SVM)	
		Positive	Negative
Actual	Positive	1761	767
	Negative	848	7331

Table 7.3: Confusion Matrix (One vs Rest) for Diagnostics, Treatment, and Care in the SVM model

One of the main contributors to the challenges in classifying adverse events in this category is the inherent complexity and domain-specific nature of the category itself. Adverse events that impact patient treatment or care processes, even if they initially belong to other categories, often get misclassified as Diagnostics, Treatment, and Care due to the shared characteristic of affecting patient care. This is highlighted in the scenario described in subsection 7.3.4. These misclassifications underscore the need for domain expertise and clinical knowledge to accurately discern and classify these events.

Furthermore, the fact that the Diagnostics, Treatment, and Care category is the majority class can have an impact even after balancing the dataset. Balancing the dataset helps address the issue of class imbalance, but it does not eliminate the challenges associated with the majority class. Models trained on the balanced dataset may still exhibit a bias towards the majority class, affecting their performance and potentially leading to a higher number of false negatives and false positives in the minority classes.

The considerable confusion and misclassifications within the Diagnostics, Treatment, and Care category have a direct impact on the recall and precision scores of the other categories. Adverse events that should be correctly classified in their respective categories may be erroneously classified as Diagnostics, Treatment,

and Care due to the overlap and ambiguity in their characteristics. This misclassification hinders the accurate assessment of the performance of other categories and may affect the overall effectiveness of the classification models.

In conclusion, the Diagnostics, Treatment, and Care category poses significant challenges in adverse event classification, with a notable presence of false negatives and false positives, impacting the recall and precision scores of the other categories. The confusion and misclassifications arise from the intricate nature of adverse events that impact patient treatment or care processes, but should be classified based on their underlying causes, which may belong to other categories. By addressing these challenges and refining the classification models, we can strive for more accurate and reliable adverse event classification within the Diagnostics, Treatment, and Care category. This will lessen its impact on the other categories while improving the discrimination between different types of events within this category.

7.4.4 Patterns in Adverse Event Descriptions: Valuable Features or Random Noise?

The EDA revealed common patterns used in the adverse event descriptions, as seen in the ten most common descriptions in Table 5.4. Initially, these descriptions raised some concerns about noise. However, when exploring these patterns further we found that approximately all the frequently used descriptions belonged to the same category, Patient Administration and Cooperation. The exact distribution of the top two of these descriptions can be found in Table 7.4 and Table 7.5.

Type of Event	**Count**
Patient Administration and Coordination	268
Documentation and Information	39
Diagnostics, Treatment, and Care	15
Laboratory Services	7
Drugs	6
Medical Equipment	1
Patient Behavior	1

Table 7.4: The distribution of the most frequent adverse event description ("se vedlegg"/"see attachment")

Type of Event	Count
Patient Administration and Coordination	59
Diagnostics, Treatment, and Care	8
Laboratory Services	6
Documentation and Information	4

Table 7.5: The distribution of the second most frequent adverse event descriptions ("annet"/"other")

We also observed a pattern of four sentences being used frequently, either individually or in conjunction with each other. Notably, these sentences, as shown in Table 7.6, sometimes featured numbering as well. These observations suggested that reporters had access to and frequently utilized a list, potentially revealing an underlying pattern. After discussions with CG1&2, we discovered that these descriptions were part of a report used for cooperation discrepancies.

1. hjemsendelse før aksept
2. manglende epikrise
3. mangler ved varsling
4. mangler ved medisinliste

Table 7.6: List used as descriptions in the adverse event reports

In addition, there were a significant number of titles and descriptions including prefixes of the reference codes to these reports, with 1278 titles and 447 descriptions including the former reference prefix "ESA" and 230 titles and 207 descriptions including the current reference prefix "Elements". The total distribution of all the patterns related to the cooperation discrepancies reports is found in Table 7.7.

In conclusion, these common adverse event titles and descriptions possess the potential to serve as both valuable, discriminatory features and noise for the classification models. A significant percentage (81.4%) of instances featuring these patterns belong to the Patient Administration and Cooperation category, so they can be considered valuable features for classifying such adverse events. However, considering the current dataset's limitations and the possibility of false labels, it is also possible that the presence of these features might confuse the models, particularly in cases belonging to the Documentation and Information and Diagnostics, Treatment, and Care categories. This confusion is evident in the results, particularly in Figure 7.1 and Figure 7.2, which displays considerable misclassification between these categories, potentially attributed to these patterns.

Type of Event	Count
Patient Administration and Cooperation	1455
Documentation and Information	183
Diagnostics, Treatment, and Care	69
Drugs	40
Laboratory Services	25
Falls and Accidents	6
Medical Equipment	2
Patient Behavior	2

Table 7.7: Distribution of adverse events with title or description related to the cooperation discrepancy report

7.4.5 Abbreviations and Common Spelling Errors

Abbreviations and common spelling errors in the dataset present notable challenges and are potential sources of error in our analysis. Throughout the EDA, we encountered several abbreviations and different spellings for the same words.

The dataset exhibited a considerable number of common spelling errors. Alsmadi and Gan [2019] highlights the prevalence of these challenges in short texts, such as tweets or quick messages. These types of texts resemble the environment in which adverse reports are written, as they are often composed in a fast-paced work setting where spelling accuracy may not be prioritized. Such spelling errors can introduce noise into the dataset, making it more difficult to extract accurate insights and affecting the overall performance of our models.

The presence of abbreviations and spelling errors in the dataset raises concerns regarding data quality and reliability. These issues may have implications for the validity of our findings. It is important to acknowledge these limitations and consider potential strategies for addressing them in future research. To reduce the impact of abbreviations and spelling errors, steps such as implementing data validation checks, employing automated spelling correction algorithms, or performing manual data cleaning can be considered. By addressing these challenges, we may enhance the accuracy and integrity of the dataset, ultimately leading to more reliable results and conclusions.

7.4.6 Assessing the Implementation of the Current Models

The successful implementation of machine learning in healthcare systems heavily relies on trust in the predictive systems used [Schwartz et al., 2022]. The accuracy of the model is a significant factor affecting the user's trust in the sys-

tem [Jung et al., 2020]. Considering the current models have an accuracy rate of 76.1% for MNB and 76.5% for SVM, concerns arise regarding the immediate implementation and the establishment of trust.

It is important to carefully assess the implications of implementing these models within a healthcare system, and also in the context of analyzing adverse events. Although these accuracy rates may be considered reasonable in a different context, it is crucial to recognize the risks associated with inaccurate predictions within a healthcare setting. Such inaccuracies can significantly impact the identification and thereby potential prevention of adverse events. Inaccuracies can also lead to a false portrayal of adverse events occurring at the hospitals, distorting the overall understanding of the situation.

The aim of these models is to reduce misclassifications and establish more consistent categorization across medical perspectives and clinics. However, considering the current performance of the models, there is no evidence that this would enhance the categorization in the adverse event reporting process. Specifically, there is no indication that the reporters misclassify adverse events in 25% of the instances, which is the level of misclassification our models would introduce.

However, both models demonstrate promising performance in certain categories. The F1-scores for the Laboratory Services and Falls and Accidents categories exceed 90% for both models, indicating reliable results that are likely to inspire trust. While the current models may not immediately enhance the overall categorization of adverse events, the successful performance in these specific categories suggests the potential for improvement and expansion. By addressing the limitations and focusing on enhancing the models' capabilities, we can work towards a more comprehensive and reliable adverse event reporting system.

The study by Schwartz et al. [2022] found that overall accuracy alone is insufficient for establishing trust in predictive healthcare systems. It highlights the importance of understanding how the decisions are made by the model, as transparency and interpretability play crucial roles in building trust. Making the classification models explainable could enhance trust and help ensure the validity of the results. By providing insights into the models' inner workings, explainability can help bridge the gap between technical predictions and human understanding, empowering clinicians to make more informed decisions based on the models' predictions.

Furthermore, the European Union (EU) published guidelines for artificial intelligence in 2019, which include ethical guidelines for transparency and explainability [Madiega, 2019]. These guidelines are in line with the General Data Protection Regulation (GDPR) regulations. The inclusion of transparency and explainability provisions in the EU guidelines demonstrates the recognition of these principles

as crucial aspects of responsible artificial intelligence deployment, including in the healthcare domain. Therefore, future efforts should focus on implementing explainability features in NB and SVM models to comply with these guidelines and enhance trust in the system.

Chapter 8

Conclusion and Future Work

In this concluding chapter, we summarize the key findings and contributions of this master thesis, which focused on the role of machine learning in adverse event reporting and analysis. Lastly, we present proposals for future work.

8.1 Conclusion

In this section, we will address the research questions presented in Chapter 1, and to which extent they are answered through this thesis.

Research question 1: *What are the challenges of the adverse event reporting and analysis process, and how can the application of machine learning aid in these challenges?*

Through interviews and discussions with clinicians, we have identified several areas where the use of machine learning potentially could contribute to improving the process of adverse event reporting and analysis. As the domain proved to be more complex than initially anticipated, more time than expected was dedicated to understanding the domain and accurately pinpointing how machine learning could be applied.

The main finding is the use of machine learning to standardize the process of categorizing adverse events. The categorization follows a national guideline, NOKUP, presented by Saastad et al. [2015], in order to produce national statistics and mon-

itor and compare different hospitals. The categorization was mentioned by all interviewees as a part of the process that had the potential for improvement, due to the different interpretations of the guideline used, the occurrences of misclassification, the time spent on categorization, and because it creates the basis for further analysis and compliance with laws and regulations.

Secondly, the use of machine learning to identify patterns and relationships within adverse event data was also discussed with the clinicians. By applying unsupervised clustering one could uncover hidden associations and provide a deeper understanding of the underlying factors contributing to adverse events, that are hard to detect for the human eye. This application of machine learning could be valuable for hospitals in order to potentially find new patterns and insights that could lead to better patient safety.

A third area discussed is the use of machine learning and NLP in order to produce summaries of adverse event reports. This could especially be valuable for hospitals and institutions that have fewer resources to use for the continuous improvement of patient safety including analysis and monitoring of adverse events. Providing clinicians with information that is more easily analyzed can contribute to improving the adverse event process. However, this application of machine learning was not a priority with the clinicians participating in our thesis and was thus not selected for the machine learning study.

Research question 2: *To what extent can a selected application of machine learning deliver reliable results for healthcare professionals working with adverse events?*

The objective of this research question was to conduct a study focusing on one of the identified areas for machine learning application in the adverse event reporting and analysis process to examine if it could produce reliable results. The selected application was the automatic classification of adverse events into the NOKUP categories.

By exploring the performance of two machine learning techniques, Näive Bayes (NB) and Support Vector Machine (SVM), in classifying adverse events, valuable insights were gained regarding the feasibility and potential of using these models in practice. The findings indicate that the models could achieve promising results in some categories, such as Laboratory Services and Falls and Accidents with F1-scores exceeding 90%. These specific results suggest that the models' predictions in these categories can be considered reliable and potentially inspire trust in the system among healthcare professionals.

However, it is crucial to address the challenges encountered in other categories, namely Document and Information and Medical Equipment, where both models

showed poor F1-scores, leading to unreliable results and misclassifications. This highlights the importance of careful consideration of the overall performance and implications of the models within the adverse event reporting system. Although the overall achieved F1-scores of 0.7182 for MNB and 0.7165 for SVM may be considered reasonable in general, their implications in a healthcare setting raise concerns. Inaccurate predictions have the potential to significantly impact the identification and prevention of adverse events, potentially resulting in patient harm and a distorted understanding of the situation.

Through the EDA and the research interviews, it became apparent that the current dataset suffers from misclassifications and inconsistent categorization of adverse events in the Type of Event column, which served as the classification labels for the machine learning study. This observation was further supported during the clinical evaluation, where representatives from Corporate Governance identified similarities between the models' results and the general pattern of misclassification by the clinicians. Addressing these issues through expert validation of the existing categories holds great potential for improving the performance of the models and achieving a reliable automatic classification system for adverse events.

In conclusion, while machine learning for automatic classification of adverse events holds promise for delivering reliable and meaningful results in the future, careful evaluation of performance, addressing limitations, and promoting transparency are essential for its successful implementation in healthcare settings. Future research and development efforts should focus on improving model accuracy, reducing misclassifications, and enhancing the interpretability of machine learning algorithms to ensure their effectiveness and trustworthiness in supporting healthcare professionals.

8.2 Contributions

This section will outline the contributions of this thesis. The contributions encompass valuable domain insights gathered from clinicians and experts, as well as a machine learning study conducted on an unexplored dataset consisting of Norwegian clinical data.

Through interviews with clinicians and experts, we have identified and documented the adverse event reporting process at St. Olav's Hospital, which was previously not available to the public. This documentation, detailed in subsection 6.1.1, provides valuable insights into the current workflow and lays the foundation for future research and improvements in the reporting process.

Furthermore, we have identified specific areas where machine learning techniques

can aid in adverse event reporting and analysis. By conducting interviews with domain experts and clinicians, we have ensured that our findings align with the needs and perspectives of the healthcare professionals involved. This alignment with the clinicians strengthens the practical relevance and potential impact of our work.

In addition, a machine learning study was conducted to further explore the feasibility of the automatic classification of adverse events based on their title and description. While the current models still need refinement, the study showed promise and future potential. Additionally, it serves as evidence of how two classification techniques can effectively work with Norwegian clinical text data. This proof-of-concept sets the stage for future improvements and research regarding automated adverse event classification.

Finally, the thesis has also contributed to the identification of two additional potential areas for future research. The first area is the clustering of adverse events to enhance the current analysis process and uncover patterns not easily recognizable by humans. The second is the summarization of adverse events for a more manageable analysis as the manual review of adverse events can be both time-consuming and resource-intensive. These identified applications of machine learning can create further opportunities for new research in the collaboration between computer scientists and clinicians.

Overall, this thesis contributes to the knowledge and practical application of adverse event reporting and analysis in Norway, bridging the gap between research and practice by involving clinicians, documenting existing processes, and showing the potential of machine learning in improving adverse event reporting and analysis process and thereby patient safety and healthcare outcomes.

8.3 Future work

This section outlines the areas that can be further developed based on the foundation laid in this thesis. It encompasses various possibilities for extending the research and opportunities for exploring new areas of research within the field of computer science in Norway.

An important aspect of extending the research conducted in this thesis involves validating the classification labels in the dataset to ensure their accuracy and reliability for adverse event classification. This validation process aims to verify that the assigned NOKUP categories align with the actual content and characteristics of the adverse events. However, it is important to acknowledge that validating classification labels in the dataset is not a straightforward task due to the inherent subjectivity involved. Therefore, the active participation of clinicians and experts

is crucial in ensuring the validity and credibility of the classification process, reinforcing the integrity of the dataset, and enhancing the overall effectiveness of the adverse event classification performance.

Additionally, to meet the requirements for adverse event categorization as presented by Saastad et al. [2015] in the national guidelines, it is important to include the subcategories of NOKUP in future classification tasks. By including the subcategories, the classification process meets the requirements for adverse event reporting, ensuring its relevance and practical use at the hospitals.

Another aspect of extending this research is to revise and improve the existing machine learning models used for adverse event classification. While the machine learning study presented promising results, there is still room for refinement and optimization. The performance of the current models can be further enhanced by fine-tuning their hyperparameters and incorporating advanced techniques such as ensembling learning. Furthermore, to enhance the performance of the models, additional preprocessing steps can be employed. For instance, addressing common spelling errors and abbreviations frequently used by reporters can help improve the accuracy of the classification process. By incorporating techniques to handle such variations in the text, we can ensure that the models are more robust and capable of handling real-world adverse event reports.

Additionally, it is crucial to explore alternative machine learning approaches for adverse event classification. One such example, as mentioned by Alsmadi and Gan [2019], is the utilization of neural networks. Whereas SVM is more often utilized in text classification tasks, neural networks have demonstrated remarkable capabilities in various natural language processing tasks and have the potential to capture intricate patterns and relationships within textual data. Investigating the suitability and performance of neural network architectures can provide valuable insights into their effectiveness in handling adverse event data.

In addition to the automatic classification of adverse events, two other potential areas where machine learning can be applied to enhance the adverse event reporting process were identified. These areas have been verified as important by clinicians, indicating their interest and potential impact. These additional use cases are discussed in subsection 6.1.2 and subsection 6.1.2. However, further research is needed to establish the theoretical foundation and assess the practical feasibility of applying machine learning techniques in these areas. Future work should focus on exploring and validating these potential solutions, considering their implications for improving the accuracy and efficiency of adverse event reporting.

To conclude, this thesis has established a foundation for future advancements in the classification of adverse events at Norwegian hospitals. While the presented

research has provided valuable insights, there is still room for improvement and opportunities for further exploration. By continuing the research, significant strides in enhancing patient safety and optimizing health care can be made by utilizing adverse event reports. The findings and methodologies presented in this thesis serve as a starting point for future researchers to build upon, ultimately contributing to the ongoing efforts to improve healthcare outcomes and ensure the well-being of patients.

Bibliography

Lee Adler, Charles R Denham, Marjorie McKeever, Robert Purinton, Franck Guilloteau, J David Moorhead, and Roger Resar. Global trigger tool: implementation basics. *Journal of Patient Safety*, pages 245–249, 2008.

Majid Hameed Ahmed, Sabrina Tiun, Nazlia Omar, and Nor Samsiah Sani. Short text clustering algorithms, application and challenges: A survey. *Applied Sciences*, 13(1):342, 2022.

Issa Alsmadi and Keng Hoon Gan. Review of short-text classification. *International Journal of Web Information Systems*, 15(2):155–182, 2019.

Murugan Anandarajan, Chelsey Hill, Thomas Nolan, Murugan Anandarajan, Chelsey Hill, and Thomas Nolan. Text preprocessing. *Practical text analytics: Maximizing the value of text data*, pages 45–59, 2019.

Janet E Anderson, Naonori Kodate, Rhiannon Walters, and Anneliese Dodds. Can incident reporting improve safety? healthcare practitioners' views of the effectiveness of incident reporting. *International journal for quality in health care*, 25(2):141–150, 2013.

Ahmed Askar and Andreas Züfle. Clustering adverse events of covid-19 vaccines across the united states. In *Similarity Search and Applications: 14th International Conference, SISAP 2021, Dortmund, Germany, September 29–October 1, 2021, Proceedings 14*, pages 307–320. Springer, 2021a.

Ahmed Askar and Andreas Züfle. Clustering of adverse events of post-market approved drugs. In *17th International Symposium on Spatial and Temporal Databases*, pages 106–115, 2021b.

G Ross Baker, Peter G Norton, Virginia Flintoft, Régis Blais, Adalsteinn Brown, Jafna Cox, Ed Etchells, William A Ghali, Philip Hébert, Sumit R Majumdar, et al. The canadian adverse events study: the incidence of adverse events among hospital patients in canada. *Cmaj*, 170(11):1678–1686, 2004.

Atreya Basu, Christine Walters, and M Shepherd. Support vector machines for text categorization. In *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*, pages 7–pp. IEEE, 2003.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

Troyen A Brennan, Lucian L Leape, Nan M Laird, Liesi Hebert, A Russell Localio, Ann G Lawthers, Joseph P Newhouse, Paul C Weiler, and Howard H Hiatt. Incidence of adverse events and negligence in hospitalized patients: results of the harvard medical practice study i. *New England journal of medicine*, 324(6):370–376, 1991.

Nitesh V Chawla. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, pages 875–886, 2010.

Hercules Dalianis and Hercules Dalianis. Characteristics of patient records and clinical corpora. *Clinical Text Mining: Secondary Use of Electronic Patient Records*, pages 21–34, 2018.

Eefje N De Vries, Maya A Ramrattan, Susanne M Smorenburg, Dirk J Gouma, and Marja A Boermeester. The incidence and nature of in-hospital adverse events: a systematic review. *BMJ Quality & Safety*, 17(3):216–223, 2008.

Xuelian Deng, Yuqing Li, Jian Weng, and Jilian Zhang. Feature selection for text classification: A review. *Multimedia Tools and Applications*, 78:3797–3816, 2019.

Elisa Dolci, Barbara Schärer, Nicole Grossmann, Sarah Naima Musy, Franziska Zúñiga, Stefanie Bachnick, Michael Simon, et al. Automated fall detection algorithm with global trigger tool, incident reports, manual chart review, and patient-reported falls: algorithm development and validation with a retrospective diagnostic accuracy study. *Journal of medical Internet research*, 22(9): e19516, 2020.

Liam Donaldson. An organisation with a memory. *Clinical medicine*, 2(5):452, 2002.

Molla S Donaldson, Janet M Corrigan, Linda T Kohn, et al. To err is human: building a safer health system. 2000.

Persephone Doupi, Helge Svaar, Brian Bjørn, Ellen Deilkås, Urban Nylén, and Hans Rutberg. Use of the global trigger tool in patient safety improvement efforts: Nordic experiences. *Cognition, Technology & Work*, 17:45–54, 2015.

Soufiane El Mrabti, Mohammed Al Achhab, and Mohamed Lazaar. Comparison of feature selection methods for sentiment analysis. In *Big Data, Cloud and Applications: Third International Conference, BDCA 2018, Kenitra, Morocco, April 4–5, 2018, Revised Selected Papers 3*, pages 261–272. Springer, 2018.

Huw Prosser Evans, Athanasios Anastasiou, Adrian Edwards, Peter Hibbert, Meredith Makeham, Saturnino Luz, Aziz Sheikh, Liam Donaldson, and Andrew Carson-Stevens. Automated classification of primary care patient safety incident report content and severity using supervised machine learning (ml) approaches. *Health informatics journal*, 26(4):3123–3139, 2020.

Extend AS. Komplett kvalitetssystem - modulbasert, enkelt og fleksibelt. `https://www.extend.no/eqs/?gad=1&gclid=CjwKCAjwg-GjBhBnEiwAMUvNW_4QNpc3d1IzKc0yqX075atQK0ep9JhHU1dO3kmNtRrmxu8TWn5IfBoCH00QAvD_BwE`. Last Accessed: 07.06.2023.

Katsuhide Fujita, Masanori Akiyama, Keunsik Park, Etsuko (Nakagami) Yamaguchi, and Hiroyuki Furukawa. Linguistic analysis of large-scale medical incident reports for patient safety. In *MIE*, pages 250–254, 2012.

Lars Ulrik Gerdes and Christian Hardahl. Text mining electronic health records to identify hospital adverse events. *Studies in health technology and informatics*, 192:1145–1145, 2013.

Paul Gill, Kate Stewart, Elizabeth Treasure, and Barbara Chadwick. Methods of data collection in qualitative research: interviews and focus groups. *British dental journal*, 204(6):291–295, 2008.

Government of Norway. Regulation on leadership and quality improvement in health and care services. `https://lovdata.no/dokument/SF/forskrift/2016-10-28-1250`, 2020. Last accessed: 28.05.2023.

Government of Norway. Norwegian board of health supervision investigation commission. `https://lovdata.no/dokument/NL/lov/2017-06-16-56`, 2021. Last accessed: 28.05.2023.

Government of Norway. Act on specialist health services. `https://lovdata.no/dokument/NL/lov/1999-07-02-61`, 2023. Last accessed: 28.05.2023.

Qiong Gu, Li Zhu, and Zhihua Cai. Evaluation measures of the classification performance of imbalanced data sets. In *Computational Intelligence and Intelligent Systems: 4th International Symposium, ISICA 2009, Huangshi, China, October 23-25, 2009. Proceedings 4*, pages 461–471. Springer, 2009.

Steve R Gunn et al. Support vector machines for classification and regression. *ISIS technical report*, 14(1):5–16, 1998.

Emma Haddi, Xiaohui Liu, and Yong Shi. The role of text pre-processing in sentiment analysis. *Procedia computer science*, 17:26–32, 2013.

Frederick Hartwig and Brian E Dearing. *Exploratory data analysis*. Number 16. Sage, 1979.

Helse Nord-Trøndelag. Om oss - nøkkeltall helse nord-trøndelag. `https://hnt.no/om-oss#nokkeltall-helse-nord-trondelag`, 2022. Last Accessed: 29.05.2023.

Julio Hernandez, Jesús Ariel Carrasco-Ochoa, and José Francisco Martínez-Trinidad. An empirical study of oversampling and undersampling for instance selection methods on imbalance datasets. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I 18*, pages 262–269. Springer, 2013.

Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

Se Young Jung, Hee Hwang, Keehyuck Lee, Ho-Young Lee, Eunhye Kim, Miyoung Kim, and In Young Cho. Barriers and facilitators to implementation of medication decision support systems in electronic medical records: Mixed methods approach based on structural equation modeling and qualitative analysis. *JMIR Medical Informatics*, 8(7):e18758, 2020.

Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan. A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1):4–20, 2010.

Stefan Knerr, Léon Personnaz, and Gérard Dreyfus. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing: algorithms, architectures and applications*, pages 41–50. Springer, 1990.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019.

Tambiama André Madiega. Eu guidelines on ethics in artificial intelligence: Context and implementation. 2019.

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Madison, WI, 1998.

Scott D McKnight. Semi-supervised classification of patient safety event reports. *Journal of patient safety*, 8(2):60–64, 2012.

Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE, 2020.

Ajinkya More. Survey of resampling techniques for improving classification performance in unbalanced datasets. *arXiv preprint arXiv:1608.06048*, 2016.

Norwegian Board of Health Supervision. Statistics about severe adverse events reported to the norwegian board of health supervision. `https://www.helsetilsynet.no/globalassets/opplastinger/tilsyn/varsel_enhet/statistikk_varselordning.pdf`. Last accessed: 28.05.2023.

Norwegian Board of Health Supervision. Tasks of the norwegian board of health supervision. `https://helsetilsynet.no/om-oss/oppgaver-organisering-statens-helsetilsyn/oppgavene-statens-helsetilsyn/`, 2023. Last accessed: 28.05.2023.

NTNU. Hunt cloud. `https://www.ntnu.edu/mh/huntcloud`. Last Accessed: 29.05.2023.

Mei-Sing Ong, Farah Magrabi, and Enrico Coiera. Automated categorisation of clinical incident reports using statistical text classification. *Quality and Safety in Health Care*, 19(6):e55–e55, 2010.

Muhammad Salman Pathan, Avishek Nag, Muhammad Mohisn Pathan, and Soumyabrata Dev. Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, 2:100060, 2022.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Natasha Rafter, Anne Hickey, Sarah Condell, Ronán Conroy, P O'connor, D Vaughan, and David Williams. Adverse events in healthcare: learning from mistakes. *QJM: An International Journal of Medicine*, 108(4):273–277, 2015.

Deepa Rani, Rajeev Kumar, and Naveen Chauhan. Study and comparision of vectorization techniques used in text classification. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2022.

Sebastian Raschka. Naive bayes and text classification i-introduction and theory. *arXiv preprint arXiv:1410.5329*, 2014.

Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 616–623, 2003.

Thomas Brox Røst, Christine Raaen Tvedt, Haldor Husby, Ingrid Andås Berg, and Øystein Nytrø. Capturing central venous catheterization events in health record texts. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 488–495. IEEE, 2018.

Eli Saastad, Tor-Arne Engebretsen, Øystein Flesland, Kristin Klem, Sverre Levernes, Namik Resulbegovic, Thomas J Riiser, Sissel Rudsro, Kristin Sandby, Frode Strømman, et al. Nasjonalt klassifikasjonssystem–sluttrapport fra prosjektet «felles nasjonalt klassifikasjonssystem for uønskede hendelser». 2015.

Jessica M Schwartz, Maureen George, Sarah Collins Rossetti, Patricia C Dykes, Simon R Minshall, Eugene Lucas, and Kenrick D Cato. Factors influencing clinician trust in predictive clinical decision support systems for in-hospital deterioration: Qualitative descriptive study. *JMIR Human Factors*, 9(2):e33960, 2022.

Foram P Shah and Vibha Patel. A review on feature selection and feature extraction for text classification. In *2016 international conference on wireless*

communications, signal processing and networking (WiSPNET), pages 2264–2268. IEEE, 2016.

Catarina Silva and Bernardete Ribeiro. The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 1661–1666. IEEE, 2003.

Gurinder Singh, Bhawna Kumar, Loveleen Gaur, and Akriti Tyagi. Comparison between multinomial and bernoulli naïve bayes for text classification. In *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, pages 593–596. IEEE, 2019.

Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19*, pages 1015–1021. Springer, 2006.

St. Olavs Hospital. Nøkkeltall for st. olavs hospital. `https://stolav.no/om-oss/nokkeltall-for-st-olavs-hospital`, 2021. Last Accessed: 29.05.2023.

David C Stockwell, Eric Kirkendall, Stephen E Muething, Elizabeth Kloppenborg, Hima Vinodrao, and Brian R Jacobs. Automated adverse event detection collaborative. *Journal of patient safety*, 9(4):203–210, 2013.

The national research ethics committees. Regional committee for medical and health research statistics. `https://www.forskningsetikk.no/om-oss/komiteer-og-utvalg/rek/`, 2014. Last Accessed: 23.05.2023.

Alper Kursat Uysal and Serkan Gunal. The impact of preprocessing on text classification. *Information processing & management*, 50(1):104–112, 2014.

Suchi Vora and Hui Yang. A comprehensive study of eleven feature selection algorithms and their impact on text classification. In *2017 Computing Conference*, pages 440–449. IEEE, 2017.

Ž Vujović et al. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6):599–606, 2021.

Justin J Waring. A qualitative study of the intra-hospital variations in incident reporting. *International Journal for Quality in Health Care*, 16(5):347–352, 2004.

Michael Waskom, Olga Botvinnik, Drew O'Kane, Paul Hobson, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, Jordi Warmenhoven, Julian de Ruiter, Cameron Pye, Stephan Hoyer, Jake Vanderplas, Santi Villalba, Gero Kunter, Eric Quintero, Pete Bachant, Marcel Martin, Kyle Meyer, Alistair Miles, Yoav Ram, Tal Yarkoni, Mike Lee Williams, Constantine Evans, Clark Fitzgerald, Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), September 2017. URL `https://doi.org/10.5281/zenodo.883859`.

World Health Organization et al. World alliance for patient safety: Who draft guidelines for adverse event reporting and learning systems: from information to action. Technical report, World Health Organization, 2005.

World Health Organization et al. Patient safety incident reporting and learning systems: technical report and guidance. 2020.

Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *Icml*, volume 97, page 35. Citeseer, 1997.

Bei Yu. An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, 23(3):327–343, 2008.

Yongli Zhang. Support vector machine classification algorithm and its application. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3*, pages 179–186. Springer, 2012.

Appendices

A Norwegian stop words

The Norwegian stop words used in the implementation were extracted from the Python library Natural Language Toolkit [Bird et al., 2009]. Following is the full list of stop words:

[“og”, “i”, “jeg”, “det”, “at”, “en”, “et”, “den”, “til”, “er”, “som”, “på”, “de”, “med”, “han”, “av”, “ikke”, “ikkje”, “der”, “så”, “var”, “meg”, “seg”, “men”, “ett”, “har”, “om”, “vi”, “min”, “mitt”, “ha”, “hadde”, “hun”, “nå”, “over”, “da”, “ved”, “fra”, “du”, “ut”, “sin”, “dem”, “oss”, “opp”, “man”, “kan”, “hans”, “hvor”, “eller”, “hva”, “skal”, “selv”, “sjøl”, “her”, “alle”, “vil”, “bli”, “ble”, “blei”, “blitt”, “kunne”, “inn”, “når”, “være”, “kom”, “noen”, “noe”, “ville”, “dere”, “som”, “deres”, “kun”, “ja”, “etter”, “ned”, “skulle”, “denne”, “for”, “deg”, “si”, “sine”, “sitt”, “mot”, “å”, “meget”, “hvorfor”, “dette”, “disse”, “uten”, “hvordan”, “ingen”, “din”, “ditt”, “blir”, “samme”, “hvilken”, “hvilke”, “sånn”, “inni”, “mellom”, “vår”, “hver”, “hvem”, “vors”, “hvis”, “både”, “bare”, “enn”, “fordi”, “før”, “mange”, “også”, “slik”, “vært”, “være”, “båe”, “begge”, “siden”, “dykk”, “dykkar”, “dei”, “deira”, “deires”, “deim”, “di”, “då”, “eg”, “ein”, “eit”, “eitt”, “elles”, “honom”, “hjå”, “ho”, “hoe”, “henne”, “hennar”, “hennes”, “hoss”, “hossen”, “ikkje”, “ingi”, “inkje”, “korleis”, “korso”, “kva”, “kvar”, “kvarhelst”, “kven”, “kvi”, “kvifor”, “me”, “medan”, “mi”, “mine”, “mykje”, “no”, “nokon”, “noka”, “nokor”, “noko”, “nokre”, “si”, “sia”, “sidan”, “so”, “somt”, “somme”, “um”, “upp”, “vere”, “vore”, “verte”, “vort”, “varte” and “vart”]

B Clustering of Adverse Events

Text clustering is a NLP technique that involves grouping together similar documents or pieces of text into clusters or categories [Jain et al., 1999]. It is an unsupervised machine learning algorithm, which means it does not require any pre-labeled training data [Xu and Wunsch, 2005]. The goal of text clustering is to find patterns and relationships within a large corpus of text, which can then be used to classify and organize the data in a meaningful way. The resulting clusters can be analyzed and visualized to gain insights into the data, such as identifying topics of discussion or trends over time [Ahmed et al., 2022]. By grouping similar documents together, it can help identify patterns, trends, and themes that might not be immediately obvious from looking at individual documents.

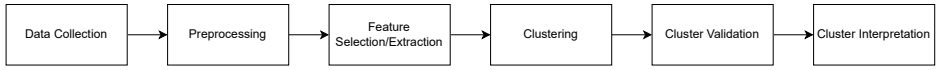


Figure 1: Text clustering steps

B.1 Experimental Plan

The clustering of adverse events was the initial focus of the machine learning study before obtaining access to the dataset. To address this specific focus, an experimental plan was developed. This subsection provides a brief overview of the original experimental plan and emphasizes the reasons behind each step rather than describing the detailed procedures.

1. **Data collection**. The dataset needs to be extracted and relevant columns and rows need to be identified.
2. **Data cleaning and text Processing**. Before clustering, it is important to prepare the data by removing any rows with invalid values, such as *Null* values, and preprocessing the text. Lowercase conversion, removing special characters and numbers, stopword removal, tokenization, and stemming are all preprocessing steps that can help reduce the dimensionality and improve the quality of the resulting clusters.
3. **Feature selection and extraction** Feature selection and extraction can be crucial for the effectiveness of a clustering method [Xu and Wunsch, 2005]. Feature selection can decrease the dimensionality and workload, simplifying the clustering process. To represent each document as a vector of features, a feature extraction method needs to be selected and performed. A good method can contribute to a simple and easily understood clustering [Jain et al., 1999].

4. **Clustering**. Text clustering is a complex process, and there is no single algorithm that is guaranteed to perform best for all problems or datasets [Xu and Wunsch, 2005]. The selection of a clustering algorithm is therefore crucial as it has a significant impact on the resulting clusters. This step will also include the selection of the similarity measure. This measure is used to determine how similar or dissimilar two documents are to each other [Jain et al., 1999]. The goal of the clustering algorithm is to create coherent and meaningful clusters that are also clearly distinct [Ahmed et al., 2022].

5. **Evaluation and interpretation**. Cluster validation is essential as clustering is subjective in nature and the same dataset can produce different clusters depending on the algorithm and use case [Jain et al., 1999]. There are three different methods to objectively assess a clustering algorithm's performance: external, internal, and relative [Xu and Wunsch, 2005]. This project would have employed an internal and relative test, the internal test as it examines the clustering structure directly from the dataset, and a relative test to compare algorithms. The main goal of clustering is to find clusters that provide meaningful insights about the data to the users [Xu and Wunsch, 2005]. To determine if the clusters provided are significant and useful, clinicians would be used as experts to analyze and discuss the results. The topic and trends each cluster would represent should also be identified.

Refining the experimental plan or repeating the process with different parameters may be required to enhance the results and gain a better understanding of the techniques and data. It is essential to adopt an iterative approach that allows for improvements as needed.

C Summary of research interviews

This appendix provides summaries of the research interviews conducted.

C.1 First Interview with Corporate Governance at St. Olav's Hospital

During this interview, we gained insights into CG-1&2's daily tasks associated with adverse events, and their role in the workflow of adverse event processing.

The interviewees clarified that their role does not involve active engagement in the processing of adverse event reports. Rather, their primary responsibility is to monitor the status of adverse events within the hospital and identify patterns that may span across various clinics and departments. By closely observing the reported incidents, they can detect recurring themes or emerging trends, allowing them to raise awareness and initiate discussions among relevant stakeholders. Their involvement lies in the proactive surveillance and identification of potential systemic issues, aiming to facilitate improvements in patient safety and quality of care across the hospital as a whole.

It was revealed that they manually review all adverse event reports, with particular emphasis on the free-text fields, as they contain the most details. They especially pay close attention to the reporter's description and the manager's assessment of the incident. In their risk and management system, EQS, they have the capability to filter, categorize, and search through the reports. This allows them to filter by specific departments or severity levels, and they can generate reports and graphs based on the available data. Furthermore, in special cases, they utilize information from the patient journal during their analysis.

Previously, they employed a dashboard that presented aggregated statistics, such as the count of different types of adverse events within a given timeframe. However, they discontinued using this system as they preferred examining the information documented in the reports rather than relying solely on aggregated data. They emphasized the importance of thoroughly reviewing all reports to obtain a comprehensive understanding of the current state of adverse events within the hospital.

C.2 Interview with Quality Advisor at St. Olav's Hospital

After our conversation with the Department of Corporate Governance, we aimed to gain insights from another perspective in the context of adverse event reporting. This led us to engage in a conversation with QA-1, who holds a distinct role

in this process. QA-1 has access to all adverse event reports related to the two clinics under their responsibility.

During the interview, QA-1 revealed that they did not utilize EQS for report analysis. Instead, they exported the data to Excel and leveraged pivot tables to thoroughly explore the reports and uncover potential trends. Among their analytical approaches, QA-1 emphasized the identification of recurring words as a crucial aspect of their analysis.

As part of their role as quality advisors, they attend Health, Safety, and Quality meetings every month, where they discuss and present patient incidents. The cases presented in the meetings require special attention, either due to their severe consequences or their perceived significance in other aspects. The objective of these discussions is to identify and propose interventions that can facilitate desired improvements.

In addition to exploring QA-1's involvement in the processing of adverse event reports, our conversation delved into the field of NOKUP. We learned that QA-1 had actively participated in the creation of NOKUP. We discussed several drawbacks of the classification system, such as disparities in category weighting and the varying number of subcategories. Moreover, they acknowledged that different healthcare professionals might interpret the categories differently based on their respective backgrounds and expertise. For instance, physicians may perceive certain incidents as treatment-related, while pharmacists may view them as drug-related. Furthermore, it was noted that not all managers had extensively studied the NOKUP categories, primarily due to their demanding schedules and time constraints. Given this inherent variability in the interpretation of NOKUP categories, QA-1 emphasized the importance of prioritizing the information conveyed through the free-text fields in the reports during analysis.

Through our interview with QA-1, we gained valuable insights into the role of a quality advisor and engaged in thought-provoking discussions regarding the intricacies of the NOKUP categories. This interview contributed to our enhanced understanding of the multifaceted nature of adverse event reporting and provided valuable perspectives to guide our subsequent research endeavors. It also confirmed the interest among clinicians in utilizing clustering to look for new patterns and categories compared to today's system using NOKUP.

C.3 Interview with Quality Advisor at Helse Nord-Trøndelag (HNT)

QA-2 works as a quality advisor at HNT, which operates the hospitals in Levanger and Namsos. The number of employees at HNT is approximately one-third that

of St. Olav's Hospital[St. Olavs Hospital, 2021; Helse Nord-Trøndelag, 2022]. By engaging with quality advisors from diverse healthcare settings, we aimed to explore different approaches and practices in evaluating and interpreting these reports. This interview thus provided us with additional perspectives on the analysis of adverse event reports, complementing the insights gained from previous interviews. This interview expands our understanding of the role of quality advisors and sheds light on variations in adverse event analysis across healthcare organizations.

QA-2's role does not involve direct handling of adverse event reports. Instead, they provide training to other employees on how to use EQS and Power BI. QA-2 primarily focuses on analyzing the numerical data rather than interpreting its meaning, which is delegated to others. For instance, clinicians may approach QA-2 to obtain statistics on trends, and QA-2 would then use EQS to generate the required statistics.

At HNT, they utilize a dashboard alongside EQS. This dashboard serves a similar purpose to the one previously employed at St. Olav's Hospital. It leverages the information from adverse event reports to present aggregated data, such as the count of incidents categorized as Infections, or the distribution of contributing factors across different reports. The dashboard offers filtering capabilities based on severity level, event location, report status, and contributing factors. The NOKUP categories provided by the reporter's manager are utilized to generate the graphs, representing the event type, location, and contributing factors. As the dashboard only includes aggregated categorical data and no patient identifiers, the dashboard can not be used to analyze the free-text fields of the reports.

During the interview, a significant portion of the discussion revolved around the NOKUP categories and their importance in the analysis of adverse events. It became evident that the categorization performed is not always accurate. This inconsistency in categorization poses challenges in accurately capturing and interpreting adverse events. It was mentioned that reporters often report patient-related incidents as non-patient-related, which leads to deficiencies in the report as it is usually not corrected by the managers.

The key takeaways from this interview were the reoccurring focus on the incorrect use of the NOKUP categories and the lack of usability in the systems applied. After this interview, the idea of using machine learning to categorize adverse events according to the NOKUP categories became more prominent. In addition, we thought of the idea of using machine learning to summarize the reports in order to make processing and comparing them less resource-demanding.

C.4 The second interview with Corporate Governance at St. Olav's Hospital

This interview took place about 4 months after the last interview and in the meantime, access to the dataset had been granted. Time had been spent understanding the contents of the dataset, starting the EDA, as well as the clustering experiment.

One of the main goals of this interview was to discuss our current idea, which was clustering adverse events, and how future work could benefit from this clustering experiment. However, when we were granted access to the dataset and realized we had access to the NOKUP categories used we started to consider if classifying adverse events using the NOKUP categories could be possible, and potentially more valuable. We thus decided to use this interview to clarify the workflow regarding adverse event processing and discuss which application of machine learning they found the most valuable and interesting.

When presenting our potential use cases, they clearly stated that using NOKUP categories and supervised machine learning to classify adverse events would provide more value to them than using unsupervised machine learning to potentially generate new categories of adverse events. The idea of summarizing the reports where as expected not as valuable to them as they have the resources to read the raw text, and a preference for this in order to obtain a comprehensive understanding of the incident and the reporter's experience.

According to the interviewees, cluster analysis can potentially be useful for identifying patterns and types of incidents at local levels, such as within a clinic. They believe that the number of adverse events at these levels is generally low, and thus extensive data-driven methods are not necessary for humans to recognize the patterns. However, when dealing with larger datasets, like at the hospital level, the interviewees stated that cluster analysis is not considered relevant. In such cases, the focus is primarily on identifying incidents that need to be reported in accordance with legal requirements, or requests made by the media, and a more standardized categorization would thus be more valuable.

In order to get a better understanding of which information in the dataset could be useful in order to detect similarities between adverse events we discussed how they use the information provided by the adverse event reports when monitoring and analyzing adverse events. The importance of the free text fields was once again emphasized. The description of the event, the title, and the description of the cause were emphasized. Based on this feedback we concluded to use the description and title as input to our model.

Key takeaways from the interview were that they found the classification of ad-

verse events using NOKUP more interesting than the use of unsupervised clustering to detect new patterns and categories, or the use of text summarization to make the reports easier to process. In addition, we got a greater understanding of how they work, and how their work is influenced by the Regulation on Leadership and Quality Improvement in Health and Care Services presented in Section 2.2.

C.5 Group Feedback

We presented our current understanding of the workflow of adverse events, the three options of classification, clustering, and text summarization, as well as our conclusion to go for classification. The feedback from the clinicians was in general positive. They agreed that the categorization of adverse events using the NOKUP is often inconsistent and that automatization of this would benefit several levels of the adverse event processing.

D Results from Experiment 3

This appendix includes the macro F1-score achieved by MNB and SVM when employing all the different combinations of preprocessing, feature extraction, and feature selection methods. The preprocessing methods techniques are stopword removal and stemming. The feature extraction methods include TF-IDF and BOW in combination with unigram or unigram and bigram. The feature selection methods are CHI2, the ANOVA F-value and MI. The combinations also encompass combinations without preprocessing and feature selection methods, but feature extraction is always included as it serves a dual purpose as the document representation method.

The F1-score with the highest value for both classification models is highlighted in bold and marked with a yellow background.

Multinomial Naïve Bayes (MNB)

Preprocessing Techniques		Feature Extraction Methods		Feature Selection Methods	macro F1-score
No Stopword Removal	No Stemming	TF-IDF	Unigram	No Feature Selection	0.68366234569324
				CHI2	0.6882154792271551
				ANOVA F-value	0.6881901967897156
				MI	
			Unigram and Bigram	No Feature Selection	0.6685425227002255
				CHI2	0.6749105281413049
				ANOVA F-value	0.6751718451088976
				MI	
		BOW	Unigram	No Feature Selection	0.7057359067785994
				CHI2	0.7149852676349955
				ANOVA F-value	0.7180536491866588
				MI	**0.7182172034231848**
			Unigram and Bigram	No Feature Selection	0.6329521163992451
				CHI2	0.6469151067744459
				ANOVA F-value	0.6499478236834262
				MI	0.6556162332031954
	Stemming	TF-IDF	Unigram	No Feature Selection	0.6740904241376109
				CHI2	0.6786068376339369
				ANOVA F-value	0.6791319697110968
				MI	
			Unigram and Bigram	No Feature Selection	0.6708481592017274
				CHI2	0.6739699589308817
				ANOVA F-value	0.6750067897015342
				MI	
		BOW	Unigram	No Feature Selection	0.6685299448862084
				CHI2	0.6789308033556003
				ANOVA F-value	0.6764445094341871
				MI	0.6785569045807817
			Unigram and Bigram	No Feature Selection	0.49739381025128554
				CHI2	0.5026507697021032
				ANOVA F-value	0.5040803310063594
				MI	0.504769368642694
Stopword Removal	No Stemming	TF-IDF	Unigram	No Feature Selection	0.6892461752526857
				CHI2	0.6900463707867275
				ANOVA F-value	0.6902892869388216
				MI	
			Unigram and Bigram	No Feature Selection	0.6856257229504491
				CHI2	0.6891322953255494
				ANOVA F-value	0.690631804378485
				MI	
		BOW	Unigram	No Feature Selection	0.7099492584254676
				CHI2	0.7155368027606233
				ANOVA F-value	0.7150990827318416
				MI	0.7148042380522537
			Unigram and Bigram	No Feature Selection	0.6567182565583407
				CHI2	0.6596016510206131
				ANOVA F-value	0.6589857670723566
				MI	0.6648335643906826
	Stemming	TF-IDF	Unigram	No Feature Selection	0.6794453826612562
				CHI2	0.6820816345697349
				ANOVA F-value	0.6826204940503005
				MI	
			Unigram and Bigram	No Feature Selection	0.6702953905797422
				CHI2	0.6760467261183889
				ANOVA F-value	0.6752195903357014
				MI	
		BOW	Unigram	No Feature Selection	0.6602393536596652
				CHI2	0.6758069362546324
				ANOVA F-value	0.6772428036785498
				MI	0.6789129434963109
			Unigram and Bigram	No Feature Selection	0.5396761346382072
				CHI2	0.5466125593794171
				ANOVA F-value	0.547948239023425
				MI	0.5514302508471405

Support Vector Machine (SVM)

Preprocessing Techniques		Feature Extraction Methods		Feature Selection Methods	macro F1-score
No Stopword Removal	No Stemming	TF-IDF	Unigram	No Feature Selection	0.70902047049966
				CHI2	0.7076825101184864
				ANOVA F-value	0.709020900049296
				MI	
			Unigram and Bigram	No Feature Selection	0.6698832785155001
				CHI2	0.6553255162474572
				ANOVA F-value	0.6052943674406175
				MI	
		BOW	Unigram	No Feature Selection	0.6756953270583155
				CHI2	0.6774739560268737
				ANOVA F-value	0.6778455969432986
				MI	0.6770747782969838
			Unigram and Bigram	No Feature Selection	0.6693622095596673
				CHI2	0.6650080938298857
				ANOVA F-value	0.664791801431137
				MI	0.6649834726666365
	Stemming	TF-IDF	Unigram	No Feature Selection	**0.7164708990301282**
				CHI2	0.7159210241320567
				ANOVA F-value	0.7122048719038252
				MI	
			Unigram and Bigram	No Feature Selection	0.6518434462379
				CHI2	0.6283833492176618
				ANOVA F-value	0.561037927241807
				MI	
		BOW	Unigram	No Feature Selection	0.6794328186804773
				CHI2	0.6816507441199011
				ANOVA F-value	0.6816507441199011
				MI	0.6814117768651007
			Unigram and Bigram	No Feature Selection	0.6700532327625368
				CHI2	0.643932905667035
				ANOVA F-value	0.643932905667035
				MI	0.6535183553855567
Stopword Removal	No Stemming	TF-IDF	Unigram	No Feature Selection	0.7092693725152116
				CHI2	0.7064656425269008
				ANOVA F-value	0.6970008832568676
				MI	
			Unigram and Bigram	No Feature Selection	0.6800792486321277
				CHI2	0.6767433933601411
				ANOVA F-value	0.5999699249396528
				MI	
		BOW	Unigram	No Feature Selection	0.6851755879676223
				CHI2	0.6851835012472816
				ANOVA F-value	0.6864977742531295
				MI	0.6852914779200582
			Unigram and Bigram	No Feature Selection	0.6835791578295791
				CHI2	0.6821526707242479
				ANOVA F-value	0.6814211688326814
				MI	0.6809140019694147
	Stemming	TF-IDF	Unigram	No Feature Selection	0.7103436084287608
				CHI2	0.7109124080801666
				ANOVA F-value	0.7077497522088441
				MI	
			Unigram and Bigram	No Feature Selection	0.6681922732692473
				CHI2	0.6515416431842846
				ANOVA F-value	0.581444273714598
				MI	
		BOW	Unigram	No Feature Selection	0.6841395906633508
				CHI2	0.6842774578339537
				ANOVA F-value	0.680617626475531
				MI	0.6841395906633508
			Unigram and Bigram	No Feature Selection	0.680617626475531
				CHI2	0.6611519394665877
				ANOVA F-value	0.6611519394665877
				MI	0.6636104531448916

E Confusion Matrices

This appendix presents the unnormalized confusion matrices for MNB and SVM, providing insights into the absolute counts of predictions made by the models. They illustrate the distribution of predicted labels compared to the true labels. While these matrices were not directly referred to in the discussion, they offer supplementary information regarding the distribution of predictions and the presence of class imbalances. It is important to note that the color distribution in the matrices may present a misleading picture of performance due to class imbalances. To gain a more accurate understanding of the models' performance, it is necessary to consider the number of reports for each category. The confusion matrix for MNB is shown in Figure 2, while Figure 3 displays the confusion matrix for SVM.

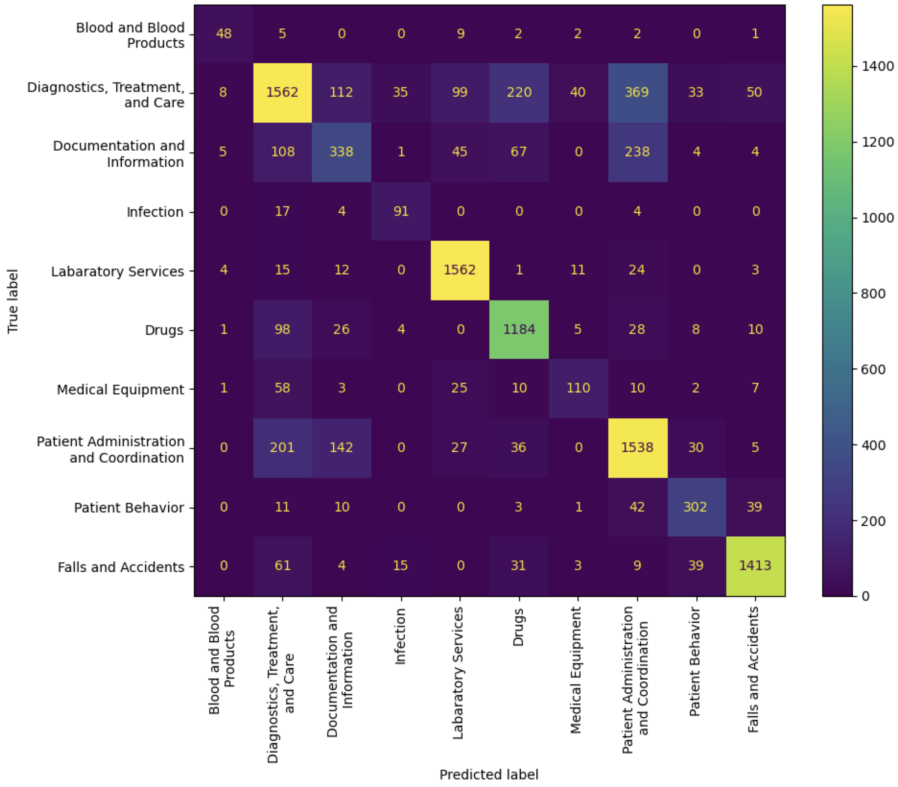


Figure 2: Confusion matrix for the MNB model

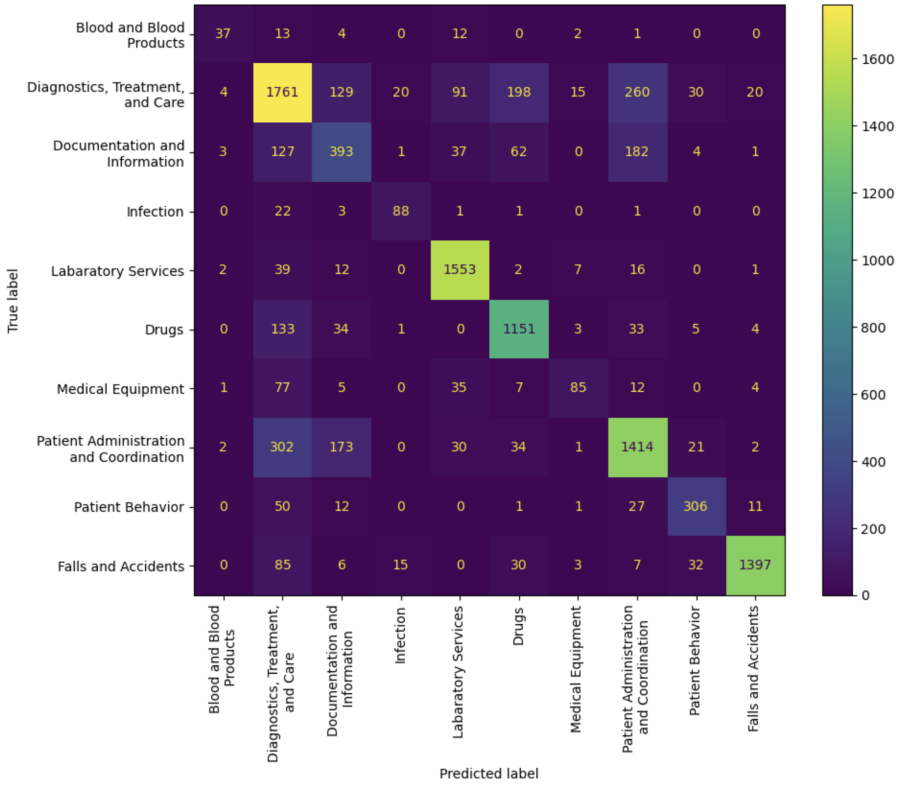
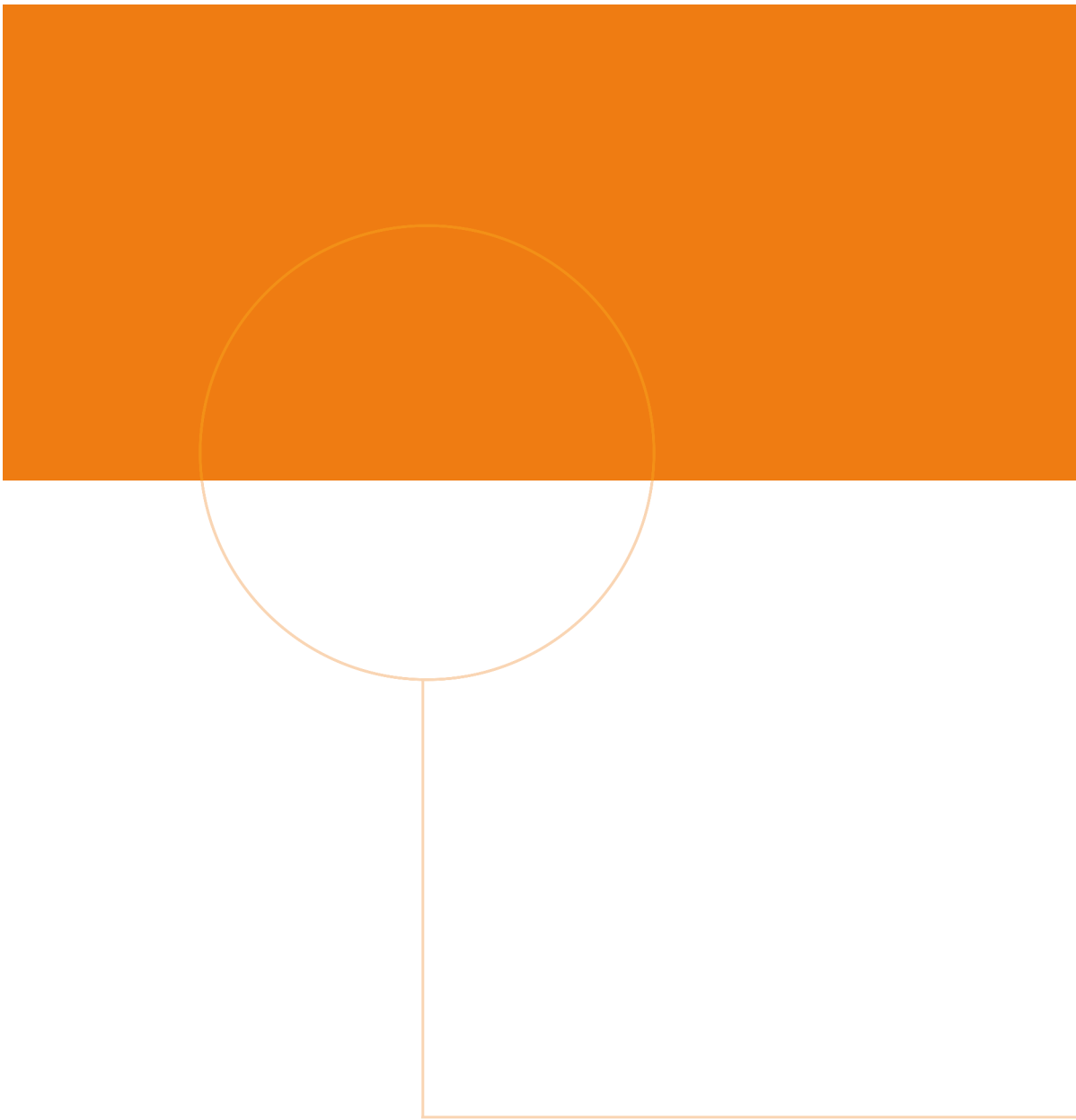


Figure 3: Confusion matrix for the SVM model



NTNU

Norwegian University of
Science and Technology