Text Classification Models in Social Media Channels owned by Pharma: Early Detection of Compliance-Relevant Information

Master Thesis submitted in fulfilment of the Degree

Master of Business Administration

MBA Specialization: Digital Marketing

Submitted to Dr Daniel Dan

Verena Foppa, MSc

00517372

Vienna, 16.02.2023

AFFIDAVIT

I hereby affirm that this Master's Thesis represents my own written work and that I have used no sources and aids other than those indicated. All passages quoted from publications or paraphrased from these sources are properly cited and attributed.

The thesis was not submitted in the same or in a substantially similar version, not even partially, to another examination board and was not published elsewhere.

01.02.2023	
Date	Signature

ABSTRACT

Nowadays, social networks and social media are all the rage. The growth of electronic communication via social media increases the necessity of reshaping the healthcare industry and the pharmaceutical sector. Thus, their marketing techniques are also in need of adaptation. Social media has always been observed with a critical eye by the Pharmaceutical Industry. Due to restricting regulations and hence limited derived benefits, up to recent years, only the big players have moved into this field. Although global events have boosted digital transformation in Pharma enhancing the need to leverage digital channels for communication and interaction with customers, many companies are reluctant to venture into social. The primary concern is the complexity of processing and archiving incoming compliance-relevant issues in a timely and standardised operating procedure-conform manner. Compliance, which includes queries regarding medical, regulatory, and patient information, pharmacovigilance as well as product complaints, requires up to 24/7 monitoring. This results either in companies paying agencies for monitoring or in limiting, and even totally disabling the chat function.

This thesis aims to tackle the use of text classification algorithms to automate the monitoring of pharma-owned social media channels, especially for compliance-relevant information. Machine learning algorithms in combination with natural language processing techniques are used to classify textual data into predefined classes using a supervised learning approach. Overall, the performance of three proposed algorithms (i.e., Linear Discriminant Analysis, k-Nearest Neighbour and fastText) is consistent with available research evidence, thus validating the project's findings and demonstrating the viability of a scaled implementation in pharma-owned social media channels on a national and international scale. Overall better performance has been observed in fastText, which is why it was selected for continued project development among the three models examined, despite the good performance of the other two given the rather small data set. This proof-of-concept study serves the purpose of highlighting the feasibility, importance, and impact of the integration of social media as a pull marketing instrument into omnichannel strategies of pharmaceutical companies, for increased compliance and drug safety monitoring as well as increased insight generation and customer-centricity.

ACKNOWLEDGEMENTS

I would like to acknowledge and extend my sincere gratitude to my supervisor, Dr Daniel Dan, Professor at the School for Applied Data Science of the private international Modul University of Vienna, who made this project possible. His advice and guidance through all the stages of my thesis were crucial to gain a well-rounded insight into the topic, working with the different methodologies and interpreting and presenting the results as clearly as possible. In addition, he has provided time and space for trial and error as well as patience as I familiarised myself with the subject.

Additionally, I want to express my gratitude to my employer and friend Dr Susanne Harzer and her family for their unwavering support and patience as I conducted my research and wrote my project. I would like to thank them for their commitment and the measures taken to enable me to acquaint myself with the necessary tools for my thesis even within my working environment.

Further, I would like to thank my work colleagues and friends, who volunteered in helping me, especially during the preparatory phases of this project (focus groups).

I also present my utmost appreciation to my MBA colleagues, an eclectic group of people I can call my peer group, whose motivation and encouragement have always kept me driven to pursue progress and finalisation of my work.

Last but not least, I would like to thank my family for their support, kindness and patience always being present during the time of my studies and my thesis.

TABLE OF CONTENTS

LIST OF FIGURES

List of Abbreviations

ADR	Adverse Drug Reaction
AE	Adverse Event
AGES	Agentur für Gesundheit und Ernährungssicherheit
AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
ASR	Automatic Speech Recognition
B2B	Business to Business
B2C	Business to Customer
BO	Bowel Open/ Bowel Obstruction
cBOW	continuous Bag of Words
CFG	Context-Free Grammar
CLD	Chronic Liver Disease/Chronic Lung Disease
CNN	Convolutional Neural Network
COVID-19	Corona Virus Disease 2019
CRAN	Comprehensive R Archive Network
CRF	Conditional Random Fields
CRM	Customer Relationship Management
DL	Deep Learning
DNN	Deep Neural Network
DRL	Deep Reinforcement Learning
DTM	Document Term Matrix
EFPIA	European Federation of Pharmaceutical Industries and Associations
EHR	Electronic Health Record
EMA	European Medicine Agency
EMBL-EBI	European Molecular Biology Laboratory-European Bioinformatics Institute
EU	European Union
FAIR	Facebook AI Research Lab
FDA	Food and Drug Administration
FTP	File Transfer Protocol
GDPR	General Data Protection Regulation
GloVe	Global Vector
GMDH	Group Method of Data Handling
GPU	Graphics Processing Unit
HCP	Health Care Professional
IBM	International Business Machines Corporation
ICT	Information and Communication Technology
IDE	Integrated Development Environment
IDF	Inverse Document Frequency
IFPMA	International Federation of Pharmaceutical Manufacturers and Associations
IGEPHA	Interessensgemeinschaft österreichischer Heilmittelhersteller und Depositeure
IMI	Innovative Medicines Initiative
IoT	Internet of Things
IT	Information Technology
KDD	Knowledge Discovery in Databases

kNN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
LSTM	Long-Short-Term Memory
MedDRA	Medical Dictionary for Regulatory Activities
MHRA	Medicines and Healthcare Regulatory Agency
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MLBDD	Machine-Learning Based Disease Diagnostics
MT	Machine Translation
NB	Naïve Bayes
NIS	Non-Interventional Study
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
NN	Neural Network
OOV	Out-Of-Vocabulary
OPDP	Office of Prescription Drug Promotion
PASS	Post-Authorization Safety Study
PHARMIG	Verband der pharmazeutischen Industrie Österreichs
PII	Personally Identifiable Information
PV	Pharmacovigilance
QSAR	Quantitative Structure-Activity Relationship
R&D	Research and Development
RF	Random Forests
RL	Reinforcement Learning
RNN	Recurrent Neural Network
Rx	the known symbol for prescription drugs (Latin "R" for "recipere")
SEA	Search Engine Advertising
SEO	Search Engine Optimization
SL	Supervised Learning
SNS	Social Networking Sites
SOP	Standard Operating Procedure
SRL	Semantic Role Labelling
SSL	Semi-Supervised Learning
STR	Speech to Text Recognition
SVD	Support Vector Decomposition
SVM	Support Vector Machine
TC	Text Classification
TF	Term Frequency
TNR	True Negative Rate
TPA	True Positive Accuracy
TPR	True Positive Rate
TTS	Text To Speech
UGC	User-Generated Content
UNIX	UNiplexed Information Computing System
US	United States

USL	Unsupervised Learning
VSM	Vector Space Model
Web-RADR	Web-Recognizing Adverse Drug Reactions
WHO	Word Health Organization
Word2Vec	Word to Vector
WSD	Word Sense Disambiguation

1 INTRODUCTION

1.1 Context and Theoretical Framework

Nowadays, over 50% of the world's population are active social media users (Dixon, 2022), and depending on the geographical location, between 55% and 85% of them search for health-related content (Eurostat Website, 2021; Jia et al. 2021). This statistic alone reveals the enormous impact social media has on our lives. While social media channels were mainly used privately until a few years ago, today it is hard to imagine business online marketing without them. Facebook, Twitter, LinkedIn, Instagram, YouTube and many others, offer excellent opportunities as pull marketing channels to reach potential customers, interact with them and meet personal customer demands through targeted marketing. While many industrial sectors are already successfully active on social media, there is still considerable reluctance among pharmaceutical marketers to integrate social channels. Several underlying reasons contribute to the current situation: (i) First, health-related topics are fundamentally sensitive. Their dissemination to the public is restricted by law which contributes to a limited field of action and risk aversion. (ii) Another reason is the strict compliance guidelines imposed on Pharmaceuticals, not only by internal company processes, but also by authorities, health institutes, and stakeholder organisations. (iii) In addition, the European Medicines Act imposes several restrictions on the advertising of prescription medicines, also known as Rx medicines, which further complicates Rx-based marketing activities. The reluctance to engage in social media strategies in Europe is apparent. In a recent Austrian survey conducted in different pharmaceutical companies in 2021, 40% of participating digital and marketing managers deploy between 1-25% of their budget for social media marketing activities, and 60% of participants stated that analysis of social media activities would be just rudimentary or not performed at all. Further, 50% mentioned digital marketing skills being on or below average (Gerfertz-Schiefer, 2021).

In the past few years, global events have boosted digital transformation worldwide across all industries thereby forcing also pharmaceutical operations to adopt trending multi- and omni-channel marketing strategies. This shift has majorly been reinforced by the changing engagement preferences of HCPs with an increase in “digital natives,” who welcome virtual contacts. Nevertheless, due to the low level of “controllability,” the chat functions of Pharma-owned channels have either limited access, are deactivated altogether, or monitoring is outsourced to external partners and agencies for whom compliance monitoring is not part of the core competencies.

Artificial Intelligence (AI) along with Natural Language Processing (NLP) have made enormous advancements over the past decade, with products such as Facebook’s face recognition and automated friend tagging application, speech recognition systems (Alexa, Amazon Echo, Siri, etc.)

or navigation systems, all implementing NLP to identify and address user requests. In addition, sophisticated text mining applications as well as machine learning and neural network algorithms have also been successfully deployed in fields as diverse as medical research, risk management, and customer care, such as spam detection and product reviews, insurance, i.e., fraud detection, and contextual (targeted) advertising. The application of AI technologies, such as natural language processing and machine learning, has become even more imperative following the growth of textual big data.

1.2 Research Aims and Objectives

This thesis is a proof-of-concept project. The aim is to investigate the use of different text classification (machine learning) algorithms to automate the monitoring of Pharma-owned social media channels, for compliance-relevant information, which includes queries pertaining to areas, such as

- pharmacovigilance (PV), for all drug safety issues and off-label use, which need up to 24/7 monitoring,
- quality, for product complaints and,
- regulatory, for queries regarding market authorization status and market availability of medical compounds.

Further, relevant queries that shall be covered are medical-related queries regarding detailed scientific information about products and diseases, which need to be internally archived and presented to authorities in case of internal or external audits.

Lastly, marketing queries (e.g., about upcoming events, visit requests etc.) and patient information including sentiment (e.g., praise and firestorm), should be included in the research though not compliance-relevant (i.e., it does not need to be reported to regulatory bodies). These informations can provide important insights into patient preferences and feedback on marketing campaigns. The analysis of the performance of the algorithms should allow to give indications as to which methodology may be best suitable for such monitoring.

The purpose of this project is to highlight the importance of social media as a pull marketing strategy for Rx Pharma-marketing by demonstrating that already proven technology can be implemented to overcome monitoring issues around social media engagement for pharmaceutical companies.

1.3 Structure of Thesis

After this introductory chapter, this thesis is structured as follows. The second chapter, Literature Evidence, is divided into two parts. The first part includes an overview of the world of artificial intelligence, the introduction and historical overview of the major subfields of machine learning, deep learning, and natural language processing, as well as the introduction of different NLP techniques and industrial examples where these algorithms are successfully applied. The second part of the second chapter is dedicated to the pharmaceutical industry. First, the digitalisation process that has already taken place with the help of AI technology is described, followed by the sector of “Customer Relationship Management” where a need for action is still apparent. Here, deep insights into the challenges regarding the integration of social media into the omnichannel strategy are addressed and illustrated by examples, and the underlying reasons for reluctance are explained in detail.

The literature review is followed by the third chapter, Methodology, in which an introduction is given to the focus area of the project (text mining and text classification), before moving on to the project design. In this part, the origin and processing of the utilised data set are presented, as the programming language used, as well as the appl of different NLP technologies. Furthermore, theoretical introductions to the three selected algorithms, namely Linear Discriminant Analysis (LDA), k-Nearest Neighbour (kNN) and fastText are given, as well as an insight into different evaluation metrics that are crucial for algorithm performance validation.

In the next chapter, the results and findings of the three proposed algorithms are reviewed and discussed. A comparative analysis shall further clarify the interpretation of the results. The conclusion will present a rounded framework of the project and its timely relevance. Lastly, a statement about the contribution of present knowledge, the limitations of the study and the outlook for future research are given.

2 LITERATURE REVIEW

2.1 The Environment of Artificial Intelligence: A Brief Overview

***Can machines think?* – Alan Turing, 1950 (Turing, 1950)**

Over the past decades, several Artificial Intelligence (AI) researchers, data scientists, linguist experts and professors have attempted to mutually agree on a common definition of AI. However, despite all efforts, no such results have been achieved yet. Further complicating matters is the development of sub-fields of AI over the course of the last decades, such as machine learning, deep learning, neural networks and natural language processing, which are terms that are often used synonymously as their technologies can be utilised simultaneously but encompass different research areas. What most of the individual definitions seem to have in common, is that AI is defined as the capability of a digital computer system to execute tasks that are typically associated with cognitive functions unique to humans, such as the capacity to reason, find meaning, generalise, or adjust by learning from experience (Rapaport, 2020). Since their advent in the 1940s, digital computing systems have repeatedly demonstrated the capability to be programmed to perform particularly complex tasks, such as playing games with great proficiency or deducing mathematical principles (Mackenzie, 1995; Silver, 2018). Figure 1 depicts an overview of the relationships and distinctions of the sub-fields in the environment of artificial intelligence.

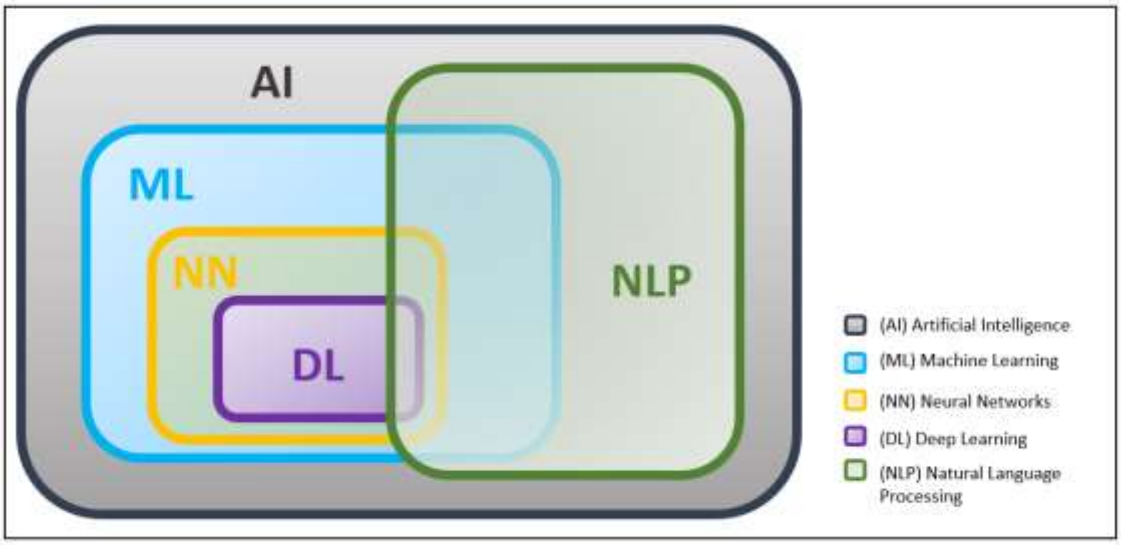


Figure 1: Visual representation of the environment of artificial intelligence. AI, Artificial Intelligence; ML, Machine Learning; DL, Deep Learning; NN, Neural Networks; NLP, Natural Language Processing; modified from Choi et al. 2020.

The key technology employed is known as Natural Language Processing (NLP), a branch of computational linguistics, computer science, and AI, and is conceived to fill the communication gap between human (natural) languages and computers. An application of AI known as Machine

Learning (ML) uses NLP techniques that can parse data, learn on their own without explicit programming, and then use what they have learned to make decisions. Nevertheless, despite ongoing improvements in computer memory and processing speed, programmes are still unable to emulate human adaptability in a wider set of tasks, or those requiring substantial everyday knowledge (Sathya, 2020).

Deep Learning (DL) is a part of machine learning, which is applied to larger data sets and is built upon Neural Networks (NN). More specifically, the number of hidden layers determines the differentiation between a neural network from a deep neural network (see section 2.4 for more details). Especially deep learning models have reached human expert-level performance in carrying out certain specific tasks. Artificial intelligence, in this limited sense, is now found in applications as diverse as medical diagnosis, computer search engines, and voice or handwriting recognition. (Houssein et al. 2021; Ni et al. 2022; Trattner et al. 2022).

2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a discipline that has gone through numerous stages throughout its history: first, rule-based, followed by the statistical revolution, and finally, the neural revolution (Chernyavskiy et al. 2021). As people recognised the importance of translation from one language to another, they hoped to create a machine that could do this sort of translation automatically. NLP, as a branch of artificial intelligence and linguistics, aims to research challenges associated with the autonomous comprehension of natural language, also called Natural Language Understanding (NLU) or Linguistics, and Natural Language Generation (NLG) by bridging the communication gap between computer and humans (Yadav, 2021).

While NLP is a hot topic of discussion today due to its many uses and recent advancements, the term wasn't existent until the late 1940s. Originating in the early ideas of Machine Translation (MT) and in the first conceptual approaches to create a “translating machine” in the mid-30ies, the first successful attempts were achieved during Second World War, with the deployment of Enigma, an encoding machine, utilised by Germans’ military to transmit secret messages about troops’ positions and strategies in combat (Johri et al. 2021). In 1950, Alan Turing, a British intelligence agent, who was majorly involved in cracking Enigma’s coding mechanism, published the paper “Computing Machinery and Intelligence” in the Journal “Mind”, in which he proposed an idea (an imitation game) for determining whether a computer would have thinking capabilities comparable to those of a human person. This notion became known as the “Turing Test” (Turing, 1950; Turing, 2012).

In a joint effort between International Business Machines Corporation (IBM) and Georgetown University, the “Georgetown experiment” took place in 1954 and featured the automatic translation of more than sixty Russian sentences into English, the two predominant languages at that time. Despite it being a small-scale experiment with only 250 words and six “grammar” rules, it raised hopes for the development of automatic translation systems with high quality in the near future (Garvin, 1967; Hutchins, 2004).

In the decades between the 50ies and 70ies, research and development of natural language processing systems were heavily influenced by the linguist Noam Chomsky and his work on structural language: When he first introduced syntactic structures in 1957, the field of NLP underwent a fundamental evolution. In its “Theory of Formal Syntax,” Chomsky refined a rule-based concept of language based on universal “context-free grammar” (CFG) and more restrictive “regular grammar” to specify text search patterns and create a well-defined, formal linguistics system (Chomsky & Schützenberger 1959): Because linguistics is directly embodied in rules and other forms of representation, it was formerly thought that machines might be trained to reason and operate similarly to the human brain, thus facilitating the automatic processing of natural languages.

The 1960s saw the development of some remarkably effective natural language processing systems, including Terry Winograd's SHRDLU (acronym of no particular significance) at the Massachusetts Institute of Technology (MIT). By utilizing phrase “blocks” inside a framework of a limited vocabulary, the system could be instructed in everyday speech, carrying out simple requests, such as “Can you put the red cone on top of the green block?”. Despite being inefficient at understanding complex as well as ambiguous situations, the system accurately recognised relations between objects in a natural language setting. The ELIZA program, a simulation created by Carl Rogers and written between 1964 and 1966 by Joseph Weizenbaum, delivered a strikingly human-like interaction between a psychologist and a patient, without having thinking or emotional capabilities. To patients’ statements going beyond the limited knowledge base, such as “My head hurts.”, ELIZA replied with a general question, such as “Why do you claim your head hurts?” (Weizenbaum, 1966). In retrospect, it is possible that ELIZA was the first conversational BOT predecessor being one of the early systems to pass the Turing test. (Joseph et al. 2016).

The 1970s saw further advancements as many programmers began developing “conceptual systems” that could translate input from the real world into data that computers could understand. In 1969 Roger Schank published the “Conceptual Dependency Theory” aimed at “illustrating how people think and process information” (Schank, 1972) with the help of tokenisation, i.e., the idea to divide each sentence into nominal (i.e., people or objects), action (i.e., activity, usually described by verbs), and modifier (i.e., quality or environment/location) constituents for a better understanding of a sentence’s meaning. A sentence is composed of one or more concepts, and for the full meaning to be understood, one concept may depend on another (Schumacher et al. 2012). Following tokenization, many different “parsing” approaches for syntactic analysis, for strings of symbols or text, its logical components, and their relationship to each other, were developed and successfully implemented up to recent years to overcome the communication gap between natural language and computers (described in more detail in section 1.4).

Standard parsing techniques that solely depended on hand-crafted rules, also called "symbolic NLP", faced two key issues due to the size, richness, unrestrictive nature, and ambiguity of natural language: (i) Most methods employed up to this point were based on CFG and restrictive syntactical structures. (ii) The computer's attempt to match and deduce sentence meaning was only achievable in the most ideal circumstances. In fact, early approaches for automated Russian to English translations were capped by language metaphors and homographs, i.e., words with the same spelling but different meanings (Nadkarni et al. 2011): Apparently the Biblical phrase "The spirit is willing, but the flesh is weak" was translated to "The vodka is good, but the meat is rotten" (Pollack, 1983).

By the mid and late 70ies, researchers began examining Chomskyan linguistic theories and found that many of them were not empirically correct, despite being compelling in writing. Further theoretical underpinnings, such as the scepticism regarding the usefulness of probabilistic models in NLP, prohibited the kind of linguistics corpus that later formed the foundation of the machine learning approach to natural language processing. The slow decline of Chomskyan theories on one hand, and the constant rise in computing power on the other, popularised by Moore's law stating that the speed and capability of computers will double every two years due to the advances in the number of microchip transistors, led researchers to converge more to empiricism and probabilistic models. The advent of the 80ies saw the introduction of machine learning-based techniques, which brought a major shift from a symbolic to a statistical and machine learning-based NLP (Johri et al. 2021).

While earlier machine learning algorithms, such as decision trees, generated rigid rule systems similar to existing handwritten rules, by the 90ies, probabilistic and statistical methods of handling natural language processing were the most common types of models.

To create models that are more resistant to unknown inputs (new terms or complex grammatical constructions) or wrong inputs, which is typical for real-world data, machine learning techniques can apply statistical inference algorithms (e.g., with omitted words or incorrectly spelled words). In contrast to rule-based (e.g., grammar) language processing systems, machine learning apprehends linguistic rules automatically by analysing large numbers of real-world data with the help of statistical inference. This allows increased accuracy by simply supplying more input data. Each rule in CFG has associated probabilities to their occurrence that are found using machine learning tools on designated corpora. As a result, a smaller, broader set of rules replaces a large number of specific rules, and statistical-frequency data is used to resolve ambiguities (Nadkarni et al. 2011).

2.3 Machine Learning (ML)

Machine Learning (ML) is a sub-field within the environment of artificial intelligence that focuses on the learning element of AI by constructing models that best represent a set of data. In contrast to traditional programming, where an algorithm may be explicitly implemented using known features, machine learning uses subsets of data as well as different combinations of features and weights for algorithm development (Choi et al 2020).

In principle, there are two categories into which machine learning models can be divided: generative and discriminative, as they reflect two different approaches to the probability of an object (e.g., text) falling into a given category. Generative models focus on the data set's pattern and distribution and simulate how data is distributed throughout the data set using Bayes' rule, according to which the probability of one or more events occurring simultaneously is based on known parameters related to the event(s). Discriminative models define boundaries in the data set and directly generate predictions on unseen data based on previously observed data. A generative model apprehends parameters by the maximisation of joint probability, i.e., that two events occur at the same time, whereas a discriminative machine learning trains a model by learning parameters that maximise the conditional probability, i.e., that one event occurs because another event has occurred. Both models are appropriate for particular applications due to their various machine learning strategies, with each, as with any method, implying advantages and disadvantages (Hsu & Griffiths 2010).

To better understand the different approaches, the first crucial initial step is to determine the type of data that need to be processed and the expected outcome. In the case of machine learning, labelled data and unlabelled are the two types of data that need to be dealt with. Although labelled data have both input and output parameters set in a completely machine-readable structure, classifying (labelling) the data initially takes a significant amount of human effort. On the contrary, for unlabelled data, none of the parameters need to be available in machine-readable form, which may eliminate the initial human work, but it compels the use of more complex solutions during the computational process (Nadkarni et al. 2011). The four most popular machine learning techniques are (i) supervised, (ii) unsupervised, (iii) semi-supervised, and (iv) reinforcement learning, which can each be used depending on the nature of the task at hand and the desired outcome.

2.3.1 Supervised Learning (SL)

Most discriminative models, also known as “conditional” models, are applied in supervised machine learning. Supervised Learning (SL) is the approach defined using labelled data, whose purpose is to “supervise” or “train” algorithms to correctly classify data or predict outcomes by learning the boundaries with the use of estimated and maximum probability. Already labelled input and output data enable the algorithm to monitor accuracy and improve performance over time. These models are not significantly impacted by outliers, and while they are often a better option than generative models, misclassification issues can be a significant disadvantage. For data mining, classification and regression are the two categories that apply to supervised learning. Classification problems deploy learning models that precisely classify test data into distinct categories, such as sorting lemons from oranges, Figure 2A.

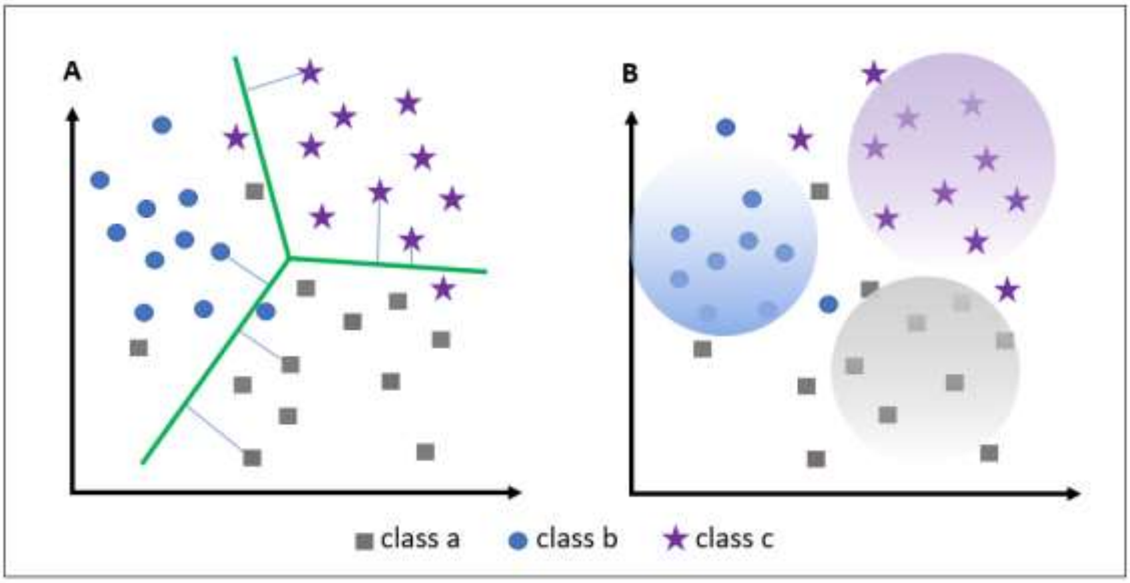


Figure 2: Schematic illustration of data subdivision with the use of A: Supervised Learning, and B: Unsupervised Learning; modified from Manning et al. 2008.

Supervised learning approaches can be used in the real world to separate and quarantine spam from relevant emails in a dedicated folder. Common classification techniques include Decision Trees (DT), naïve Bayes (NB), Random Forests (RF), Conditional Random Fields (CRF), linear classifiers, such as Linear Discriminant Analysis (LDA) and non-linear classifiers, such as k-Nearest Neighbour (kNN). Regression is another supervised learning algorithm which aims at comprehending the link between dependent and independent variables. Regression models are most useful when applied for forecasting budgets and sales revenue predictions which are based on several data. The most common regression algorithms include linear regression, logistic regression, and polynomial regression. Support Vector Machines (SVM) can be used in classification as well as regression problems for supervised learning (Choi et al. 2020; Jiachong, 2019).

2.3.2 Unsupervised Learning (USL)

In contrast to discriminative models, generative models, as their name suggests and previously mentioned, can originate new data. These models are typically, though not exclusively, such as LDA, applied to unsupervised machine learning problems. Unsupervised learning (USL) approaches have the capability to identify hidden patterns in large data sets without human interference or -thus the term "unsupervised"- and are therefore often used for analysis and grouping of large unlabelled data sets. Instead of modelling merely the decision border between classes, generative models go into great detail to mimic the real data distribution and learn the different data points. The sole disadvantage these models have in comparison to discriminative models is their propensity to outliers (Johri et al. 2021). Clustering, association, and dimensionality reduction are the three main tools utilised in unsupervised learning models. Unlabelled data can be grouped based on their similarities or differences using clustering as a data mining approach, Figure 2B. As an illustration, k-means clustering algorithms divide related data points into groups, where the k-value (cluster centre) denotes a grouping's size and granularity, e.g., useful for market segmentation and image compression. Another unsupervised learning technique is association, which employs various parameters to detect relations between variables in a given data set. Recommendation engines, i.e., "Customers Who Bought This Item Also Bought" and market basket analyses both regularly apply these techniques. Dimensionality reduction is a learning strategy that is utilised when the number of features (or dimensions) in a given data set is excessively large. Data integrity remains intact while the amount of data inputs are brought down to a computationally manageable level. This method is also frequently applied during data pre-processing, such as when auto-encoders clean up visual data to optimise image quality (Manning et al. 2008; Roldan-Baluis & Vasquez 2022).

2.3.3 Semi-Supervised Learning (SSL)

Semi-Supervised Learning (SSL), as the name implies, is a hybrid technique between supervised and unsupervised learning. It is a broad category of machine learning techniques that makes use of both labelled and unlabelled data. The fundamental principle of semi-supervision is to treat each data point differently depending on whether it has a label or not: Supervised training adjusts model weights to reduce the average difference between predictions and labels. With limited labelled data, however, the algorithm may find a decision boundary that is valid for the labelled points but does not apply to the entire distribution. On the other hand, unsupervised learning tries to cluster data points based on feature similarities, thus minimizing the variance in predictions. An unsupervised algorithm, however, might identify sub-optimal clusters in the absence of labels for direct training (Johri et al. 2021; Mor, 2022). Both supervised and unsupervised approaches have the potential to fall short in case of challenging clustering environments or in the absence of enough labelled data respectively. The semi-supervised setting, however, can leverage the advantages of both: Labelled data serve as a quality control tool, grounding the model predictions and adding structure to the learning phase by identifying the number of classes and which clusters belong to which class. Unlabelled data provide context by introducing a large amount of data to the model, thus increasing accuracy and precision. In classification or clustering problems, the concepts of continuity and cluster assumption suggest that data of the same cluster and data which tend to be grouped into high-density clusters are likely to share the same label. Thus, a decision border should not be in areas where data are closely spaced but rather located between high-density areas, creating distinct cluster divisions. The so-called manifold assumption holds that a high-dimensional data distribution can be represented in an embedded low-dimensional space, also called data manifold (Mor, 2022; Li & Liang 2019*)*. Another frequently employed method for semi-supervised learning is consistency regularisation, which promotes network predictions to be close to the training samples, thus not being significantly affected in the output even in presence of a large amount of unlabelled data (Englesson & Azizpour 2021). Pseudo labelling, first described by Lee in 2013, is a process that trains a model on a small, labelled data set initially to predict pseudo labels for the large unlabelled data set thereby enhancing the model's performance (Lee 2013).

It becomes apparent that the real-world challenges related to data collection promote techniques that exploit unlabelled data. In these cases, data set engineering is more effective when it takes the use of as much unlabelled data as feasible. Labelling data is notably laborious, and time-consuming, and might often require domain knowledge in many NLP tasks, such as webpage categorization, audio analysis, or named-entity recognition, or less conventional machine learning applications, such as protein sequence classification. However, one must note that data labelling improves data context and quality for individuals, teams, and businesses. The major gains are more accurate predictions and increased usability. Accurate data labelling improves quality control in machine learning algorithms, as it enables the model to be trained properly and produce the desired results, thus providing "truthful data" for further testing and iterations and eliminating scenarios of "trash in, trash out." Labelling data also increases data usability, i.e., by reclassifying a categorical variable into a binary variable or by data aggregation, thus lowering the number of model variables or allowing the inclusion of control variables which can improve the model's performance. High-quality data is a key concern whether they are used to create computer vision models or NLP models, such as text classification or sentiment analysis.

2.3.4 Reinforcement Learning (RL)

Reinforcement Learning (RL) is undoubtedly growing as a result of AlphaZero's success, a single system that defeated world-champion programs in each of the games Go, shogi, and chess after teaching itself from scratch (Silver et al. 2018). RL is a machine learning method that uses trial and error to learn how to address a multi-level problem. The system is educated on real-life scenarios for decision-making. For the actions it takes, it either obtains rewards or penalties. Its objective is to maximise the overall reward. RL techniques can be classified according to different perspectives: model-based and model-free approaches, value-based and policy-based methods (or a combination of the two), Monte Carlo methods and temporal-difference methods, as well as on-policy and off-policy methods. However, the two main categories are model-based and model-free techniques (Zhang & Yu 2020): The model-based technique is used if the algorithm (agent) can foresee the reward for a certain action before taking it, thus planning what it should do. The technique is model-free if the agent needs to do the action to observe and learn from the outcome. Temporal Difference Learning (TD-Learning) and Q Learning are the most common model-free methods employed.

2.4 Deep Learning (DL)

The origins of Deep Learning (DL) can be found in 1943 when Walter Pitts and Warren McCulloch developed a "computer model based on the neural networks of the human brain" (Schmidhuber, 2015). They emulated the human thinking process using a set of mathematical formulas and algorithms they named "threshold logic." Deep learning has progressed since then, halted by only two important gaps in its evolution, also referred to as the infamous AI winters (McCulloch & Pitts 1943; Schmidhuber, 2015).

Frank Rosenblatt, a psychologist, created the first Artificial Neural Network (ANN), the Perceptron, in 1958, designed to simulate how the human brain processed visual information and learned to recognise objects (Rosenblatt, 1958). Thereby, ANNs contain nodes, which emulate neuronal cell bodies interacting with other nodes via connections (synapses) through axons and dendrites. Following the Hebbian theory of "nerves that fire together, wire together," connections between nodes in an ANN are weighted according to their capacity to provide a desired result. A layer of input nodes, a layer of output nodes, and several "hidden layers" between the two "outer" nodes are typical components of ANNs. Simple ANNs consist of an input layer with one to three hidden levels and an output layer, whereas Deep Neural Networks (DNN) have tens or hundreds of hidden layers (LeCun et al. 2015). Alexey G. Ivakhnenko and Valentin G. Lapa (1966) made the first attempts to construct deep learning algorithms in 1965 using models (Group Method of Data Handling - GMDH) with polynomial activation functions (Kolmogorov-Gabor polynomials), which were afterwards statistically analysed. The best statistically selected features from each layer were then transmitted to the next layer, a slow, manual process.

The first Convolutional Neural Networks (CNN) were employed by Kunihiko Fukushima following the first AI winter of the 1970s, which lasted for almost a decade. Fukushima created perhaps the first ANN that deserves the attribute "deep" by including several convolutional and pooling layers. He created the Neocognitron in 1979, which had a hierarchical, multi-layered design, that used a reinforcement technique to allow the computer to learn visual pattern recognition and grow stronger over time (Fukushima & Miyake 1982). Furthermore, Fukushima's design allowed for human adjustment of significant elements by raising the "weight" of specific connections. ANNs feed information forward for most tasks. In these so-called feed-forward neural networks, data from each node from the previous layer is transmitted to each node in the subsequent layer, processed, and then forwarded to each node in the subsequent layer (Choi et al. 2020). Although the fundamentals of a continuous backpropagation model, i.e., the backward transmission of errors for training deep learning models, were conceptualised by Henry J. Kelley (1960) and underwent further theoretical developments in the 70ies, the first practical demonstration was introduced by Yann LeCun (1989), applying backpropagation and convolutional neural networks on “handwritten” zip codes. This method was later used to read numbers on handwritten checks.

The second AI winter, which lasted from 1985 to 1990, stalled progress on deep Learning and neural networks. In the 90ies, many practical and commercial pattern recognition applications were dominated by non-neural machine learning methods such as Support Vector Machines (SVMs), a system for identifying and mapping related data, developed by Vladimir Vapnik and Dana Cortes (1995). Sepp Hochreiter and Juergen Schmidhuber (1997) developed the LSTM (long-short-term memory) utilised in Recurrent Neural Networks (RNN) to overcome the challenges of learning and keeping information for a long time. In contrast to feed-forward neural networks, in recurrent neural networks information can be transmitted in cycles or loops between nodes within a layer or to previous layers, where it is processed and forwarded again (Choi et al. 2020). LSTM models have proven to be more effective than traditional RNNs (LeCun et al. 2015). As computers began decreasing data processing time and graphics processing units (GPUs) were created in 1999, deep learning took its next big evolutionary milestone with the first feed-forward neural network-based “language” model developed by Yoshio Bengio and his group in 2001 (Bengio et al. 2003).

The term “deep learning” became popular in the mid-2000s after Hinton and Salakhutdinov (2006) solved the “vanishing gradient problem”—a fundamental issue in gradient-based learning methods that prevented upper layers from learning “features” formed in lower layers due to signal loss—which was the main cause of the computational slowdown. They demonstrated that layered feed-forward neural networks could be retrained one layer at a time by treating each layer as an unsupervised constrained Boltzmann's machine, followed by the use of supervised backpropagation for final adjustment (Selvam et al. 2016). Faster processing and GPUs boosted computing speed by 1000 times during the following decade and allowed neural networks to compete against support vector machines, especially in terms of output results. Another benefit of neural networks is that they continuously improve over time as more training data is provided. In face of the availability of big data, deep learning has received further attention from NLP researchers. Deep learning can perform difficult NLP tasks due to higher processing power availability: “When it comes to NLP, there is no way to solve the ambiguity of language.” (Johri et al. 2021). It is impossible to describe every conceivable meaning of words using rules or decision trees. Deep learning effectively resolves this issue because the algorithm itself infers the process of mapping an input to an output, avoiding the need for a programmer to supply the rules for decision-making (Johri et al. 2021).

2.5 Artificial Intelligence-based NLP Techniques

NLP tasks cannot be solved using one single approach. Rule-based (symbolic approach), traditional machine learning (statistical approach), and neural network-based approaches are now the three most popular methods. Although research focused increasingly on statistical models for probabilistic decisions, many of them require the text to be pre-processed using symbolic NLP methods, thus building up on one another to address problem-specific tasks. This kind of processing is known as a pipeline (Khurana et al. 2022). Pipelines are still widely utilised for various NLP tasks. This section displays some typical pipeline stages.

2.5.1 Rule-based NLP

Tokenization is known as the procedure used to break down character strings into smaller, more manageable units, called *tokens*, to facilitate further analysis when processing unstructured text. Tokens can be words, numbers (integers) or other characters. *Case conversion*: A frequent pipeline step is to convert all characters to lowercase, thereby treating “Good” and “good” as one word; otherwise, computers would treat them as different words, which can impact further text analysis. *Stopword removal* indicates the elimination of words like “the”, “and”, “in” etc. since they do not add to any meaningful interpretation and potentially slow down computation due to their frequency (Brants, 2003). The process of *stemming* involves removing a word's suffix to reveal the word's root form. For instance, after stemming, the words “consulting” and “consultant” are changed to the word “consult.” In contrast, *lemmatization* produces the source word rather than removing a word's suffix with help of a vocabulary. Lemmatization tries to return the words’ right basic form, which might be “drive” or “driven” depending on the context (e.g., drive a car, or as the attribute being driven), whereas stemming returns “driv” in the case of “driver” or “driven.” *Removal of punctuation* and *numbers* are other common basic NLP techniques, aimed at eliminating any punctuation or number(s) respectively from text data (Khurana et al. 2022). Of important note when using these basic removal and trimming techniques is, that some might be more appropriate than others, depending on the data type and the task at hand. For instance, it would not be appropriate to remove numbers and punctuation from medical, historical, or financial data.

Many researchers worked on NLP developing tools and techniques that have helped to shape the field to the present. Tools like Parts of Speech (POS) Taggers, Chunking, Named Entity Recognition (NER), Sentiment Analysis, Emotion detection/Emotion Recognition, and Semantic Role Labelling have significantly impacted NLP opening interesting research areas. Grammatical tagging, also known as *Part-Of-Speech tagging* (POS tagging), designates a word in a text (corpus) to a specific part of speech (e.g., nouns, verbs, etc.) based on its context and definition. This process is made more difficult by homographs (like “set”) and gerunds (verbs that finish in “ing” that are used as nouns). *Chunking*, also known as “shadow parsing”, is the method of extracting phrases from unstructured text by analysing a sentence to determine its elements (noun groups, verbs, verb groups, etc.) However, neither their internal organization nor their function in the main sentence is specified. Building on top of POS tagging it uses POS tags as input and output chunks. Chunking can separate sentences into phrases that are more helpful than individual words/tokens, as they might not accurately reflect the meaning of the text and yield meaningful results when trying to extract information from text, such as names of places or people (Collobert et al. 2011). When extracting information, Named Entity Recognition (NER) is often utilised to identify name entities and then classify them into various groups. Presently, NER faces some challenges when it comes to the right attribution of words when dealing with phrase/word order variations (e.g., "exacerbated eosinophilic asthma" vs "eosinophilic asthma, exacerbated"), name derivation (mediastinum as noun vs mediastinal as adjective), inflexions (bigger/biggest or cough/coughed), synonyms ("heart" vs "cardiac", "cirrhosis" vs "end-stage liver disease"), homographs ("dermatome" as skin area supplied by a specific nerve root or as a dermatological instrument to cut the skin), and abbreviations with different or ambiguous meaning ("CLD", chronic liver disease/chronic lung disease; "BO", bowel open or bowel obstruction). Identifying the correct meaning of a homograph is also tackled by a process called *Word Sense Disambiguation (WSD*), see section 2.5.2. *Spelling/grammar error detection and recovery* is primarily interactive because far from perfect. Highly synthetic sentences are prone to false positives (right words labelled as errors) and to false negatives, such as wrongly utilised homophones, i.e., words that sound the same but are spelled differently, such as sole/soul, their/there (Nadkarni et al. 2011).

Sentiment analysis refers to the process of computationally recognizing and categorizing opinions represented in a piece of text, thus determining the writer's attitude toward a specific topic, or product as positive, negative, or neutral. Investigating and classifying different emotional states from voice, gestures, facial expressions, and text is known as *emotion detection* or *emotion recognition*. *Semantic role labelling (SRL)*, also known as shallow semantic parsing or slot-filling, is a key method of *semantic analysis*. The approach identifies the semantic role of arguments to predicates in a sentence, such as experiences, aims, or outcomes, and assigns labels to them that represent the specific role, identifying "who did what to whom" (Kao & Poteet 2007). The association between the labels and their roles and the context in which they were placed in the sentence is then utilised to conclude the meaning of the text (Jurafsky & Martin, 2021). There are several projects that illustrate the arguments and roles for a given verb, including FrameNet (Baker et al. 1998), VerbNet (Kipper et al. 2000), and PropBank (Palmer et al. 2005). The most specialised representation is in FrameNet, and the most comprehensive representation is in PropBank. Despite having the lowest apparent utility, PropBank's more inclusive representation makes it easier to train a reliable semantic role labelling system (Palmer et al. 2005).

The pre-processing of the textual data using symbolic NLP methods such as the ones mentioned above is a fundamental element of the statistical techniques covered in the following section.

2.5.2 Machine Learning-based NLP

Collocation is referred to the frequency of a word combination in a text, often measured by bigram (two-words) counts. For instance, the bigram "New York" or trigram counts (three words sequences), such as "Upper West Side" will probably be used frequently in texts about the city.

In general, n-gram counts track the frequency at which an n-word sequence occurs. So-called *co-occurrences* are also important in this context, as they reveal details about a text's meaning: The Oxford dictionary, for instance, lists 19 definitions for the verb “make” (Oxford Website, n.d.). If the verb “make” appears in the same text document as the term “factory,” or better yet, in the same sentence, it most likely alludes to” production”. Such phrases-based machine translation dominated machine translation for applications like Google Translate up to 2018 (Kane, 2020). *Word sense disambiguation*: In NLP the identification of a word’s meaning has always been a challenge. The Lesk algorithm (Lesk, 1986) is a method that can be used to determine the right word sense as in dictionary meaning in a particular context by analysing the other words in the sentence or paragraph. *Word frequency*: The number of times each word appears in the text, also called term frequency (TF), is the most basic statistical metric that can be used to analyse a text. According to Zipf's law, which states that the most common term will appear twice as often as the second most common, which will then appear twice as frequently as the third most common word, and so on, function words like “the,” “and,” “a,” and “to” are the most often used terms in most documents. So, these words are frequently disregarded in analysis, e.g., by applying stopword removal, as they are negligible for text classification.

Common statistical methods for word frequency representations are one-hot encoding, Term Frequency (TF) vectors and Term Frequency – Inverse Document Frequency (TF-IDF). *Term Frequency-Inverse Document Frequency (TF-IDF)* is utilised to assess the significance of words or n-grams in a text. Considering the situation where the objective is to find the most pertinent word in each article of a group of articles, as mentioned previously stop words, like “the,” present an issue. The TF-IDF metric weights word frequency over the full corpus rather than word frequency alone: The TF statistic essentially represents individual word frequency in the singular text document divided by the total word count in the text. The IDF figure generally represents the percentage of text documents containing those words. The individual scores for TF and IDF are then multiplied to combine these two statistics, determining the words’ relevance (Shwartz, 2020). *Dimensionality reduction*: Latent Semantic Analysis (LSA) and Latent Semantic Indexing (LSI) are methods for determining a document's main topic. LSA is based on the distributional hypothesis, according to which it is possible to identify a word's semantics by looking at documents, identifying the ones in which the specific term appears, and determining the frequency of other terms mentioned in those documents; collocation and syntax are ignored as well as potential multiple meanings. Only co-occurrence is counted. Using a bag of words method embedding documents into a vector space, a large word-by-document co-occurrence matrix for the corpus is created, with the table's columns representing the documents, while the rows represent the words. Each word's TF-IDF score is represented in single cells. The relationship between words and documents is analysed by a mathematical method called Singular Vector Decomposition (SVD), which reduces the dimensionality of the corpus. The difference between LSA and LSI is that the former creates low-dimensional vectors using SVD describing the documents, while the latter computes low-dimensional vectors that describe each word (Kao & Poteet 2007; Shwartz, 2020).

Word embeddings were developed as an alternative dimensionality reduction technique to LSI, by Yoshua Bengio and his associates (Bengio, 2003). They investigated the possibility of using neural networks to generate low-dimensional representations of input terms. In contrast to sparse word representations (TF representations), which convert categorical variables to numerical vectors of thousands or millions of dimensions, in word embedding, each word is represented by a numerical vector with only tens or hundreds of dimensions. Aside from the computational advantages of low-dimensional word representations, researchers discovered the increased performance of several natural language processing applications when employing the word embedding technique to initiate neural network models (Chen & Manning, 2014; Sutskever et al., 2014; Qi et al, 2018).

2.5.3 Neural Network-based NLP

Google researchers (Mikolov et al. 2013a) developed Word2Vec (Word to Vector), a model for word representations, that was trained by either learning to predict a word in the data set (target word) from the immediately following and preceding words (called surrounding words or context words), called “continuous bag of words model” or CBOW, or by learning to predict the following and preceding “context” words from the current “target” word, also called “Skip-gram model” (Figure 3). The word embedding vector for the respective predictions, target word for CBOW and context words (Skip-gram), is then created from the vector in the single hidden layer (Shwartz, 2020).

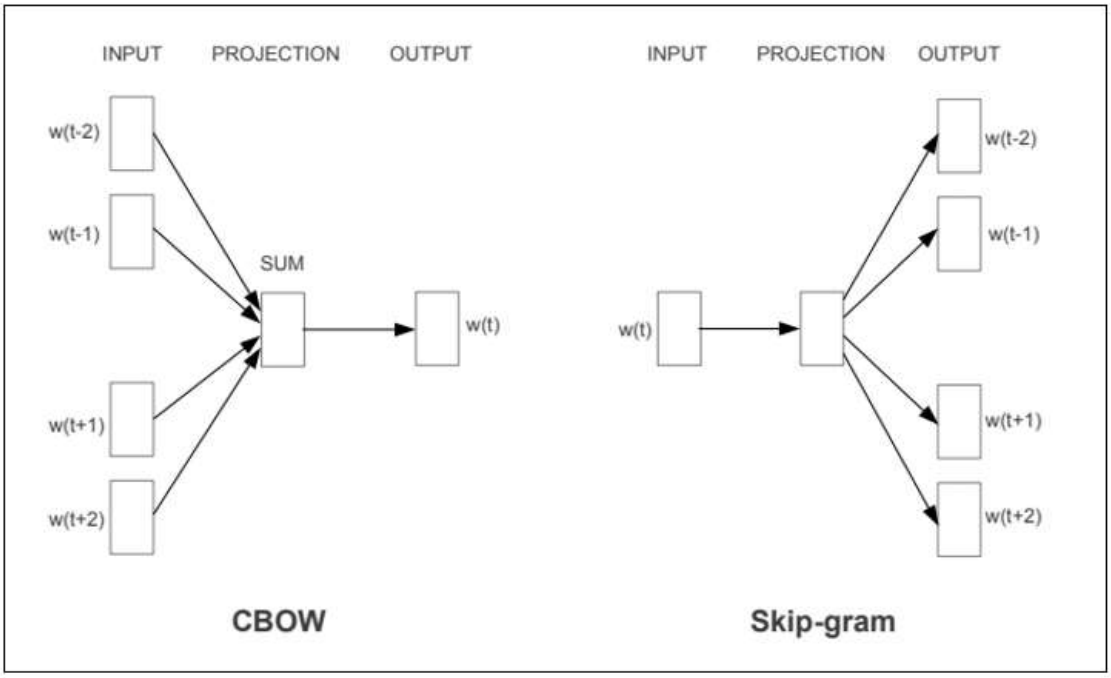


Figure 3: Schematic illustration of CBOW and Skip-gram model for word representation; from Mikolov et al. 2013a.

Interestingly, word embeddings seem to encode male/female relationship and capture some of the semantics (meaning) of words, after discovering that related terms, such as the vectors for “king,” “queen,” “man,” and “woman,” were close together. Further, the vector that results from the vector math “king (minus) man (plus) woman” is remarkably similar to the vector for “queen.” Related to this, encodings of the same term in singular and plural behave similarly, as the vector produced by "apple - apples" and “car – cars” are nearly identical (Collobert et al. 2011; Mikolov et al. 2013a). Global Vector (GloVe) created by Stanford academics, is another widely utilised type of word embedding (Pennington et al, 2014) utilising a different approach: While Word2vec makes use of co-occurrence in the local context (neighbouring words), GloVe relies on global word-to-word co-occurrence counts across the entire corpus. In reality, both of these models produce comparable outcomes for a variety of classification and regression tasks and both share two major challenges: First, the inability to handle words not found in the training corpus, out-of-vocabulary-words, or OOV, and the need for multiple vectors for morphological word variants (Shwartz, 2020).

The next step in the development of the word representation model was fastText, an open-source word embedding library created by Facebook researchers at FAIR, Facebook AI Research Lab (Bojanowski et al, 2017). FastText presents a different approach to deal with pending issues from the previously mentioned models: As opposed to learning word vectors for individual words, the system creates a vector for each n-gram in a word, whereby n-grams are the parts of which a word consists. To give an example, if the value of n is 3, then the tri-gram representation for the word “Italy” is “Ita”, “tal” and “aly.” The whole word “Italy” is then equal to the sum of its mastered n-grams for that word. The n-grams utilised in fastText in the default setting are trigrams to hexagrams (n=3 to n=6) or the word itself with word beginning- and finishing symbols. N-grams have one advantage over whole words: the system operates better with uncommon, technical, and out-of-vocabulary words. Nowadays, most techniques based on the co-occurrence frequencies of words in documents, like LSI, have been replaced by word embeddings. Due to the time and effort required to train task-specific, customised word embeddings, researchers take advantage by utilising pre-existing ones, such as Word2Vec, Glove, or fastText.

In summary, a very useful NLP project requires several supporting tasks, whereby the implementation of “lower-level” activities of the NLP pre-processing as described in the upper sections must be completed sequentially before “higher-level” tasks, described in the subsequent sections, can be performed (Nadkarni et al. 2011). A modular, pipelined system design that supports “mixing-and-matching”—where the result of one analytical module becomes the input to the next—allows for the employment of several algorithms for a given task, optimizing computing power and the quality of the results.

2.6 Common Applications of Machine Learning and Deep Learning Algorithms

Nowadays, artificial intelligence as well as machine learning have become buzzwords in technology, and their fields are expanding rapidly. Without the aid of AI and ML applications like Google Maps, Google Assistant, Alexa, etc., it would be unthinkable to get through daily life. The following list includes major industries which employ some of the most popular real-world machine learning applications.

Face recognition: One of the most popular machine learning implementations is image identification. It is used to identify e.g., digital images, people, places, or objects. Automatic friend tagging suggestion on Facebook is the most well-known application of image recognition and facial identification (Stone et al. 2008). *Automatic Speech recognition* (ASR), also referred to as "Speech to text recognition" (STR), is the process of turning spoken commands into text. Deep learning, neural networks, and machine learning technologies process speech, and analyse grammar, syntax, structure, audio, and voice signal composition. Currently, a variety of voice recognition applications, such as virtual personal assistants like Google Assistant, Siri, Cortana and Alexa make extensive use of machine learning methods to carry out vocal commands. Text data along with a substantial amount of recorded speech, are used to train deep neural networks in voice synthesis to create synthetic speech from text, also known as neural "Text to Speech" (TTS) (Ni et al. 2022).

Deep neural network approach for *traffic prediction*: Robot navigation such as Google Maps or car navigation systems show different navigation routes, predict the traffic conditions and average travel time with the help of real-time location of the vehicle via the Google Map app and sensors and compare it with previous data taken. All application users help to improve its performance, as it processes information from the user and sends it back to its database (Chan et al. 2021).

The *food industry* displays a wide range of AI applications: CNNs have achieved outstanding results by analysing impact factors such as UV light, salinity, heat, and water in the agriculture sector, thus increasing productivity. Similarly, AI advancement has also impacted the commercial food processing industry. Food processing and monitoring automate operations involving food analysis, such as colour, texture, and size selection, as well as packaging processes. Several machine vision systems, such as x-rays, thermal imaging and magnet resonance imaging are employed for image recognition. To analyse and interpret the image data, algorithms like k-nearest neighbour, support vector machine, neural network, fuzzy logic, and a genetic algorithm can be used (Zhu et al. 2021).

Recommender systems: Nowadays, all major digital service providers and e-commerce corporations rely on recommendation algorithms to give a personalised user experience and boost sales performance or advertising revenues. Data mining, deep learning and neural network algorithms are employed for product recommendations or search results (Amazon, Facebook Marketplace), suggested music (Spotify), videos or channel subscriptions (YouTube), movie recommendations (Netflix), friend suggestions (Facebook), travel (TripAdvisor) or restaurant (Yelp) recommendations and job postings related contacts or courses (LinkedIn) by mining a user's browsing history and personal ratings given, or according to qualifications published online, experience, and propensity to respond to direct messages (Kumar, 2018).

Spam detection: To successfully combat the threat posed by email spam, global email providers like Gmail, Yahoo Mail, and Outlook have merged a range of machine learning techniques (such as text classification) with neural networks in their spam filters. By analysing large numbers of spam and phishing emails across vast computer networks, these machine learning algorithms are capable to identify and classify these emails. Since machine learning can adjust to changing circumstances, spam filters on Gmail and Yahoo mail do more than just analyse spam emails using pre-existing criteria, i.e., as they continue to filter spam, they develop new rules independently based on what they have learned (Puri et al. 2013; Dada et al. 2019).

Fraud detection and online transaction security are improved by machine learning algorithms, as fraudulent transactions could take place whenever an online purchase is made, including the use of fake accounts or IDs or money theft. Feed-forward neural networks assist by determining whether the transaction is legitimate or fraudulent. Support vector machines and decision trees among other machine learning models have also demonstrated good performance in terms of accuracy and coverage (Minastireanu et al. 2019). *Stock market trading, price prediction and risk management*: The inherent temporal and state awareness of architectures like LSTM and Deep Reinforcement Learning (DRL) makes them more widely employed. In particular, LSTM is advantageous due to its capabilities for long-term memory, which is required for price prediction, market simulation and trade strategy applications. For stock pricing, RNN in combination with LSTM or even CNN are used in price predictions, due to its ability to identify local time series patterns and extract useful features in high-frequency market data (Olorunnimbe & Viktor 2022).

Entertainment and gaming: DRL in particular, has made significant advances in video games, AlphaGo and AlphaGo Zero have demonstrated outstanding performance in computer vision, natural language processing, decision-making support in high-dimensional state space and substantially enhancing the generalisation and scalability of classic RL algorithms (Silver et al. 2018; Shao et al. 2019).

Weather forecasting: AI algorithms are increasingly improving weather, climate, and disaster predictions. Human mistake is not a constraint on the predictive capacities of AI and machine learning algorithms, Human mistake is not a constraint on the predictive capacities of AI and machine learning algorithms, analysing a huge number of historical weather maps to learn weather patterns for more accurate predictions rather than solving a set of challenging physical equations like standard models do. Time series-based recurrent neural networks, artificial neural networks, and support vector machines are being explored for the task (Singh et al. 2019).

Customer segmentation and targeted marketing: Clustering (k-means) is an unsupervised learning technique that assists customer segmentation by classifying similar customers into a pre-set of segments, such as demographic segmentation (e.g., age, gender, occupation, marital status), geographic segmentation (e.g., city of residence, country), technographic segmentation (e.g., software, technology), psychographic segmentation (e.g., personal attitudes and traits, values, interests) or behavioural segmentation (e.g., action/inaction, consumption habits, feature use, average order/value, session frequency etc.). In a business context, it empowers marketers to interact with every customer in the most efficient approach, employing specific advertising to a specific segment (targeted marketing) thus maximizing customer benefits and optimizing sales (Kansal et al. 2018). Increasingly, customer experience is being improved through the usage of AI technologies. Several machine learning and deep learning methods and algorithms, neural networks, and natural language processing are used by many of the chatbots found on e-commerce websites. These chatbots are programmed to instantly respond to a variety of common customer queries.

Healthcare: Machine learning methods are also gaining popularity in healthcare and disease diagnosis. The promise of inexpensive and time-efficient Machine Learning-Based Disease Diagnostics (MLBDD) has been widely demonstrated in the past years (Ahsan & Siddique 2022). For classification and regression-related challenges, support vector machines are employed in medicine for disease detection, such as malignant tumours. The Naïve Bayes classifier is used to project disease likelihood as well as early breast cancer detection (Houssein et al. 2021). Over the years, deep learning has been extensively employed in the medical fields of pathology and radiology to diagnose diseases (Hayashi, 2019). Convolutional neural networks are widely utilised for biological image detection and recognition as well as the localisation of human organs (Yap et al. 2017). Data mining, machine learning, semantic text analysis or predictive analytics can be used for health promotion and disease prevention (e.g., increasing personalised Healthcare) at all levels of research, surveillance, and intervention: The aim is to gain a better understanding of human health and disease and exactly identify “at-risk” groups considering health factors, such as environment, genetics and lifestyle, detect threats and improve disease monitoring and decision-making processes through real-time analytics, as well as minimise waste of resources, which can have an impact on healthcare expenditures. Integration and harmonization of existing medical data sources (e.g., health organisations and insurance databases) with big data sources, spanning from social media and internet queries to healthcare apps and wearable electronic devices, can support public health and enhance personalised care and intervention. Longitudinal information on individuals retrieved via text mining, semantic text analysis and sentiment analysis (social listening tools) creates deeper insights, especially when paired with socioeconomic determinants of health (Infodemiology and infoveillance) (Eysenbach, 2016). Based on data gathered from real-time social media updates, satellites, historical data, and other sources, AI and ML technologies are also being implemented to monitor and predict worldwide epidemic (pandemic) disease outbreaks. One current application of AI technology (e.g., support vector machines and artificial neural networks) is the prediction and monitoring of malaria outbreaks, especially in third-world countries, by taking into consideration environmental information such as temperature, average monthly rainfall, and the total number of positive cases (OECD Health Policy Studies, 2019).

2.7 Digitalisation Process of the Pharmaceutical Industry

2.7.1 Introduction to Artificial Intelligence Implementation in Pharma

Although the academic and scientific field of artificial intelligence has existed since the 1950s, AI technology has only recently become fundamental due to its performance development to human-level capabilities. In the previous section (2.6), some examples of AI and machine learning implementations in major industries were mentioned, such as travel, finance, gaming, retail, media, and general healthcare. Albeit slowly, over the past decade, artificial intelligence and machine learning have also continuously gained a foothold in the Pharmaceutical Industry, becoming the much anticipated “breakthrough technology” to have a transformative effect on pharmaceutical drug discovery, research and development (R&D), clinical trial design and conduct and manufacturing processes.

2.7.1.1 Research and Development (R&D)

Over the last decade, there has been a significant increase in the number of pharmaceutical businesses adopting AI in drug research and development. Several global players, such as Novartis, Roche and Pfizer have either collaborated with or purchased AI technology, such as IBM Watson, a processor that NLP for data analytics purposes such as disease prediction: Roche, for instance, used Watson to forecast the 3-year-risk-rate of kidney failure in diabetics with a 79% rate of success (Damiati, 2020). The application of AI is influenced in part by dramatic developments in computational technology and the simultaneous elimination of barriers regarding the gathering and processing of massive amounts of data. Furthermore, it is becoming prohibitively expensive to introduce new medications to the market and patients with drug approval fees increased by up to 5% amounting to US$3.2 million in 2023 (EMA; FDA Website, 2022), and drug development from molecule to market authorization costing up to US$5 billion per drug (Schlander et al. 2021). The pharmaceutical business is attracted to AI/ML approaches because of their automated nature, predictive power, and increased efficiency. Especially the therapeutic areas of oncology, neurology and infectious diseases are targeted by companies to enhance drug research and development with the support of AI, Figure 4.

Drug discovery accounts for a substantial portion of machine learning applications in pharmaceutical sciences, owing to the usage of “high-throughput screening, combinatorial chemistry, and computer-aided drug design.” (Jumper et al. 2021). Quantitative Structure-Activity Relationship (QSAR) studies were one of the earliest fields in which artificial neural networks were used. In a nutshell, “QSAR methodology links a compound's physicochemical characteristics, i.e., how a protein's amino acid sequence results in its folded three-dimensional structure - also known as the protein structure prediction problem, to its related chemical or biological activities” (Jumper et al. 2021). Significant advancements in this difficult computational challenge were announced by DeepMind, a London-based artificial intelligence company now part of Alphabet Inc. since 2020. The “AlphaFold 2” program produced protein models of a quality approaching that of laboratory experimental research. Although still in the developmental stage, the AlphaFold Protein Structure Database at the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) already comprises around 350,000 protein models (Jumper et al. 2021; Thornton et al. 2021).

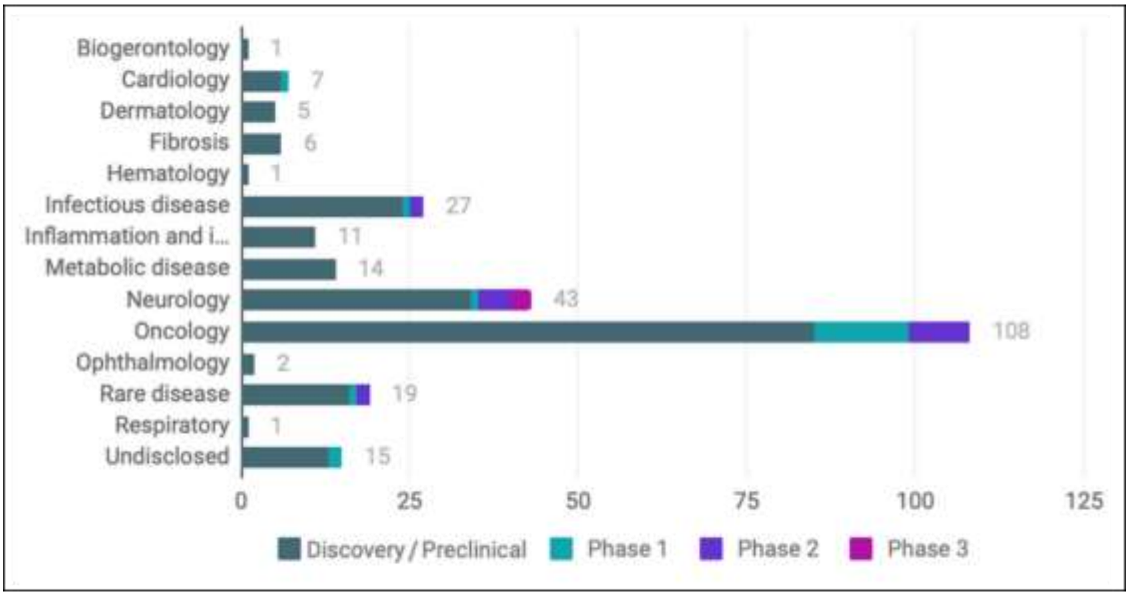


Figure 4: Number of AI-assisted drug research and development related to its development stage (Discovery/ Preclinical Phase, Phase 1, Phase 2 and Phase 3) and therapeutic area; adapted from BenchSci, 2021.

For reasons of drug efficacy and safety, it is crucial to identify the physicochemical characteristics of a pharmacological substance during the pre-formulation stage, including its solubility, stability, interactions with excipients (Han et al. 2019), as well as bioavailability, such as absorption, distribution, metabolism, and elimination (Gaisford & Saunders 2012, Chan et al. 2019). Here, the application of deep learning using ANNs has demonstrated outstanding success in the accurate prediction of solubility-enhancing effects of hydrotrope compounds including possible toxicity risks (Damiati et al. 2017). The process of turning pure drug ingredients into patient-administered pharmacological products is known as the formulation stage of drug development. Artificial neural networks have attracted a lot of attention in this field and have emerged as the most widely used machine learning tool in the prediction of pharmaceutical formulations (Yang et al. 2019): Examples include the prediction of the physical stability of solid dispersions, the biophysical properties of therapeutic monoclonal antibodies, and the dissolution characteristics of pharmaceutical tablets (Damiati, 2020).

Clinical trial design, conduct, and results analysis are among the most recent fields of drug research and discovery where beneficial disruption from AI and ML is beginning to occur (Kolluri et al. 2022). Especially the Corona Virus Disease 2019 (COVID-19) outbreak enhanced the use of artificial intelligence and machine learning in clinical trials, due to the need for greater reliance on digital technologies for clinical trial conduct. Several ML applications can be used to improve clinical trial efficiency, including an effective selection of sample size and patient population, faster adjusting to variations in patient recruitment sites, and utilising electronic health records (EHR) to minimise data errors, such as duplicate entries or missing values (Seddon et al. 2012). For greater patient safety, ML can also be utilised for real-time data access and remote monitoring, such as monitoring of medical and other health-related signals for any sign of harm to study participants (Bate & Luo 2021). These applications offer the opportunity to address the inefficiencies and uncertainties that arise in conventional drug development approaches while minimising bias and human intervention. Other challenges can now be addressed using AI methodologies, such as the analysis of big data sets and laborious compound screening while minimising standard error and reducing R&D costs amounting to over US$3 billion (averaged) over a decade (Mak & Pichika 2019).

2.7.1.2 Manufacturing Processes

The spread of COVID-19 had a significant impact on global manufacturing bases, which are now working to boost productivity and increase supply chain connectivity by adopting digital manufacturing technologies harnessing in addition to artificial intelligence also the Internet of Things (IoT). Due to the highly controlled and sensitive nature of the pharmaceutical production environment, even the smallest errors could have a profound negative influence on the manufacturer's business, legal situation, and reputation (Sehlstedt et al. 2016). For instance, a few years ago, problems in human monitoring and packaging forced a major pharmaceutical producer to recall nearly half a million tablets. Companies are already reducing such missteps thanks to digitalisation and automation, thus reducing financial and reputational harm (Kitson et al. 2018): The Big Pharma players have made significant investments in automated production lines that make it simpler to handle electronic batch records, maintain hygienic standards, and monitor process performance towards the implementation of Pharma 4.0 (Hole et al. 2021; Manzano & Langer 2020).

2.7.2 Emergence of a New Need: Digitalisation of Customer Relationship Management (CRM)

Despite considerable advancements in R&D and production, pharmaceutical businesses are "still in an experimental phase when it comes to supplying digital services beyond traditional products" (Parida et al. 2019). Given that pharmaceutical businesses operate in both B2B and B2C markets, AI solutions and digital technology influence both internal and especially external media-related parts of a company's business operations. Before diving further into the topic, three important definitions must be mentioned, as these terms are often used interchangeably, but, despite being closely related to each other, all imply different meanings: digitization, digitalisation, and digital transformation.

Digitization relates to the process of converting analogue information and information management from a physical into a digital form, which computers can store, process, and transmit (Verhoef et al. 2021). Digitalisation refers to the process of adapting an existing business model to new, digital information technologies. It embraces its ability to collect data and find hidden patterns, facilitating intra- and inter-organizational decision-making, processes, and architectures. Digitization and digitalisation work together to transform businesses into digital enterprises that are more agile and perform better (Khan et al. 2020). Especially for Pharma, digital technology helps to streamline operational processes to optimise internal and strategic efficiency, such as drug discovery and development, organisation and conduct of clinical trials, drug manufacturing processes and distribution (logistics) as well as compliance with internal Standard Operating Procedures (SOP) and adherence to institutional regulations. One fundamental aspect in the digitalisation process of a company is the exploitation of external opportunities, i.e., the attention to communication and interaction with the customers and stakeholders, capturing the potential of new markets as well as accessing market and customer information and interpreting the insights gained. The implementation of digital marketing strategies and the integration of social media bear enormous potential for pharmaceutical companies to become digital businesses (Verhoef et al. 2021).

The most prominent stage, known as digital transformation, is characterised by a company-wide shift in the adoption of a new business model. Beyond digitalisation, this step implies customer-centric organisational reforms that are supported by leadership and driven by radical corporate culture challenges: Digital transformation entails the use of digital technologies to enable interactions across borders with suppliers, customers, and even competitors to achieve a competitive advantage and transform the organisation by leveraging existing core competencies or developing new ones (Singh & Hess 2017). Hence, the establishment of new business models brought about by the adoption of digital technology is inextricably tied to digital transformation (Sebastian et al. 2020).

The impact of COVID-19 confirmed the importance of digitalisation: Facing increasing customers' demand and competition, patent expirations, and rising pricing pressures, digital (AI) implementations in several sectors allowed to improve performance in manufacturing productivity and competition, as well as financial sustainability by exerting more accurate planning and forecasting (Faraj et al. 2021). Furthermore, the COVID pandemic challenged the traditional product-oriented pharma business model, especially in the marketing sector and external communication in general, promoting a re-orientation to customer-centric strategies: Integration of online and offline environments, application of omnichannel approach and better customer and market segmentation allow the creation of a seamless process of a company's interaction with partners and intermediaries, as well as end customers. Over the last few years, the pharmaceutical sector has seen various changes in its approach to communication, from traditional product-based marketing to the implementation of digital strategies in a multi-channel approach and now the intent to transition from the "multichannel" to an "omnichannel" environment integrating push and pull marketing strategies for a targeted customer-centric approach.

"Multichannel" essentially means employing multiple channels rather than just one. Many pharmaceutical businesses began developing their digital presence with the implementation of brand and patient websites, CRM/email marketing systems along with traditional printed media and visits. Since each channel was independently developed, they tended to exist on their own, frequently repeating the same content in the same or just slightly modified manner. Although some pharmaceutical companies applied Search Engine Optimization (SEO) strategies for improved visibility, customers or patients were supposed to find the information and decide what they wanted (or didn't want) to absorb. Less emphasis was placed on the experience they provide and more on the channel with unilateral information flow. So, although using multiple channels to reach customers, the multichannel method, has lost its efficacy, mostly because the channels are highly independent of one another (Azovev et al. 2019). The challenge behind the keyword "omnichannel marketing" in the Pharmaceutical Industry lies in combining and integrating traditional and digital marketing channels so that the target audience constantly feels addressed through the topics that are pertinent to them. Individual measurements, i.e., push and pull marketing strategies, are coordinated like instruments in an orchestra to give a consistent impression to the outer world, by using the appropriate channel and content to reach the individual customer at the right time. If done correctly, this generates a flow where all marketing initiatives interact with one another, and the consumer relationship is improved with each interaction. The success of omnichannel is demonstrated by the experiences of global healthcare organisations such as Johnson&Johnson, who established a "flexible, yet secure digital IT organization to support the faster development of smart healthcare products and the improvement of customer and patient experience with the company" (Azovev et al. 2019; Cordon et al. 2016).

Push marketing (or outbound marketing) channels are used to push goods and information onto markets, so the focus of these activities lies on distribution channels. For instance, new medications and treatment options are extensively promoted to sick funds, doctors, pharmacies, and wholesalers. Push marketing channels include sales representatives and medical science liaisons, traditional post mailings and print media as well as email newsletters. Pull marketing (or inbound marketing) channels, in contrast to push marketing, aim to ignite the target audience's interest via tailored messages so that they independently seek more information on products or compound research. Channels for pull marketing are typically SEO strategies for better visibility of articles and papers published on company-owned websites, online platforms for Healthcare Professionals (HCP) for training and exchange of scientific information and expertise, such as Medscape, coliquio, Doximity, MomMD, platforms for webinars, etc., Search Engine Advertising (SEA) and social media.

Social media offers enormous potential as a marketing channel, especially as the shift from push to pull is at a “one question” distance. Thereby, it allows to collect more insightful information, about a customer’s interests, sentiments and desires than a website's simple click rate. By providing information upon request, as opposed to the “flood of information,” the result can be a successful collaboration that benefits both parties.

2.7.2.1 The Rise of Social Media in Pharma

The evolution of social media, from an electronic information exchange site in the early 2000s to a virtual meeting location, shopping platform, and now essential 21st century marketing instrument in less than a generation, has substantially altered conventional one-way, marketing-controlled communications. Nowadays, the “balance of power” has moved from businesses to customers, who utilise social networking sites to get substantial brand awareness and product knowledge, thus gaining a major role for industries regarding customer relationship management (CRM). In the past two years, social media users increased by 18% from 3.9 billion in 2020 to 4,59 billion in 2022 worldwide (Dixon, 2022). A Eurostat survey on Information and Communication Technology (ICT) in households and by individuals revealed that, on average, 55% of Europeans between the ages of 16 and 74 have looked for health-related information online, with percentages above 70% in Finland, the Netherlands, Denmark, and Germany (Eurostat Website, 2021). The numbers are similar in the US, where a survey on adults (≥18 years) reported that 59.2% of the participants researched healthcare information on social media at a weekly basis (Neely et al. 2021). Another US study revealed that while 61.2% of the population sought health information online first back in 2008, the percentage had risen to 74.4% in 2017 (Finney Rutten et al. 2020). The percentage of people looking for health information online is even higher in some Asian nations: 79%, 80%, 85%, 85%, and 86% for mainland China, the Philippines, Hong Kong, Indonesia, and Vietnam respectively (Jia et al. 2021). Needless to mention, the Pharmaceutical Industry must recognise once more the great potential of B2B and B2C interaction on these platforms.

As anticipated, early adopters and quick followers in digitalisation would see significant revenue and profit growth: Top-tier pharmaceutical companies, such as Pfizer, Eli-Lily, Böhringer Ingelheim, Johnson&Johnson and Merck, are already present on social media and use ML-based monitoring technologies for customer analysis and maximisation of existing marketing opportunities. Among them, Novartis has demonstrated a focus on boosting social media interaction, increasing engagement by 41.5% with the help of data analysis. Current trends focus on outcome-based care, customer engagement, information availability, and process improvements which are paving the way for digital transformation through personalised, 24/7 medicine, omnichannel interactions, data-driven insights, and real-time responsiveness (Henstock 2019). Even though most businesses indeed understand the strategic relevance of digital transformation, few are adequately prepared for the disruption that digital technology will bring to the industry. Barriers include data governance, silo-ed implementation, insufficient skill set, lack of urgency/absence of investment and above all regulatory concerns.

2.7.2.2 Barriers to Social Media Strategy Implementation

The Pharmaceutical Industry is still reluctant to embark on the journey of implementing social media strategies due to multiple reasons, such as a perceived lack of urgency, scarcity of specialised workforce, a silo-ed implementation of digital marketing strategies and a lack of unified targeted marketing, strongly differing regulations regarding product promotion, and strict data governance guidelines.

Lack of urgency: Although routinely adopting new technologies, pharmaceutical businesses are slow on the uptake regarding the implementation of digital channels and campaigns: Some businesses may be strategically waiting for the field to stabilise, and leverage the successful strategies developed by big pharma and other early adopters, while others wait for formal, explicit guidelines from regulatory authorities, relying meanwhile on the traditional expert-driven scientific and product-based approach.

Lack of skillset: The reluctance to engage in social media strategies in Europe is apparent. In a recent survey conducted with 30 digital and marketing managers of different pharma companies in Austria in 2021, 60% of the participants stated that metrics analysis of social media activities would be just rudimentary or not performed at all. Further, 50% mentioned digital marketing skills being on or below average (Gerfertz-Schiefer, 2021). The pharmaceutical sector lacks a skilful workforce, who understands both the industry and the new digital environment and opportunities and drives digital change. Further, it lacks a clear vision for the successful deployment of marketing solutions as well as an internal unification of the different departments, e.g., marketing, medicine, and sales, on common agreement on tactics and sharing of data. The same Austrian survey, as mentioned above, shows that nearly 30% of the managers' companies deploy 26-75% of the marketing budget into digital media strategies (nearly 40% deploy between 1-25%) (Gerfertz-Schiefer, 2021). As increased data sizes have resulted in difficult analysis and utilization of digital strategies, it is crucial to implement organization-wide governance that aims to promote a culture of unified data-driven decision-making (Parekh et al. 2016).

Geo-based Pharmamarketing – silo-ed implementation of digital strategies: When developing their product marketing plan, pharmaceutical companies must keep in mind the various restrictions governing the advertising of branded prescription drugs. This is especially important when products are being marketed in several geographies with varying local legislation. While advertising of prescription drugs is acceptable in the United States, it is strictly prohibited in Europe by an intricated regulatory framework, including the European Medicine Agency (EMA), the UK's Medicines and Healthcare Regulatory Agency (MHRA) and SwissMedic, pharmaceutical organizations such as the European Federation of Pharmaceutical Industries and Associations (EFPIA) and the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA). Further, there are national regulatory authorities and stakeholders in each European member state: For Austria, the national regulatory authority is the “Agentur für Gesundheit und Ernährungssicherheit” (AGES), and the two national organizations representing the pharmaceutical industry “Verband der pharmazeutischen Industrie Österreichs” (Pharmig) and “Interessensgemeinschaft österreichischer Heilmittelhersteller und Depositeure” (IGEPHA). This framework already makes a significant difference in fully utilising the possibilities of social media.

While the US Food and Drug Administration (FDA) has issued several guidelines on the promotional use of digital media over the past years, the European Medicines Agency (EMA) has published a guidance document on the use of social media and digital channels only in September 2022 (IFPMA, 2022): Helmed by the general restrictions on advertising prescription (Rx) products, in the case of implementation of digital marketing campaigns, reliance is placed on the standard guidelines for advertising in respect to the target audience (i.e., HCP or patients), or guidance provided by national health authorities and organisations. Guidelines provided by IGEPHA on a national level and IFPMA in collaboration with EFPIA on the European level refer to the “Act on the Advertising of Medicines”. Hence, based on the target audience global pharma companies must develop geography-based social media strategies: Novartis, for example, connects with the public for its brand Gilenya® (a medication created to treat multiple sclerosis) via a special Twitter handle named @GILENYAGoUSOnly. In the introduction, the Twitter handle clearly states that it is conceived for a US audience further outlining various engagement parameters, such as the response window, how to provide personal details, and the discretion Novartis would exercise in response to specific tweets (Limaye & Saraogi 2018).

According to the FDA guidance on “Internet/Social Media Platforms with Character Space Limitations - Presenting Risk and Benefit Information for Prescription Drugs and Medical Devices”, issued in June 2014, when promoting products, a company must inform the consumer about both benefits and potential risks: A challenging task when it comes to 140 tweet characters’ limitation (FDA Guidance Document, 2014). One prominent example was when the FDA's Office of Prescription Drug Promotion (OPDP) sent a warning letter for misconduct in social media practices to the Canadian company Duchesnay: They missed to mention the risks of the morning sickness drug Diclegis® in a tweet posted by Kim Kardashian promoting its advantages. In September 2017, Twitter announced that the site's 140-character restriction had been increased to 280 characters, also to allow room for pharma marketers to oblige authorities' regulations (Rosen, 2017).

Data governance - Handling of compliance-relevant information: The Pharmaceutical Industry is not only discouraged by laws about the handling of promotional material on social media, but major concerns revolve around the challenge of timely evaluation and processing of incoming compliance-relevant information in adherence to corporate Standard Operating Procedures (SOP) (Limaye & Saraogi 2018). Areas such as quality (product complaints), regulatory (market availability, market authorization), and especially drug safety, known as pharmacovigilance (PV), which implies the detection and processing of Adverse Drug Reactions (ADR), need up to 24/7 monitoring. Further, inquiries pertaining to medical and patient information, as they can contain drug safety information (e.g., off-label use – use outside marketed indication), need to be processed, and archived, especially because they must be presented in the context of internal and institutional regulatory audits. As a result, the chat function on pharma-owned social media channels is either limited or disabled, such as the Facebook accounts of Johnson&Johnson or Chiesi Pharmaceuticals: Such measures might eliminate the possibility of PV mentions, but they also remove the main function of social media by cutting off its biggest asset, namely the interaction with and among customers. Other companies outsourced monitoring to third parties (agencies), which often scroll through the comments and postings manually at regular intervals to screen for potential adverse reactions but, to the author's knowledge, not for compliance-relevant data. Unreported adverse events on social media and mis-reporting or non-reporting of compliance-relevant information in general offer an increasing danger of operational and reputational harm across the pharmaceutical business. Audits, penalties, regulatory sanctions, and legal actions resulting in hefty fines may occur from non-compliance. Since 2000, over US$95,244,937,209 billion have been paid to the authorities for offences, such as off-label or unapproved promotion of medical products, drug or medical equipment safety violations, or false claims among others (Violation tracker Pharmaceuticals, 2022).

2.7.2.3 Special Case: Pharmacovigilance

An Adverse drug reaction (ADR) is by definition of the World Health Organization, (WHO) "any unintended response (either harmful or beneficial) related to the pharmacological properties of a drug occurring at doses normally used in humans" (WHO, 1971). Adverse events (AE) are "any unfavourable medical development (unintended sign, symptom, or disease) in a patient or clinical trial participant after administration of a pharmaceutical product, which must not necessarily be related to the treatment" (EMA Website, 1995; Ventola, 2018). In relation to other significant causes of death including the top three heart disease, stroke, and cancer, adverse drug reactions (ADRs) are estimated to be between the fourth and the sixth most common cause of death worldwide (Le Louët 2022; WHO, 2020) with medical costs amounting from US$30 billion to US$130 billion annually only in the US (American Medical Forensic Specialists – AMFS, 2019). Because such reactions or events are not always recorded in clinical trials, pharmaceutical companies rely on post-market safety surveillance, which includes medical literature screening, Electronic Health Reports (EHR), Post Authorization Safety Studies (PASS), Non-Interventional Studies (NIS), and un-solicited (spontaneous) reports, to detect adverse reactions that may occur as the medication is used by a broader spectrum of patients. Unfortunately, the reality looks different: Underreporting, a lack of geographic diversity, and gaps in time between occurrence and reporting make all these measures rarely efficient. The Commission Report on Pharmacovigilance provided by EMA states that up to 96% of ADR and AEs go unreported and that reporting by patients and HCPs is incredibly low (EMA Website, 2016); A recent Medline systematic Review from 2000 to 2022 states that overall underreporting still reaches over 90% (Li et al. 2022). First attempts at improving adverse event detection on an international scale by applying natural language processing have been made in 2015 following the Innovative Medicines Initiative (IMI) project of the European Union (EU): Novartis and other pharmaceutical companies implemented Web-RADR - Web-Recognizing Adverse Drug Reactions (IMI Website, n.d.), a still ongoing project, in efforts to address this unmet need. Out of over three million postings taken from Facebook and Twitter screenings (excluding 55% as spam), 2% of them have been classified as possible adverse events (AEs) or “proto-AEs” using the Medical Dictionary for Regulatory Activities (MedDRA) (Limaye and Saraogi 2018). The discovery of “Crix belly” syndrome, also known as lipodystrophy syndrome, which occurs using an antiretroviral treatment for HIV, is an excellent example of an ADR being discovered through social media platforms and not during clinical trials because the studies only lasted 48 weeks, and the adverse reaction appeared after that period.

In the case of social media monitoring, companies exert caution about social media reports potentially exaggerating issues as happened to Sanofi, which had to shut down its Facebook page after being flooded by comments about severe hair loss from a patient who reacted to cancer treatment Taxotere®, eliciting strong emotions from the greater patient community. Later, Sanofi did re-open the page with terms of use. Although the European Union does not require social media monitoring for AEs, it does demand for observed occurrences to be reported (Limaye & Saraogi 2018).

Presently, companies are not required to monitor social media for adverse events and report them in the EU. They must notify regulators, though, if they are discovered during scanning. Regardless, firms are required to monitor and disclose any drug side effects reported on their own sponsored websites. By The General Data Protection Regulation (GDPR), pharmaceutical businesses are prohibited from collecting any personally identifiable information (PII), provided that they disclose on their websites and social media accounts that they engage in social listening and that the data contributed by users may be used for this purpose. The demands of the ethics committee differ from nation to nation, and it is essential to investigate them as well to ensure compliance (Limaye & Saraogi 2018).

2.8 Conclusion

This extensive literature review serves several purposes: (i) It provides an overview of the development and establishment of artificial intelligence, (ii) it illustrates its classification into different sub-fields, (iii) it highlights constant improvements and research successes, (iv) and its widespread use in various industries, as well as (v) its influence in a person's daily life. Further, it shows that (vi) although AI has entered the Pharmaceutical Industry, the use of AI technology is still hesitant, especially in the field of digital marketing, and the use of social media as a pull marketing instrument due to several aspects described in detail.

The method section shall present that already existing and widely used machine learning technology can be used in digital pharm-marketing to overcome the barriers that currently prevent the exploitation of the full potential of social media by pharmaceuticals.

3 METHODOLOGY

3.1 Introduction to Data Mining and Text Mining

Over the last decade, technological advancements made processing, analysis, and storage of big data possible, especially driven by the introduction of Web 2.0 and social media platforms, which brought a significant growth of User-Generated Content (UGC) on Social Networking Sites (SNS), such as blogs, forums, or social media (dos Santos, 2021). Data mining, also known as Knowledge Discovery in Databases (KDD), the process of discovering hidden patterns and valuable information in big raw data sets, has boosted business operations by supporting organisational decision-making processes. Further advances in NLP and machine learning have generated effective techniques and algorithms for processing and understanding natural language creating new possibilities for text analysis and social media mining (Gandomi & Haider 2015). Text mining, a combination of NLP with data mining (Tekin et al 2018), is commonly characterised as the transformation of unstructured or semi-structured textual data, into structured and normalised data sets suitable for analysis to uncover hidden patterns and information. The methodology evolved from the need to examine vast amounts of text including human language, which can then be mined for insights supporting data-driven decision-making (Pilipiec, 2022). Nowadays, text mining is a vast field of research with a variety of techniques each designed to meet the challenges given by specific situations. In relation to the project reported in this thesis, the classification of textual material, Text Classification (TC), is of interest.

3.1.1 Text Classification (TC)

Fragoudis et al. (2005) defined text classification as “the task of assigning one or more predefined categories to natural language text documents, based on their contents”. Text can be divided in several ways, depending on the number of labels associated with the text data: Single-labelled text classification describes a document that can only be assigned to one specific label. Binary text classification, also known as the two-class (positive or negative) method, assigns a document either to a certain predefined category (single label) or to the complement of that category (Joachims, 2002), whereas multi-class classification, refers to the circumstance in which each document is allocated a category from a collection of n classes (i.e., n > 2) (Garcia Constantino, 2013). A document can be associated with more than one label in the case of multi-labelled text classification. Classification does not require a computer, but manual classification is more time-consuming, laborious, and thus expensive to scale. Here machine learning-based text classification becomes fundamental as a time- and cost-saving approach, as in this case, a text classifier's set of rules, decision criterion in general, is automatically learned from the training data set. Nevertheless, human intervention is required as training data need to be labelled according to specific necessities (Manning et al. 2008; Garcia Constantino, 2013).

3.2 Project Design

3.2.1 Data Collection

Secondary data selection was used for this study to provide access to a data set that already contains a significant amount of information crucial for training the text classification models. Existing data in Excel format were pulled from electronic databases maintained by several pharmaceutical companies containing processed compliant-relevant information and medical queries. The raw data set contained a total of 10,993 text elements that were already divided into predefined classes concerning the topic they addressed, for instance, pharmacokinetic and pharmacodynamic properties, dosage adjustments for special patient populations, product complaints, requests for off-label use or reports of adverse events. A characteristic of such data sets is that they frequently include predefined classes according to internal operational necessities but are either not required for classification purposes or they can be summarised under a more appropriate umbrella term: For instance, pharmacodynamic and pharmacokinetic processes can be summarised with product complaints and stability data into the class "quality." For this task of data aggregation and re-classification, three focus group meetings between July and August 2022 were organised and attended by several pharma employees, who are all members of departments in charge of processing compliance-relevant information in their respective companies. As the quality of a supervised machine learning classifier depends on labelled training data, extensive human expert knowledge is therefore required throughout the workflow. Hence, it was necessary to create a new data set using a modified, pre-determined labelling standard which led to a final data set consisting of 1548 entries.

3.2.2 Research Instrument: R Program Language

According to the Online Historical Encyclopaedia of Programming Languages (HOPL Website, n.d.), up to date around 8,945 coding languages have been created. While between 250 and 2,500 coding languages are in use today, only a few are the most used and common ones, among them python, R, Java, JavaScript, C, C++, Go and PHP. Each language is created with a particular platform, operating system, coding approach, and use in mind.

The programming language used for this project was R since it provides several ready-to-use features that automate most text classification procedures. It is often the preferred language by quantitative analysts and data scientists as it offers a broad range of statistics-related libraries, outperforming python in terms of statistical support, and offers a supportive environment for statistical computing and design, by displaying enhanced computational power and visualization capabilities (Kuhn et al. 2016). Influenced by two existing languages, namely *Scheme* created by Steel and Sussman in 1975 and *S* created by Becker, Chambers and Wilks in 1985, the advantageous features of the syntax and vector-based data type of S, and the semantics of Scheme were combined and further developed to create a new language named R (Ihaka & Gentlemen 1996). R is available as a free software environment conceived and designed for statistical computing and graphics. Depending on the operations required, R users can install packages, which are extensions to the R statistical programming language, containing code, data, and documentation in a standardised collection format. There are two categories of packages: BASE packages come with the R download and installation, thus being already available, while contributed or third-party packages need to be downloaded, installed, and loaded separately from storages called repositories. Even though there are local repositories owned by individuals or companies, most are online and open to anyone. The top R package repositories are called CRAN, CRANTASTIC!, Bloconductor, and Github. For the project, all packages needed have been downloaded from the official repository CRAN (Comprehensive R Archive Network Repository), which is a global network of ftp (file transfer protocol) and web servers managed by the R foundation and run and maintained by the R community. For a package to be published here it must pass numerous verification checks for compliance with CRAN regulations. To present more than 9000 packages are available. R operates on a broad range of UNIX platforms, Windows and MacOS. Of important note is that R is subject to regular updates: The version used for this project was the "R version 4.2.2 (2022-10-31 ucrt) – 'Innocent and Trusting', Copyright 2022, The R Foundation for Statistical Computing, Platform: x86_64-w64-mingw32/x64 (64-bit)" (R project Website, n.d.) on a Windows Operating system. The R software was downloaded from the R project website via the (nearest) Austrian CRAN mirror, from the university of economics Vienna (CRAN mirror - Wirtschaftsuniversität Wien, n.d.). Further, R-Studio is an open-source, integrated development environment (IDE) used for R and Python (and other programming languages) for a more user-friendly environment thus facilitating code development. R-Studio Desktop was downloaded from the official website on a Windows Operating System (RStudio Desktop Website, n.d.). All packages were downloaded and installed from CRAN using the command `install.packages()` in R, and the command `library()` to read and operate with the package. For this project, the following packages were downloaded:

Package `Tidyverse`, version 1.3.2 published 18.07.2022 by Hadley Wickham, comprises a set of packages that work in harmony because they share common data representations and 'API' design. This package is designed to make it easy to install and load multiple "tidyverse" packages in a single step, such as ggplot2 for graphics, dplyr for data manipulation, tidyr for tidying data, readr for reading different document formats, stringr for string manipulations, tibble for data frames, other core tidyverse packages are purr, which provides toolsets for working with functions and vectors, and forecast for solving factorization problems of variables (Wickham et al. 2019; Package tidyverse, n.d.).

Package `tm`, version 0.7-10 published 13.12.2022 by Ingo Feinerer, is an R interface containing several text mining applications, such as tokenization, tm_map functions, removal functions (e.g., words, punctuation, numbers), weight functions (e.g., TF-IDF), VCorpus and Document

Term Matrix (DTM) among other (Feinerer, 2022). tm: Text Mining Package. *R package* version 0.7-10 (Package tm, n.d.).

Package `SnowballC`, version 7.0 published 01.04.2020, by Milan Bouchet-Valat, is an R interface to the C 'libstemmer' UTF-8 library that implements Porter's word stemming algorithm for collapsing words to a common root to aid the comparison of vocabulary. Currently supported in 15 languages (Package SnowballC, n.d.).

Package `textstem`, version 0.1.4 published 09.04.2018, by Tyler Rinker, is an R interface containing tools for text stemming and lemmatizing, whereby stemming removes endings such as affixes, and lemmatization groups inflected forms together as a single base form (Package textstem, n.d.).

Package `caret`, for Classification And Regression Training, version 6.0-93 published 09.08.2022 by Max Kuhn, is an R interface that contains powerful functions to streamline the model training process for complex regression and classification problems. There are over 230 models included in the package including various tree-based models, neural nets, and deep learning among many others (Package caret, n.d.).

Package `fastText`, version 1.0.3 published 10.08.2022 by Lampros Mouselimis, is an R interface to the library for efficient learning of word representations and sentence classification. The library is explained in detail in the works of Armand Joulin et al. (2016a & 2016b) and Piotr Bojanowski et al. 2017 (Package fastText, n.d.).

3.2.3 Sample procedures: Data Set Pre-Processing

This first step of data preparation and text correction can be done with the help of regular expressions from the basic package in R that comes with the download and installation of the program. Since the data set partly consists of extremely sensitive and confidential data, a first manual pre-processing was performed as follows: First, all company-relevant, internal information, such as case numbers, names of employees processing the respective enquiry, as well as the type of information source, e.g., whether the enquiry was received via email, telephone or through social media, was removed. Further, since trade names are not decisive for category identification, all trade names were replaced with their corresponding substance class, such as “antibiotics,” “antibody therapy” or “antidepressant” to name a few examples, if combination preparations were mentioned, the term “combination therapy” was added to the substance class, e.g., “anti-hypertensive combination therapy.” The data set was then divided according to predefined main topics (categories) for this project and assigned to the respective department representatives for overview purposes. The main topics were based on the type of information that must be processed and stored by pharmaceutical companies in a compliance-relevant manner: Pharmacovigilance, Quality and Regulatory queries. In addition, three further main topics were identified, primarily medical enquiries, followed by marketing queries, and a sixth and final category commonly referred to as communication. The latter category comprised mostly comments collected from social media channel screenings of various companies including any potentially relevant information - praise or firestorm from patients and consumers - for the company's communication department, nowadays often renamed patient centricity department. During the data set processing, the data entries were reclassified according to the new categories and anonymised and duplicate entries were removed. After the first manual pre-processing, the data set contained the following number of text entries for each predefined category: 709 entries for Medicine, 575 entries for Pharmacovigilance, 103 entries for Quality, 69 entries for Communication, 68 entries for Regulatory and 24 entries for Marketing, and was ready to be further pre-processed using various natural language processing techniques.

Before the beginning of coding, other preparatory steps had to be carried out in R-Studio itself: After opening R and R-Studio, the necessary packages - because not already installed - were installed on the computer using the function `install.packages()`. To access the packages' functions, the function `library()` was used to load them into the R environment. As the second preparatory step work directory was set to provide R with the correct path to the folder where the files to be processed are located and new documents can be stored. The complete code sequence of this project can be found in the Appendix section of this document (7. Appendix).

To avoid factorization as strings, the option was set to false at the beginning of the code because columns containing characters (i.e., text strings) are by default converted into the factor data type when creating or importing a data frame. Depending on further analysis it might be beneficial to keep the text elements as plain text strings (character), as it is in this case. At his point, the prepared classified data set in excel format was uploaded from the work directory into the R code-writing environment with the help of the `read_excel()` function from the readxl package of tidyverse. The package supports both the `.xls` format and the modern xml-based `.xlsx` format. There is no need to import this package separately, as it is part of the tidyverse. Further, the data set's object was renamed - in this case `mr` - to facilitate further coding. The next step was to utilise the `trimws()` function, which is used to remove leading white spaces from the column "classification" of the data set: The outermost rows and columns with the same values were removed to reduce an object's size. The function `view()` allows to visualise the resulting table, while the `dim()` function returns the number of dimensions, i.e, the number of rows and columns present in a data frame. In this case, the console pane shows a data set consisting of 1548 rows, i.e., the number of text strings, and two columns, i.e., the classification column containing all pre-defined categories and the content column with the text strings. To view the number of classes in the data frame the function `table(mr$classification)` can be utilised. This function allows to get a first view of the number of text strings assigned to the pre-defined categories.

```
mr$classification <- trimws(mr$classification)
View(mr)
dim(mr)
[1] 1548    2
table(mr$classification)
```

Communication	Marketing	Medicine	Pharmacovigilance	Quality	Regulatory
69	24	709	575	103	68

As the next step, a seed was set: The function `set.seed(value)` specifies a random number as the beginning value, in this case, the random integer value was set to 123. More specifically, pseudo-random numbers are produced when “random numbers” are generated in R and an algorithm used to produce these numbers needs a seed to initialise it. Being pseudo-random as opposed to completely random allows prediction and replicability of a certain outcome knowing the seed and the generator. Hence, the primary goal of utilising `set.seed()` is to be able to replicate a certain sequence of “random” numbers to enable result replicability. Before developing a statistical or machine learning model, it is usual to divide a data set into training and testing sets. There is no clear direction, however, on how much data should be used for training and testing, respectively. A typical ratio is 80:20, which states that 80% of the data is for training and 20% is for testing and derives its justification from the well-known Pareto principle (Joseph, 2022). To create a train and test training set, a new object, i.e., `mr$type`, was created to divide the data set `mr` according to the Pareto principle into 80% train- and 20% test subsets.

A contingency table is a type of frequency distribution table that simultaneously presents the association between two categorical variables. R uses the `table()` function to build contingency tables.

```
# check in the contingency table the distribution of the brands over
Type variable
table(mr$classification, mr$Type)
```

	test	train
Communication	14	55
Marketing	2	22
Medicine	145	564
Pharmacovgilance	103	472
Quality	20	83
Regulatory	17	51

After data preparation in R, the data set has not yet been converted into a computer-readable form. Following the initial step of data preparation, data set pre-processing aims to eliminate unnecessary text fragments that could mask important patterns, decrease classification performance, and introduce redundancy into the analysis. Out of several possibilities in this case pre-processing has been applied as follows in R:

The `VCorpus()` function is used to generate a corpus from the data source. A Corpus, which is a collection of text organised into databases that need to be evaluated, is the primary organisational structure for managing documents in the `tm` package. A corpus, or plural corpora, representing R objects are fully maintained in memory, whereas a Volatile Corpus, or `VCorpus`, is lost once the Corpus is removed (Kobayashi et al. 2018). To look at the corpus the `inspect()` function can be used to see the content and character number of the individual text strings, in this case, string number 1.

```
inspect (corpus[[1]])
<<PlainTextDocument>>
Metadata:  7
Content:  chars: 38
I would like to work for your company.
```

Once the documents, in this case, the text strings, have been gathered in the corpus, several pre-processing techniques can be used. The `tm` package includes many widely used pre-processing techniques (Kobayashi et al. 2018).

3.2.4 Text Transformation

Upon Corpus creation, its data can be subjected to a variety of pre-processing techniques (NLP techniques): They consist of lower-case conversion, number removal, punctuation removal, unnecessary whitespace removal etc. In the `tm` package, this functionality is encompassed in the concept of text transformation. The `tm_map()`function applies (maps) a given pre-processing step to each element, i.e., individual text strings, of the corpus to perform a transformation (Feinerer, 2022). Of important note is, that the pre-processing steps must be carefully selected for the individual project, as they can significantly impact results. All pre-processing steps can be bound together with other related operations using the `%>%` operator to create a DocumentTermMatrix`(dtm)` in one go, each for the training and the test data set, respectively.

Any character processing operation can be used for transformation if it results in a text document; `content_transformer()`, for instance, offers tools to access and modify a document's content. Of note, text manipulation functions in basic R, such as gsub(), can also be utilised directly. Lower case conversion, `tolower()`, is important, as computers handle the same word written in with lower- or upper-case letters, e.g., “cough” or “Cough,” as two distinct words, which can impact word count results. The functions, `removePunctuation()` and `remov–Numbers()`eliminate any punctuation and numbers respectively from the text as they do not inflict results. Further, in handling text written on social media, there is a strong need to remove any extra whitespace from the text data, which can be accomplished using the `stripWhitespace()` function. The other functions `removeWords(stopwords)` and stemming, `stemDocument()`, are already described as examples of basic NLP techniques in section 2.5.1. The output of such rule-based pre-processing is a cleaned data set comprised of a volatile corpus, ready for further analysis.

Vector Space Models (VSM) are the simplest and possibly most straightforward technique of text feature quantification and thus serve as an ideal starting point in the use of text classification (Kobayashi et al. 2018). The `DocumentTermMatrix()` function of the `tm` package can be used for this transformation to create features depending on the specific terms in the corpus. TermDocumentMatrix and DocumentTermMatrix – depending on whether using terms as rows and documents as columns or vice versa - use sparse matrices for corpora. For typical-sized data sets, document-term matrices frequently grow to enormous sizes. Hence, the application of the `removeSparseTerms` function eliminates rarely mentioned terms, thereby significantly reducing the matrix's size without sacrificing key relationships between the variables (Feinerer, 2022). This function eliminates terms of a given percentage of sparse elements, i.e., no appearance in a document. Further, for memory-saving reasons, the function `inspect(dtm)` allows to view a sample of the matrix without calling the full data set.

```
inspect(dtm)
<<DocumentTermMatrix (documents: 1548, terms: 111)>>
Non-/sparse entries: 8407/163421
Sparsity           : 95%
Maximal term length: 10
Weighting              : term frequency - inverse document frequency
(normalized) (tf-idf)
Sample             :
 Terms
```

Docs	can	data	drug	inform	medic	patient	request	tablet	therapi
1150	0.0000000	1.99074	0.000000	0	0	0	0.000000	0	0.0000000
1255	0.0000000	0.00000	0.000000	0	0	0	0.000000	0	0.0000000
1498	0.0000000	0.00000	1.202502	0	0	0	0.000000	0	0.0000000
1525	0.0000000	0.00000	1.202502	0	0	0	0.000000	0	0.0000000
181	0.0000000	1.99074	0.000000	0	0	0	0.000000	0	0.0000000
213	0.5049455	0.00000	0.000000	0	0	0	0.000000	0	0.4451005
371	0.0000000	0.00000	0.000000	0	0	0	1.907415	0	0.0000000
437	0.5049455	0.00000	0.000000	0	0	0	0.000000	0	0.0000000
664	0.0000000	0.00000	0.000000	0	0	0	0.000000	0	0.0000000
95	0.6311818	0.00000	0.000000	0	0	0	0.000000	0	0.0000000

```
dim(dtm)
[1] 1548  111
# understanding the dtm: the number of dimensions changes based on
stemming or lemmatizing
```

Machine learning algorithms frequently require numerical data for processing, thus when dealing with textual data or any natural language processing (NLP) operations, data must first be vectorised, i.e., converted to numerical data or vectors. One possibility is to look at the number of times a word appears in a document (raw count). For instance, the raw term frequency count (TF) is the default weighting selection for the DocumentTermMatrix() function. Another typical text analysis method is to search for keywords in documents using the TF-IDF method: The concept of Term Frequency (TF) – Inverse Document Frequency (IDF), as mentioned in section 2.5.2 Machine Learning-based NLP, states that the frequency of a term in documents is inversely correlated with its relevance across the documents (corpus). While a term's frequency in a document is revealed by TF, its relative rarity within the corpus of texts is revealed by IDF, thus reducing the weighting of frequent terms and the impact of infrequent terms, which can be more important to reveal the topic of a document, can be increased. The final TF-IDF value results by multiplying these two values together. The word with the highest TF-IDF weight serves as the document's keyword. The result after statistic-based pre-processing steps was a constructed `dtm`, serving as the basis for further analysis. For computational convenience, the `dtm` was transformed into a plain matrix, before being converted into a data frame.

With the function `bind_cols()` the two columns “classification” and “type” were added to the `dtm`. The last preparatory step was to subdivide the data set into a train and test subset for the classification algorithms, in this case, Linear Discriminant Analysis (LDA) and K-Nearest-Neighbours (kNN) as explained in the next section.

3.3 Text Classification Algorithms

There is no one “best” classification algorithm that works well with all data mining tasks. Possible reasons can be found in the individuality of the data set due to (i) the datatype, (ii) its size, and (iii) the distribution and number of classes among the data sets. Over the past decades, data mining research has produced a variety of strategies that may be suitable for diverse situations (see sections 2.3 and 2.4), such as data sets comprised mostly of numeric values rather than text strings, small data sets, or data sets featuring large numbers of classes. To become familiar with the topic, as well as after inspection of the data set, three different methods were chosen, the Linear Discriminant Analysis (LDA), k-Nearest-Neighbour (kNN) and fastText, which will be elucidated in more detail below. These methods are well-known and widely used classification algorithms with promising performance for classification rates (e.g., low false positive rate) as well as CPU speed and memory usage for the training and classification phases.

As the last step before applying the data set to the LDA and kNN classifiers a validation scheme was set, to deal with so-called overfitting. Over-fitting is a risk in every learning approach: The model may fit the example data flawlessly but produce poor predictions for new, previously unseen examples. This occurs due to the possibility that the training data may teach it random noise as opposed to only the necessary, desired features. Further, procedures like cross-validation, which divides the training data set randomly into training and test sets to internally evaluate the model's predictions, reduce the risk of overfitting. Cross-validation aims at calculating the potential practical accuracy of a predictive model, as small data sets especially can lead to too optimistic prediction error rates. Cross-validation can be performed in different ways, including leave-p-out and k-fold cross-validation (Manning et al. 2008). In this case, a 10-fold cross-validation was performed.

```
# set the validation scheme, 10-fold cross-validation
ctrl <- trainControl(method="cv", number = 10, classProbs = TRUE)
```

At this point, the data set was ready to be applied to the algorithms LDA and kNN (see Appendix 1-3 for the complete code line).

3.3.1 Linear Discriminant Analysis (LDA)

Originally described by Ronald A. Fisher in 1936, also known as Fisher's Discriminant Analysis, the Linear Discriminant Analysis was originally conceived as a two-class method (Fisher, 1936). Later, Calyampudi R. Rao developed a generalised, multi-class version of Fisher’s Discriminant analysis, calling it multiple discriminant analysis (Rao, 1948), although collectively known as the Linear Discriminant Analysis. Linear discriminant analysis is a generative linear model -it uses joint probability distribution- for dimensionality reduction and for supervised classification of two or more classes, which works on continuous variables. Under the assumption that all classes are linearly separable, LDA calculates numerous linear discrimination functions also called hyperplanes, i.e., decision boundaries, in the feature space to separate the classes: For n-classes, LDA draws n minus 1 (n-1) hyperplanes projecting data for classification. The requirements that are taken into consideration when creating these hyperplanes are to (i) maximise between-class-variance, (ii) minimise within-class-variance and (iii) project a data set into a smaller dimensional space with strong separable classes (Chen et al. 2020):

- Between-class-variance, calculated by the distance between the means of different classes,
- Within-class-variance, calculated by the distance between each class’s mean and sample and,
- Creation of a lower-dimensional space that maximises the variance between classes and reduces the variance within classes, resulting in the highest possible class separability.

Further, this methodology prevents challenges such as over-fitting and cuts down on computational costs (Chen et al. 2020).

3.3.2 k-Nearest Neighbour (kNN)

The simple, and best-known nearest neighbour method is the k-Nearest Neighbour classification, a non-linear model for classification that has been extensively studied since its development (Cover & Hart 1967). kNN makes use of proximity for classifying and predicting the groupings of single data points. Although the algorithm can be applied to both classification and regression issues, it is commonly employed as a classification algorithm because it relies on the assumption that similar data points can be located in proximity to one another. The similarity is provided by representing each instance (i.e., data point) as a point in an n-dimensional space, where an unseen instance is classified based on the nearest classified instances. The Euclidean distance is typically used to calculate the distance between points; other popular distances are the Manhattan distance, the Minkowski distance, or the Hamming distance (Chen et al. 2020). The kNN method categorises unknown examples in a feature space according to how closely they resemble their nearest training instances. The k value in the kNN algorithm specifies the number of neighbouring data points (i.e., neighbours) that will be examined to determine a particular classification. As an illustration, if k=1, the instance will be put in the same class as its one nearest neighbour. Defining k can be a balancing act because values that are too high or too low can lead to overfitting or underfitting. Greater values of k can thereby result in strong bias and low variance, while smaller values of k can imply high variance but low bias. The input data have a significant impact on the choice of k since data with more noise or outliers generally perform better with larger values of k. While computationally more expensive than other methods, kNN is employed in this thesis for testing reasons because of its ongoing popularity (Manning et al. 2008).

3.3.3 FastText

The Facebook AI Research lab (FAIR) created the open-source library fastText for rapid and accurate processing of enormous data sets while finding scalable solutions for text representation and text classification (fastText Website, n.d.). As a word embedding technique, it resembles word2vec, however, it shows one major distinction: The fastText model can be thought of as a shallow neural network (containing one hidden layer) that gains its capabilities by increasing the number of learnable vector (word) embeddings using character n-gram information supplied into the network. The word representations (weight matrix) are then averaged into a text representation, i.e., a hidden variable, which is subsequently supplied to a linear classifier (I.e., multinomial logistic regression). The probability distribution over the predefined classes is computed using the hierarchical softmax function in cases of many classes. The difference to a “classic” linear classifier is, that the hierarchical softmax function uses a binary tree for label representation, whereby every node of the binary tree is representative of a probability. Thus, a label “x” is represented by the given path of probabilities to that given label “x.” Instead of computing probability scores of a given text element for all labels in search for the highest score, fastText calculates the probability on each node along the path to the correct label, thus vastly decreasing the number of computations for each text element in a document and increasing speed in presence of multiple labels. The detailed model architecture is described elsewhere (Joulin et al. 2016a; Joulin et al. 2016b; Bojanowski et al. 2017).

The packages used in support of the fastText model are `tidyverse`, `textstem` and `fastText`. As with the previous models, fastText also requires data pre-processing, especially because the data must be converted into a readable format for fastText. As the first step, the data set was uploaded into the R environment via the `read_excel()` function and renamed `mr` to facilitate further coding. The next step consisted of creating a second data set, called `mr2`, in which all necessary pre-processing steps were performed. For such data manipulation, the `mutate()` function comprised in the dplyr package is used (see section 3.2.2 for package description). Among others, this function is specifically used to add new variables to a given data frame while keeping the existing ones. As for `read_excel()` there was no need to import the package separately, as it is part of tidyverse. Alternatively to the remove functions used for the LDA and kNN data pre-processing steps, in this case, substitution and replace functions were used to manipulate single characters in a text string: While `sub()` and `gsub()`, i.e., substitution and group substitution respectively, are R basic functions, `str replace()` and `str replace all()`, i.e., string replace, come with the stringr package comprised in tidyverse (see section 3.2.2 for package description). Thereby unnecessary characters were all replaced by other characters or removed as needed. The string mutations comprised lowercase transformation (tolower), removal of additional or invalid WhiteSpace ("[\r\n\t]", “x000D_" and " " to " respectively), removal of numbers ("[0-9]", ""), words (stopwords), punctuation ("[[:punct:]]", "") and stemming (stem_words). Additionally, as required for fastText, at the beginning of each text string the predefined label was added using `paste0("__label__", classification)` and both columns of the data set, classification and content, unified to a single “text” column by applying `text=paste(classification, content)`, further the row numbers were then changed to `id`. The last step comprised the selection of the relevant columns, “id” and “`text`,” of the newly created data frame. The `print()` function allows a view into the data frame; the example displays a selection of text strings for each predefined label. For sample selection, one example for each label has been extracted.

```
print(mr2)


 1 "__label__communication I would like to work for your company"

 42 "__label__marketing Please send me the product information"

 111 "__label__medicine Can the administration of migraine medication
 trigger an epileptic seizure"

 1210 "__label__pharmacovigilance Occurrence of actinic keratosis un-
 der NSAIDs"
```

```
1431 "__label__quality A tablet in the blister is broken"

1498 "__label__regulatory Why is the drug no longer available"
```

After `set.seed()`the data was divided into 80% training and 20% test subsets. The `anti_join()` function in the test subset returns all rows in the data frame, which are not matching in the train subset (80%). The next step was to create a `train`, `test`, and `alldata` objects containing only the pre-processed text strings to facilitate further analysis. Data from spreadsheets or database programs are frequently imported into analytics using delimited text files; Each row from the given data set becomes a row in the delimited text file, with a line separator (delimiter) between each row, and each row contains one text string. After these steps, the data frame was ready to be fed to the fastText algorithm (see Appendix 4 for the complete code line).

3.4 Evaluation Metrics for Text Classification

A variety of evaluation measures are reported to assess how effectively a classification technique performed, including Accuracy, Recall/Sensitivity, Precision/Confidence, Specificity and Cohen's Kappa. As all terms (except for Cohen's Kappa) are calculated in relation to the confusion matrix, the latter is explained first along with its related values (Kulkarni et al. 2020):

Confusion Matrix: Confusion matrices are commonly used for attempting to solve classification problems. Both binary classification and multiclass classification problems can be addressed with it. In the field of machine learning a confusion matrix, also called error matrix, is a table that is used to evaluate the performance of the supervised learning algorithm. The actual classes are listed along the Y-axis (the rows), while the anticipated classes are listed along the X-axis (the columns) - both variants are documented in the literature (Powers, 2020). True Positive (TP) and True Negative (TN), represent the number of elements classified correctly as positive or negative respectively, False Positive (FP) and False Negative (FN), represent the number of elements wrongly classified as positive or negative, respectively.

Accuracy: Accuracy is the most basic and straightforward of the evaluation measures. The percentage of successfully categorised examples determines the accuracy of a classifier, basically the quality indicator of a classification algorithm. However, despite being conceptually simple and straightforward to calculate, the accuracy metric might be deceptive since it ignores the distribution of the classes. Therefore, it is good to practise utilising at least one additional evaluation metric in addition to accuracy to present a comprehensive, accurate picture of a classifier's performance (Kulkarni et al. 2020).

Recall/Sensitivity: Recall, also known as sensitivity, describes the percentage of Real Positive cases that are accurately predicted Positive, or the “True Positive Rate” (TPR). In Machine Learning and Computational Linguistics (where the emphasis is on how confident we can be in the rule or classifier), recall is often overlooked or averaged away (Kulkarni et al. 2020).

Precision/Confidence: On the other hand, precision, or confidence, as they are referred to in data mining, refers to the percentage of predicted positive cases that are actually real positives. Machine Learning, Data Mining, and Information Retrieval are all concerned with this. In an analogous manner, it can be referred to as True Positive Accuracy (TPA), which is a measurement of the accuracy of Predicted Positives in contrast to the rate of discovery of Real Positives (TPR) (Kulkarni et al. 2020).

Specificity: Even though both measures capture to some extent information regarding the rates and types of errors made, they primarily concentrate on the positive cases and predictions. Neither of them, however, displays information regarding negative cases. Inverse Recall or Specificity assesses the percentage of true negatives that are accurately categorised as negatives, often known as the “True Negative Rate” (TNR) (Kulkarni et al. 2020).

Cohen's Kappa (κ): Developed in 1960 by Jacob Cohen, the Cohen’s kappa coefficient is a statistical measure to evaluate inter-rater reliability for qualitative (categorical) items (Cohen, 1960). It ranges from -1 to +1, with -1 indicating no rater agreement to a given classification, to +1 indicating full rater agreement. 0 indicates that the raters agreed precisely as frequently as they would have if they had both made random guesses. Further, Cohen suggested further ranges of interpretation: κ 0.01-0.20 indicate no to slight agreement, κ 0.21-0.40 indicate fair agreement, κ 0.41- 0.60 indicate moderate agreement, κ 0.61-0.80 indicate substantial agreement and 0.81-1.00 denote almost perfect agreement. Cohen’s kappa is easily adaptable to assess agreement on more than two labels, as is often the case in text classification (McHugh, 2012).

3.5 Data Analysis

Although the data collection was simplified using secondary data collection and its pre-classification, the new category assignment was complicated by the multiple interpretations of the individual text data. Months and years after the initial processing, the entries were often no longer clearly assignable. Thus, an overall agreement had to be found in the focus group as to when, for example, an entry no longer belonged to “Medicine” but to “Pharmacovigilance.” This also gave the group a deeper understanding of the complexity of the task for machine learning and neural network algorithms.

At first glance, the final data set appears to be unbalanced, as more weight is given to “Medicine” and “Pharmacovigilance” compared to the other two compliance-relevant topics “Regulatory" and "Quality". Nevertheless, it was decided to continue with the data set as such, especially since (i) the data set should not be reduced to adjust the number of text entries across the classes, (ii) duplications to increase the size of the data set should be avoided and (ii) because the data set reflected real-life conditions: indeed, most of the requests that reach a company are of medical or PV nature.

3.6 Conclusion

At the beginning of this project, a comparative study of different algorithms was not foreseen, as the first choice immediately fell on the neural network-based fastText. However, as a first dive into AI and machine learning algorithms for text classification, it was important to get a broader overview of different NLP techniques and models, which more than justified the use of the other two older, but widely used algorithms LDA and kNN. The next chapter displays how the algorithms cope with the data set, as well as the models' accuracy and overall performance in the training and test phase.

4 RESULTS AND DISCUSSION

4.1 Introduction

After all pre-processing steps were performed on the data set by using various NLP methods, from rule-based methods, such as lowercase conversion, removal of numbers, punctuation, stopwords and whitespace, to statistical methods, such as TF-IDF as discussed in the previous chapter, the data set is now ready to be fed into the three different algorithms. For LDA and kNN the last preparatory step was to set a validation scheme to increase algorithm validation, while that was not necessary for fastText due to the different algorithm approach as discussed in the previous section.

4.2 Visual Presentation of the Findings

4.2.1 Linear Discriminant Analysis (LDA)

After setting the validation scheme, the next step was to train the LDA algorithm using the train data set contained in the `dtm`. `Set.seed(seed)`, whereby seed was set before as random number 123, was applied again to ensure replicability of the model's results. The LDA performance on the training data can be inspected by simply writing the model's name `lda` (see Appendix 2 for the complete code line).

```
# train the data set with the lda method on the column classification,
using the dtm_train data set

set.seed(seed)
lda <- train(as.factor(Classification) ~ . , data = dtm_train, method =
"lda", trControl = ctrl, metric= "Accuracy", tuneLength = 5, na.action
= na.pass)

# check the model
lda

Linear Discriminant Analysis
1247 samples
111 predictor
6 classes: 'Communication', 'Marketing', 'Medicine', 'Pharmacovigi-
lance', 'Quality', 'Regulatory'
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1122, 1122, 1121, 1122, 1122, 1124, ...

Resampling results:
 Accuracy     Kappa
```

```
0.6041085  0.364645
```

With a data set of six classes, i.e., Communication, Marketing, Medicine, Pharmacovigilance, Quality and Regulatory, the 10-fold cross-validation on the training data set resulted in an accuracy of 60%, with a Cohen's Kappa value of 0.36, indicating "fair" agreement. The prediction performance of the lda model on the test data set can be viewed by applying the `confusion-Matrix()` function followed by the `print()` function.

```
# predict on test data
lda.predict <- predict(lda, newdata = dtm_test)

# check model performance with the test set
cfm <- confusionMatrix(lda.predict, as.factor(dtm_test$Classification)

print(cfm)
Confusion Matrix and Statistics
```

	Reference					
Prediction	Communication	Marketing	Medicine	Pharmacovigilance	Quality	Regulatory
Communication	4	0	3	1	1	0
Marketing	2	0	4	0	0	1
Medicine	7	1	102	34	11	10
Pharmacovigilance	1	1	26	64	0	2
Quality	0	0	6	3	6	2
Regulatory	0	0	4	1	2	2

```
Overall Statistics
Accuracy : 0.5914
95% CI : (0.5335, 0.6474)
No Information Rate : 0.4817
P-Value [Acc > NIR] : 8.72e-05
Kappa : 0.3431
Mcnemar's Test P-Value : NA
```

	Class:	Communication	Marketing	Medicine	Pharmacovigilance	Quality	Regulatory
Sensitivity		0.28571	0.000000	0.7034	0.6214	0.30000	0.117647
Specificity		0.98258	0.976589	0.5962	0.8485	0.96085	0.975352
Pos Pred Value		0.44444	0.000000	0.6182	0.6809	0.35294	0.222222
Neg Pred alue		0.96575	0.993197	0.6838	0.8116	0.95070	0.948630
Prevalence		0.04651	0.006645	0.4817	0.3422	0.06645	0.056478
Detection Rate		0.01329	0.000000	0.3389	0.2126	0.01993	0.006645
Detection Prevalence		0.02990	0.023256	0.5482	0.3132	0.05648	0.029900
Balanced Accuracy		0.63415	0.488294	0.6498	0.7349	0.63043	0.546500

In relation to the six categories, the linear discriminant analysis achieved the following True Positives (TP): Medicine 102, Pharmacovigilance 64, Quality 6, Communication 4, Regulatory 2 and Marketing 0. As expected, the sensitivity scores for the two labels "Medicine" and "Pharmacovigilance" are the highest, with values of 0.7034 and 0.6214 respectively, and the lowest for Regulatory and Marketing with 0.1176 and 0.00 respectively. Thus, the specificity, also defined as True Negative Rate (TNR), is also lowest for these two labels, with 0.5962 and 0.8485 respectively, and the highest for "Regulatory" (0.9753), "Marketing" (0.9765) and "Communication" (0.9825). The balanced accuracy for each label reached the highest percentage for "Pharmacovigilance" and "Medicine" with approx. 74% and 65% respectively, the lowest percentage was achieved by "Regulatory" (≈55%) and "Marketing" (≈49%); the overall model accuracy over all labels was 59% (95% CI: 0.5335, 0.6474) with a κ of 0.34 indicating "fair" agreement.

4.2.2 k-Nearest Neighbour (kNN)

Resembling the procedure for the LDA model, after setting the validation scheme, the next step was to train the kNN algorithm using the training data set contained in the `dtm` preceded by `set.seed(seed)` for result replicability. To view the algorithm's performance on training data, the code `knn` was applied (see Appendix 3 for the complete code line).

```
# train the data set with the lda method on the column classification,
using the dtm_train data set

set.seed(seed)
knn    <-    train(as.factor(Classification)    ~    .,    data    =
data.frame(dtm_train), method = "knn", trControl = ctrl, metric="Accu-
racy", tuneLength = 5)

# check the model
knn

k-Nearest Neighbours
1247 samples
111 predictor
6 classes: 'Communication', 'Marketing', 'Medicine', 'Pharmacovigi-
lance', 'Quality', 'Regulatory'
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1122, 1122, 1121, 1122, 1122, 1124, ...

Resampling results across tuning parameters:

k      Accuracy      Kappa
5     0.5309888     0.2146491
7     0.5205959     0.1852585
9     0.5125439     0.1584476
11    0.5204662     0.1668810
13    0.5100725     0.1457563
Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

The non-linear classifier achieved the highest accuracy level with 53%, which was used to select the optimal operating model for the given data set, determining a k of 5, i.e., assigning a label to a given text string by observing its 5 closest neighbours. Cohen's Kappa value of 0.21, indicates "fair" agreement. The prediction performance of the kNN model on the test data set can be viewed by applying the `confusionMatrix()` function followed by the `print()` function.

```
# predict on test data
knn.predict <- predict(knn, newdata = data.frame(dtm_test))

# check performance with the test set
cfm <- confusionMatrix(knn.predict, as.factor(dtm_test$Classification))
print(cfm)
Confusion Matrix and Statistics
```

	Reference					
Prediction	Communication	Marketing	Medicine	Pharmacovigilance	Quality	Regulatory
Communication	4	0	4	4	0	1
Marketing	0	0	0	0	0	0
Medicine	9	2	104	49	13	10
Pharmacovigilance	1	0	27	47	2	4
Quality	0	0	10	2	5	2
Regulatory	0	0	0	1	0	0

```
Overall Statistics
Accuracy : 0.5316
95% CI : (0.4735, 0.589)
No Information Rate : 0.4817
P-Value [Acc > NIR] : 0.04724
Kappa : 0.2222
Mcnemar's Test P-Value : NA
```

	Class:	Communication	Marketing	Medicine	Pharmacovigilance	Quality	Regulatory
Sensitivity		0.28571	0.000000	0.7172	0.4563	0.25000	0.000000
Specificity		0.96864	1.000000	0.4679	0.8283	0.95018	0.996479
Pos Pred Value		0.30769	NaN	0.5561	0.5802	0.26316	0.000000
Neg Pred alue		0.96528	0.993355	0.6404	0.7455	0.94681	0.943333
Prevalence		0.04651	0.006645	0.4817	0.3422	0.06645	0.056478
Detection Rate		0.01329	0.000000	0.3455	0.1561	0.01661	0.000000
Detection Prevalence		0.04319	0.000000	0.6213	0.2691	0.06312	0.003322
Balanced Accuracy		0.62718	0.500000	0.5926	0.6423	0.60009	0.498239

The k-Nearest-Neighbour analysis achieved the following True Positives (TP): Medicine 104, Pharmacovigilance 47, Quality 5, Communication 4, Regulatory 0 and Marketing 0. As expected, the sensitivity scores for the two labels "Medicine" and "Pharmacovigilance" are the highest, with values of 0.7172 and 0.4563 respectively, and the lowest for Regulatory and Marketing both with 0.00. Specificity is lowest for "Medicine" and "Pharmacovigilance" labels, with 0.4679 and 0.8283 respectively, and the highest for "Regulatory" (0.9964), "Marketing" (1.000) and "Communication" (0.9686). The balanced accuracy for each label reached the highest percentage for “Pharmacovigilance” and “Communication” with approx. 64% and 62% respectively. Interestingly, “Medicine” achieved among the lowest percentages for label accuracy with ≈59%; the overall model accuracy over all labels was 53% (95% CI: 0.4735, 0.589) with a κ of 0.22 indicating “fair” agreement.

4.2.3 FastText

The third algorithm to be used for this project was the fastText model. By default, the word vectors will account for character n-grams between 3 and 6 characters. The fastText package for R provides different functions which can be adapted to a specific need, depending on the type of data set and the instance in which the model must be trained. In this case, for the provided data set the command “supervised” is chosen (Mouselimis, 2022).

```
Read 0M words
Number of words:  4696
Number of labels: 6
Progress: 100.0% words/sec/thread:  546702 lr:  0.000000 loss:  0.641651
ETA:   0h 0m
Elapsed time: 0 hours and 0 minutes and 2 seconds.
```

Of important note is, that the fastText functions have mandatory and optional arguments. The arguments input and output are always mandatory, while others, such as learning rate `lr`, the size of word vectors `dim`, verbosity level (displays details about processing information) verbose, number of threads `thread`, and `MilliSecs`, specifying the time delay of printing the output file, are optional. The last two arguments set to TRUE are Booleans, whereby `remove_previous_file` deletes any existing file with the same output name if the path output is not an empty string (""), and `print_process_time` prints the model’s process time on screen. FastText needed two seconds of training time for 4696 words and six predefined labels.

Based on the output model, the subsequent command can be used to `predict` new data. This output will take the form '__label__pharmacovigilance' (in the case of the pharmacovigilance label) with each line denoting a new label (input and output data must be equal as in number of lines). The probability of the labels can also be obtained with the command `predict-prob`. In the sample selection, text strings were attributed with a given probability to the different labels. Interestingly, there were no mentions of test set elements attributed to the label “Marketing.”

```
__label__communication 0.255397
__label__quality 0.857385
__label__medicine 0.799403
__label__pharmacovigilance 0.899873
__label__regulatory 0.276361
```

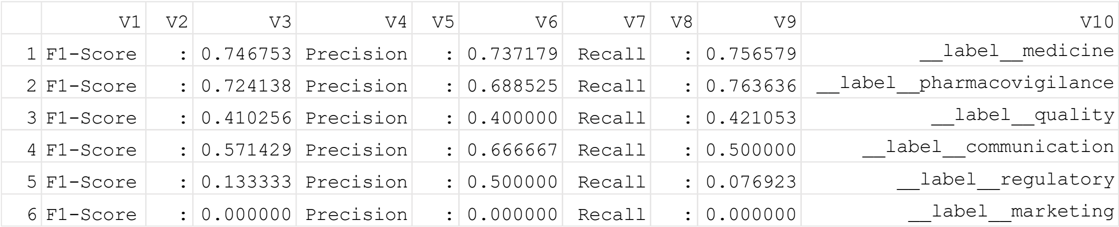
The model’s performance can be determined by calculating the `Precision` and `Recall` with the `test function` in this case at `k=1`, i.e., single most likely label, on a test set after the model has been trained (Mouselimis, 2022). The R session's metrics are only printed by the `test` command. For the fastText model, Recall is defined by the proportion of correctly predicted labels out of the correct labels, while precision is defined by how many accurate labels were predicted. For `k=1` Precision and Recall were both calculated at 69%.

```
# 'test' function
#----------------
N     310
P@1   0.690
R@1   0.690
```

The result of the `test-label` command, which is equal to the test function, can be saved to the `test label valid.txt` file, which contains the Precision and Recall for each distinct label on the data set (`alldata.txt`). This means that the test label valid.txt data set's number of rows must match the number of distinct labels in the `alldata.txt` data set. Using the code line `verify snippet`, this can be confirmed (see Appendix 4 for the complete code line).

```
# number of unique labels of data equal to the rows of the 'test_la-
bel_valid.txt' file
length(unique(res_pharma)) == nrow(test_label_valid)
[1] TRUE

head(test_label_valid)
```



	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	F1-Score	:	0.746753	Precision	:	0.737179	Recall	:	0.756579	__label__medicine
2	F1-Score	:	0.724138	Precision	:	0.688525	Recall	:	0.763636	__label__pharmacovigilance
3	F1-Score	:	0.410256	Precision	:	0.400000	Recall	:	0.421053	__label__quality
4	F1-Score	:	0.571429	Precision	:	0.666667	Recall	:	0.500000	__label__communication
5	F1-Score	:	0.133333	Precision	:	0.500000	Recall	:	0.076923	__label__regulatory
6	F1-Score	:	0.000000	Precision	:	0.000000	Recall	:	0.000000	__label__marketing

The F1-score is an alternative to the more widely used arithmetic mean and is defined as the harmonic mean of precision and recall. It is frequently utilised when calculating an average rate. The highest F1-score was achieved for the label “Medicine” with a 75% probability of text elements being correctly attributed to the label, followed by ”Pharmacovigilance” with a 72% probability, “Communication” with 57%, “Quality” with 41%, and “Regulatory” with a low F1-score of 13%. For “Marketing” no scores could be calculated.

4.3 Comparative Analysis and Interpretation of the Results

In this project, three different, well-known, and widely used machine learning algorithms for supervised learning in text classification problems were investigated: The generative linear classifier Linear Discriminant Analysis (LDA), the discriminative non-linear classifier k-Nearest Neighbour (kNN) and the shallow neural network fastText.

FastText achieved the highest hit probability on the test set with 69% precision and recall, followed by the linear generative model LDA with 59%, while the non-linear kNN achieved the lowest overall accuracy with 53%. With regard to individual classes, "Medicine" and "Pharmacovigilance" have been classified best, which was to be expected as these classes had the highest number of text entries, namely 709 and 575 for “Medicine” and “Pharmacovigilance” respectively. For the class “Medicine”, both LDA and kNN correctly classified the highest number of text elements with 102 and 104 true positives, respectively, according to the confusion matrix. This is also reflected in the sensitivity, where “Medicine” has the highest value in both models, about 70% for LDA and 72% for kNN. Interestingly, this is not reflected in the balanced accuracy: The highest balanced accuracy in both models was calculated for “Pharmacovigilance” with approximately 74% for LDA and 64% hit probability for kNN. The sensitivity for “Pharmacovigilance” is, however, interestingly lower for kNN with 46% than for LDA with 62%. FastText shows the highest sensitivity scores for the labels “Medicine” and “Pharmacovigilance” both with approximately a 76% hit probability as well as the highest F1-scores with 75% and 72% respectively. In contrast to LDA and kNN, fastText also achieves a higher sensitivity/recall score for the classes “Communication” with a 57% and “Quality” with a 41% hit probability, although both LDA and kNN do not exceed 30% here.

Looking at the data set, it becomes clear that the number of text entries (1548) is very heterogeneously distributed across the six classes: Accordingly, 45.8% of the entries were attributed to the class "Medicine", the other 37.1% to "Pharmacovigilance", 6.7% to "Quality", 4.5% to "Communication", 4.3% to "Regulatory" and 1.6% to "Marketing".

In principle, requests for compliance-relevant information are received with varying frequency in the daily pharmaceutical business. Most of the requests are of medical nature and are based on the fact that doctors in particular ask for scientific information on products and diseases, either directly or through sales representatives - after discussions with doctors. Pharmacovigilance reports usually come directly from patients, who call the manufacturing company to request reassuring information on certain side effects, so-called spontaneous reports, and less from physicians. Whether these are medically verified or not, they must also be recorded. Quality and Regulatory enquiries are less frequent: Quality enquiries mostly concern pharmacists and hospitals, who enquire about alternative dosage forms for special patient cases or inquire about extended stability data, that is for instance if a product that has been stored outside the recommended parameters for a longer period is still usable or needs to be disposed of. Regulatory enquiries mostly concern the marketing authorisation status or availability status of a medicinal product in a certain country or the possibility of making the product available across borders for certain patient cases. Thus, the heterogeneity of the class distribution also results from the frequency of the enquiries received. Although the data reflect real-life conditions, the uneven distribution certainly impacted the efficiency of the algorithm. To compare the different classification models in the best possible way, the data set was pre-processed in the same way. Differences between the three models are most apparent considering the classes that had fewer text elements, namely “Quality”, “Regulatory”, and “Marketing”, with 20, 17, and 2 test entries, respectively. FastText could not deliver evaluation values for the label "marketing", which was to be expected given the low number of text entries, whereas it showed better performance overall given the same pre-processed data set for the other classes. kNN and LDA were able to deliver results for all classes, whereby LDA overall seemed to perform better, although marginal, than kNN.

It is commonly understood that non-linear classifiers are generally superior to linear classifiers (Manning et al. 2008): First, several nonlinear models include linear models as an exceptional case. For instance, kNN, a nonlinear learning technique, will occasionally result in a linear classifier. Second, nonlinear models can sometimes be less complex than linear models (e.g., with fewer dimensions). Third, regardless of whether the classifier is linear or nonlinear, the difficulty of learning does not affect the type of classifier. There are instead numerous aspects, e.g., (i) data pre-processing, and (ii) computational processes, such as regularization or margin maximization, that can make a learning method more or less effective.
However, this does not imply that nonlinear classifiers should always be used for statistical text classification: A key concept in machine learning is the bias-variance trade-off, which describes why there isn't a single, perfect learning strategy, and the choice of a suitable learning strategy is a necessary step in resolving a text classification challenge. Bias is an error that results from false presumptions made by the learning algorithm. An algorithm may fail to simulate the proper relationships between input and output if it has a significant bias (underfitting). Variance is the error calculated from the training data's sensitivity to minor fluctuations. Overfitting results from a high variance, which models the noise in the training data rather than the desired result. Linear models are known to have high bias and low variance, while non-linear classifiers tend to have low bias and high variance. Since most randomly generated training sets provide similar decision hyperplanes, the low computational variance could be an underlying reason for a slightly better LDA performance over kNN.

In addition to the heterogeneous data distribution over the classes, another point regarding the data set that should not be neglected is that queries do not always allow a clear, singular classification from the beginning. Natural language processing is a difficult problem mainly because of the richness, ambiguity, and expressive power of human language. Often a medical query hides a side effect report that must be forwarded to the pharmacovigilance department, quality requests can also hide medical questions. The clarity depends not only on the question itself but on the vocabulary used, which is often very similar, often with terms used synonymously over most classes due to the scientific subject matter. This makes text comprehension even more important. Thereby, the low variance of vocabulary in the data set could be another reason why LDA as a generative model performs better than kNN.

On the same note, it is interesting to note that the class "Communication" with comparatively few text strings (69) performed surprisingly well in relation to “Regulatory” (68) and “Quality” (103), with an accuracy of 50% and an F1-score 57% ranking even better than “Quality” (F1-score of 41%) with more text strings. This finding supports the assumption that the similar scientific vocabulary of the text entries makes it more difficult for the algorithm to assign entries to the correct class. The class "Communication" consists of aggregated text entries that mainly express sentiments, such as praise or firestorm, but also enquiries about possible career opportunities, thus differing rather substantially from the other scientific heavy text data shared by all other classes.

4.4 Conclusion

In general, the performance of all three algorithms is in line with literature evidence disclosed in previous chapters thus validating the outcome of this project and supporting the feasibility of a scaled implementation on a national and international level for pharma-owned social media channels. As expected fastText was superior concerning overall performance, which is why out of the three models investigated the one to be chosen for further project development is fastText, even though also the other two performed well given the limited and imbalanced data set. The subsequent section provides closing remarks on the important ideas that have been added to theory and practice rather than just restating the research’s findings.

5 CONCLUSION

5.1 Summary

Digital marketing in pharmaceuticals has unlocked a new possibility for targeted communications, with technology unleashing the ability to respond – at scale — to individual customer demands. Companies can now interact with physicians with a precision that was unimaginable even a few years ago thanks to the combination of more channel options, the adaptability of digital content, and the gain of improved customer insights. The multichannel approach is being gradually replaced by omnichannel – using the most effective channels, which must be mutually integrated into a unified system with the goal of seamless and uninterrupted communication with the client. This increasingly entails the use of social media.

The use of social media by pharmaceutical corporations is a double-edged sword and the majority are still cautious mainly due to the many restrictions and legal requirements for processing and archiving compliance-relevant information. In the meantime, increasing numbers of customers and patients have turned to social media to share their experiences and preferences, insightful information that is of utmost relevance for Pharma B2B and B2C interactions. Unfortunately, comprehensive solutions haven't evolved at the same pace as social media: In many cases, pharma brands still rely on internal social media teams that are also charged with launching marketing campaigns, or, in other cases, they involve third parties who might employ creativity and innovation in the realisation of a marketing campaign, but do not include monitoring of compliance-relevant information within their core competencies. Unfortunately, Pharma is missing a special source of information, firstly in terms of valuable insight into customer preferences, and secondly in terms of better drug safety monitoring and adverse drug reaction detection.

5.2 Contribution to Knowledge

Pharma companies have long recognised the potential of social network sites. However, given all these restrictions, regulations, hurdles as well as financial implications potentially causing reputational harm, it comes as no surprise that Pharma is especially careful when embarking on new initiatives. Additionally, the growing use of social media as part of the digital marketing strategy carries the need to monitor these channels in terms of compliance-relevant information. Literature evidence available on well-known platforms for scientific papers, such as PubMed and Medbase, predominantly lists work of explorative nature, pharmacovigilance monitoring on patient forums, directed at specific substances, e.g., bupropion (antidepressant) or methylphenidate (Ritalin – neuronal stimulant for attention deficit disorders), or general social media monitoring projects for pharmacovigilance at the international level, such as the WEBRadr project (see description above). According to the author's research, only anecdotal reports of monitoring practices on Pharma-owned Social Media channels regarding compliance-relevant information. Hence, this study aims to implement a Machine Learning-based classification model that automatically detects compliance-relevant information in text comments posted on pharma-owned social media channels and classifies it according to the type of information to forward it to the appropriate department for further processing.

5.2.1 Study Limitations

As different methods were investigated in this work, the final data set, fit for "any of" classification was modified to fit "one of" classification: Classification of classes which are not mutually exclusive is also known as, "any-of", multi-label, or multi-value classification, whereby a document or text element can belong to multiple classes, just one class, or no classes at all. Mutually exclusive classes are described in "one-of" classifications, also known as multiclass, and single-label classification, with more than two classes but with each element belonging to one single class. LDA and kNN are examples of such "one-of" classifiers, whereas fastText is easily capable of handling multi-label classification. For comparison reasons, the data set was modified so that each text element was assigned to a single class, and in case of ambiguity, to the most prominent one.

However, in text categorization, true one-of problems are less frequent than any-of problems. Especially in real life, when processing incoming queries, internal consultation is often required to establish the nature of the enquiry and to forward it to the respective department. Thus, not all text data can always be clearly assigned to one single class. As already mentioned, particularly the classes "Medicine" and "Pharmacovigilance" sometimes have fluid, marginal transitions.

Although all three algorithms achieved results above 50%, especially fastText with 69% after 2 seconds of learning time, and thus the proof of concept was successful, it must still be considered that the data set was rather small. A larger data set could show the performance differences between the two classical models and fastText even more clearly.

5.3 Implications for Relevant Stakeholders

This project consisted of mining comments (chat logs) on pharma-owned social media channels for compliance-relevant information via a text classification model, and forwarding it to the respective department for further processing and archiving. The purpose was to highlight the importance and the outreach of the idea behind this project by demonstrating that already proven technology can be used to overcome hesitation and caution from pharmaceutical companies to engage in social media, encouraging the full use of this pull marketing tool and the exploitation of its enormous potential as part of the omnichannel strategy.

Used across countries, this could capture a larger number of adverse event reports as well as quality and regulatory requests, thus contributing to increased drug safety and compliance monitoring. Through the use of already proven and widely used machine learning methods, the proof of concept could show that a scaled implementation is possible. Further investigations are necessary to establish the project further and to use it in cooperation with pharmaceutical companies.

Furthermore, as the drug development process grows more collaborative, and patients' perspectives become more valuable, social networks are becoming recognised as an effective means of gathering and disseminating information. Although direct promotion of prescription drugs is prohibited in Europe, social media platforms are an invaluable asset for pull marketing strategies in terms of indirect sales impulses. Indirect sales impulses increase trust in your company as an expert and problem solver and significantly influence whether you are included in the doctors' decision-making process. The long-term effect of indirect impulses should therefore not be underestimated, especially with regard to long-term customer loyalty. A strong brand creates trust, which is reflected in relationships with doctors and ultimately in company turnover.

5.4 Future Research

It should be mentioned that while "Pharmacovigilance", "Medicine", "Quality" and "Regulatory" belong to compliance-relevant areas, as they must be processed and archived either due to company-internal SOPs or official regulations, text data under the aggregated term "Communication" and "Marketing" do not belong to them. "Communication" and "Marketing" include all text data that may be important for the "patient centricity" and "marketing" departments, i.e., praise, firestorms, and customer preferences. While these insights are not subject to compliance guidelines, they can still have a significant impact on marketing strategies and campaign set-up as well as feedback. Although both classes are not the focus of this work and only comprise a small amount of the text data, they could be included for further development of the project. Another aspect in relation to future feature extension of this project is to extend such a text categorisation to input sources other than social media. Customer enquiries also come in directly to the company, either as a phone call or via a central designated e-mail. Usually, in these cases, the reception desk is responsible for sorting them out and sending them to the right department. Especially the sorting and forwarding of incoming mail could easily be taken over by the text classification algorithm. In this way, all input channels could be controlled automatically, thereby shortening processing times and simplifying and streamlining operations.

6 BIBLIOGRAPHY

Ahsan, M. M., & Siddique, Z. (2022). Machine learning-based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 102289.

AMFS – American Medical Forensic Specialists. (2019, June 15). Adverse Drug Reactions and Drug-Drug Interactions: Consequences and Costs. Pharmacology Expert. Available: https://www.amfs.com/adverse-drug-reactions-and-drug-drug-interactions-consequences-and-costs/. Last retrieved: 14.01.2023.

Azoev, G., Sumarokova, E., & Butkovskaya, G. (2019, December). Marketing communications integration in healthcare industry: digitalization and omnichannel technologies. In International Scientific and Practical Conference on Digital Economy (ISCDE 2019) (pp. 635-640). Atlantis Press.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics.

Bate, A., & Luo, Y. (2022). Artificial Intelligence and Machine Learning for Safe Medicines. Drug Safety, 45(5), 403-405.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., & Kandola, J., Hofmann, T., Poggio, T., & Shawe-Taylor, J. (Eds.). (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3(6), 1137–1155. https://doi.org/10.1162/153244303322533223

Bojanowski P, Grave E, Joulin A, Mikolov T (2017). “Enriching Word Vectors with Subword Information.” Transactions of the Association for Computational Linguistics, 5, 135–146. doi:10.1162/tacl_a_00051.

Brants, T. (2003). Natural Language Processing in Information Retrieval. CLIN, 111.

Chan, H. S., Shan, H., Dahoun, T., Vogel, H., & Yuan, S. (2019). Advancing drug discovery via artificial intelligence. Trends in pharmacological sciences, 40(8), 592-604.

Chan, R. K. C., Lim, J. M. Y., & Parthiban, R. (2021). A neural network approach for traffic prediction and routing with missing data imputation for intelligent transportation system. Expert Systems with Applications, 171, 114573.

Chen, D., & Manning, C. D. (2014, October). A fast and accurate dependency parser using neural networks. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 740-750).

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. Journal of Big Data, 7(1), 1-26.

Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021, September). Transformers: "The End of History" for Natural Language Processing?. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 677-693). Springer, Cham.

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. Translational Vision Science & Technology, 9(2), 14-14.

Chomsky, N., & Schützenberger, M. P. (1959). The algebraic theory of context-free languages. In Studies in Logic and the Foundations of Mathematics (Vol. 26, pp. 118-161). Elsevier.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12(ARTICLE), 2493-2537.

Cordon C. Garcia-Milà Pau Vilarino T. F. & Caballero P. (2016). Strategy is digital: how companies can use big data in the value chain. Springer, IX, p144. Available: https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1231568. Retrieved on: 14.01.2023.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

CRAN mirror - Wirtschaftsuniversität Wien. (n.d.). Available: https://cran.wu.ac.at/. Retrieved on: 14.01.2023.

Dada, E. G., Bassi, J. S., Chiroma, H., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802.

Damiati, S. A. (2020). Digital pharmaceutical sciences. AAPS PharmSciTech, 21(6), 1-12.

Damiati, S. A., Martini, L. G., Smith, N. W., Lawrence, M. J., & Barlow, D. J. (2017). Application of machine learning in prediction of hydrotrope-enhanced solubilisation of indomethacin. International Journal of Pharmaceutics, 530(1-2), 99-106.

Daniel Jurafsky & James H. Martin. (2019). Semantic Role Labelling and Argument Structure. Speech and Language. (pp. 1-24). Processing.Copyright © 2021. All rights reserved. Draft of December 29, 2021. Available: https://web.stanford.edu/~jurafsky/slp3/. Retrieved on: 14.01.2023.

Dixon, S. (2022, December 16). Number of social media users worldwide from 2017 to 2027. Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Retrieved on: 14.01.2023.

dos Santos, M. L. B. (2021). The "so-called" UGC: an updated definition of user-generated content in the age of social media. Online Information Review, 46(1), 95-113.

EMA Website. (1995, June). Note for Guidance on Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. Available: https://www.ema.europa.eu/en/documents/scientific-guideline/international-conference-harmonisation-technical-requirements-registration-pharmaceuticals-human-use_en-15.pdf. June. Retrieved on: 14.01.2023.

EMA Website. (2016, September 19). Workshop: How could social media be relevant to regulatory decision making? Available: https://www.ema.europa.eu/en/documents/presentation/presentation-how-could-social-media-data-be-relevant-regulatory-decision-making-june-m-raine_en.pdf. Retrieved on: 14.01.2023.

EMA Website. (2022, April 1). Explanatory note on general fees payable to the European Medicines Agency. Available: https://www.ema.europa.eu/en/documents/other/explanatory-note-general-fees-payable-european-medicines-agency-01-april-2022_en.pdf. Retrieved on: 14.01.2023.

Englesson, E., & Azizpour, H. (2021). Consistency Regularization Can Improve Robustness to Label Noise. arXiv preprint arXiv:2110.01242.

Eurostat Website. (2021, April 6). One in two citizen look for health information online. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20210406-1. Retrieved on: 14.01.2023.

Eysenbach, G. (ed.) (2016), "Garbage in, garbage out: Data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection", Journal of Medical Internet Research, Vol. 18/2, p. e41, http://dx.doi.org/10.2196/jmir.4738

Faraj, S., Renno, W., & Bhardwaj, A. (2021). Unto the breach: What the COVID-19 pandemic exposes about digitalization. Information and Organization, 31(1), 100337.

fastText Website. (n.d.). L Available: https://fastText.cc. Retrieved on: 14.01.2023.

FDA Guidance Document. (2014, June). Internet/Social Media Platforms with Character Space Limitations - Presenting Risk and Benefit Information for Prescription Drugs and Medical Devices. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/internetsocial-media-platforms-character-space-limitations-presenting-risk-and-benefit-information. Retrieved on: 14.01.2023.

FDA Website. (2022, October 27). Prescription Drug User Fee Amendments. Available: https://www.fda.gov/industry/fda-user-fee-programs/prescription-drug-user-fee-amendments. Retrieved on: 14.01.2023.

Feinerer, I. (2022, December 14). Introduction to the tm Package Text Mining in R. Available: From: https://cran.r-project.org/web/packages/tm/tm.pdf. Retrieved on: 14.01.2023.

Finney Rutten, L. J., Blake, K. D., Greenberg-Worisek, A. J., Allen, S. V., Moser, R. P., & Hesse, B. W. (2019). Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. *Public Health Reports*, *134*(6), 617-625.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.

Fragoudis, D., Meretakis, D., & Likothanassis, S. (2005). Best terms: an efficient feature-selection algorithm for text categorization. Knowledge and Information Systems, 8(1), 16-33.

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In Competition and cooperation in neural nets (pp. 267-285). Springer, Berlin, Heidelberg.

Gaisford, S., & Saunders, M. (2012). Essentials of pharmaceutical preformulation. John Wiley & Sons.

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

Garcia Constantino, M. (2013). *On the use of text classification methods for text summarisation* (Doctoral dissertation, University of Liverpool).

Garvin, P. (1967). The Georgetown-IBM experiment of 1954: an evaluation in retrospect. Papers in linguistics in honor of Dostert, 46-56.

Gerfertz-Schiefer, N. (2021, July 1). Digitale Transformation in Pharmaunternehmen. PharmAustria, MedMedia. Available: https://www.medmedia.at/pharm-austria/digitale-transformation-in-pharmaunternehmen/. Retrieved on:14.01.2023.

Han, R., Xiong, H., Ye, Z., Yang, Y., Huang, T., Jing, Q., ... & Ouyang, D. (2019). Predicting physical stability of solid dispersions by machine learning techniques. Journal of Controlled Release, 311, 16-25.

Hayashi, Y. (2019). The right direction needed to develop white-box deep learning in radiology, pathology, and ophthalmology: A short review. Frontiers in Robotics and AI, 6, 24.

Henstock, P. V. (2019). Artificial intelligence for pharma: time for internal investment. Trends in pharmacological sciences, 40(8), 543-546.

Hinton, G.E. & Salakhutdinov, R.R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science (New York, N.Y.). 313. 504-7. 10.1126/science.1127647.

Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. Neural computation, 9, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

HOPL Website. (2023). Online Historical Encyclopedia of Programming Languages. Computer Language Taxonomy and Genealogy. Available: https://hopl.info/. Retrieved on: 14.01.2023.

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

Houssein, E. H., Emam, M. M., Ali, A. A., & Suganthan, P. N. (2021). Deep and machine learning techniques for medical imaging-based breast cancer: A comprehensive review. Expert Systems with Applications, 167, 114161.

Hsu, A. S., & Griffiths, T. E. (2010, February). Effects of generative and discriminative learning on use of category variability. In 32nd annual conference of the cognitive science society.

Hutchins, W. (2004). The Georgetown-IBM Experiment Demonstrated in January 1954. 3265. 102-114. 10.1007/978-3-540-30194-3_12.

IFPMA.(2022, September 28). Joint Note for Guidance on social media and digital channels. Available: https://ifpma.org/publications/joint-note-for-guidance-on-social-media-and-digital-channels/. Retrieved: 14.01.2023.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. Journal of computational and graphical statistics, 5(3), 299-314.

IMI Website. (n.d.). Innovation Medicines Initiative. WEB-RADR Project. (2022) Web Recognizing Adverse Drug Reactions. Available: https://www.imi.europa.eu/projects-results/project-factsheets/web-radr. Retrieved on: 14.01.2023.

International drug monitoring: the role of national centres, report of a WHO meeting held in Geneva from 20 to 25 September 1971. World Health Organization. 1972. [2022-09-12]. https://apps.who.int/iris/handle/10665/40968. [PubMed: 4625548]

Ivakhnenko, A. G., & Lapa, V. G. (1966). Cybernetic predicting devices. Purdue Univ Lafayette Ind School of Electrical Engineering.

Jia, X., Pang, Y., & Liu, L. S. (2021). Online Health Information Seeking Behavior: A Systematic Review. Healthcare (Basel, Switzerland), 9(12), 1740. https://doi.org/10.3390/healthcare9121740

Jiachong Li. 2019. Regression and Classification in Supervised Learning. In Proceedings of the 2nd International Conference on Computing and Big Data (ICCBD 2019). Association for Computing Machinery, New York, NY, USA, 99–104. https://doi.org/10.1145/3366650.3366675

Joachims, T. (2002). Text classification. Learning to Classify Text Using Support Vector Machines, 7-33.

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural Language Processing: History, Evolution, Application, and Future Work. In Proceedings of 3rd International Conference on Computing Informatics and Networks (pp. 365-375). Springer, Singapore.

Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016a). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016b). Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Joyce, James (2003), "Bayes' Theorem", in Zalta, Edward N. (ed.), The Stanford Encyclopedia of Philosophy (Spring 2019 ed.), Metaphysics Research Lab, Stanford University,

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kane, V. L. (2020). Interpretation and machine translation towards google translate as a part of machine translation and teaching translation. Applied Translation, 15(1), 10-17.

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018, December). Customer segmentation using K-means clustering. In 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS) (pp. 135-139). IEEE.

Kao, A., & Poteet, S. R. (Eds.). (2007). Natural language processing and text mining. Springer Science & Business Media.

Kelley, H. J. (1960). Gradient Theory of Optimal Flight Paths. ARS Journal, 30(10), 947-954. Precursor of modern backpropagation.

Khan, Z. H., Siddique, A., & Lee, C. W. (2020). Robotics utilization for healthcare digitization in global COVID-19 management. International journal of environmental research and public health, 17(11), 3819.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications, 1-32.

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-based construction of a verb lexicon. AAAI/IAAI, 691, 696.

Kitson, P. J., Marie, G., Francoia, J. P., Zalesskiy, S. S., Sigerson, R. C., Mathieson, J. S., & Cronin, L. (2018). Digitization of multistep organic synthesis in reactionware for on-demand pharmaceuticals. Science, 359(6373), 314-319.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text classification for organizational researchers: A tutorial. Organizational research methods, 21(3), 766-799.

Kolluri, S., Lin, J., Liu, R., Zhang, Y., & Zhang, W. (2022). Machine learning and artificial intelligence in pharmaceutical research and development: a review. The AAPS Journal, 24(1), 1-10.

Kuhn, M., Yates, P., Hyde, C. (2016). Statistical Methods for Drug Discovery. In: Zhang, L. (eds) Nonclinical Statistics for Pharmaceutical and Biotechnology Industries. Statistics for Biology and Health. Springer, Cham. https://doi.org/10.1007/978-3-319-23558-5_4

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In data democracy (pp. 83-106). Academic Press.

Kumar, S. (2018). Survey on Personalized Web Recommender System. International Journal of Information Engineering & Electronic Business, 10(4).

Le Louët, H., & Pitts, P. J. (2023). Twenty-First Century Global ADR Management: A Need for Clarification, Redesign, and Coordinated Action. Therapeutic Innovation & Regulatory Science, 57(1), 100-103.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.

Lee, D. (2013). Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.

Lesk, M. (1986, June). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th annual international conference on Systems documentation (pp. 24-26).

Li, R., Curtis, K., Zaidi, S. T., Van, C., & Castelino, R. (2022). A new paradigm in adverse drug reaction reporting: consolidating the evidence for an intervention to improve reporting. Expert Opinion on Drug Safety, 21(9), 1193-1204.

Li, Y. F., & Liang, D. M. (2019). Safe semi-supervised learning: a brief introduction. Frontiers of Computer Science, 13(4), 669-676.

Limaye, N., & Saraogi, A. (2018). How social media is transforming pharma and healthcare. Applied Clinical Trials, 27(2).

Mackenzie, D. (1995). The automation of proof: A historical and sociological exploration. IEEE Annals of the History of Computing, 17(3), 7-29.

Mak, K. K., & Pichika, M. R. (2019). Artificial intelligence in drug development: present status and future prospects. Drug discovery today, 24(3), 773-780.

Manning, C. D., Raghavan, P., & Schutze, H.(2008) Introduction to information retrieval. Cambridge University press. Companion website to the book. Available: https://nlp.stanford.edu/IR-book/. Retrieved on: 14.01.2023.

Manzano, T., & Langer, G. (2020). Getting ready for pharma 4.0. Pharmaceutical Engineering.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276-282.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Yih, W. T., & Zweig, G. (2013b, June). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 746-751).

Minastireanu, E. A., & Mesnita, G. (2019). An Analysis of the Most Used Machine Learning Algorithms for Online Fraud Detection. Informatica Economica, 23(1).

Mor, L. (2022). Introduction to Machine Learning, International Journal of Science and Research (IJSR), 11(3), 1522-1525, https://www.ijsr.net/get_abstract.php?paper_id=SR22328110600.

Mouselimis L (2022). fastText: Efficient Learning of Word Representations and Sentence Classification using R. R package version 1.0.3, https://CRAN.R-project.org/package=fastText.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.

Neely, S., Eldredge, C., & Sanders, R. (2021). Health Information Seeking Behaviors on Social Media During the COVID-19 Pandemic Among American Social Networking Site Users: Survey Study. Journal of medical Internet research, 23(6), e29802. https://doi.org/10.2196/29802

Ni, J., Wang, L., Gao, H., Qian, K., Zhang, Y., Chang, S., & Hasegawa-Johnson, M. (2022). Unsupervised Text-to-Speech Synthesis by Unsupervised Automatic Speech Recognition. arXiv preprint arXiv:2203.15796.

OECD Health Policy Studies (November 21, 2019). Health in the 21st Century. Putting Data to Work for Stronger Health Systems. OECDilibrary. Available: https://doi.org/10.1787/e3b23f8e-en. Retrieved on: 14.01.2023

Olorunnimbe, K., & Viktor, H. (2022). Deep learning in the stock market—a systematic survey of practice, backtesting, and applications. Artificial Intelligence Review, 1-53.

Oxford Dictionary Website. (n.d.). Available: https://www.oxfordlearnersdictionaries.com/definition/english/make_1. Retrieved on: 14.01.2023.

Package caret. (n.d.). caret: Classification and Regression Training. Available: https://CRAN.R-project.org/package=caret. Retrieved on: 14.01.2023.

Package fastText. (n.d.). fastText: Efficient Learning of Word Representations and Sentence Classification. https://CRAN.R-project.org/package=fastText

Package SnowballC.(n.d.). SnowballC: Snowball Stemmers Based on the C 'libstemmer' UTF-8 Library. From: https://CRAN.R-project.org/package=SnowballC. Retrieved on: 14.01.2023

Package textstem. (n.d.). textstem: Tools for Stemming and Lemmatizing Text. Available: https://CRAN.R-project.org/package=textstem. Retrieved on: 14.01.2023.

Package Tidyverse. (n.d.) tidyverse: Easily Install and Load the 'Tidyverse'. Available: https://CRAN.R-project.org/package=tidyverse. Retrieved on: 14.01.2023.

Package tm. (n.d.). tm: Text Mining Package. Available: https://CRAN.R-project.org/package=tm. Retrieved on: 14.01.2023.

Palmer, Martha & Kingsbury, Paul & Gildea, Daniel. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics. 31. 71-106. 10.1162/0891201053630264.

Parekh, D., Kapupara, P., & Shah, K. (2016). Digital pharmaceutical marketing: A review. Research Journal of pharmacy and technology, 9(1), 108.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Pollack, A. (1983, April 28). Technology; The Computer as Translator. The New York Times, Section D, page 2. Digitized version of print archive. Available: https://www.nytimes.com/1983/04/28/business/technology-the-computer-as-translator.html. Retrieved on: 14.01.2023.

Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061.

Puri, S., Gosain, D., Ahuja, M., Kathuria, I., & Jatana, N. (2013). Comparison and analysis of spam detection algorithms. International Journal of Application or Innovation in Engineering and Management, 2(4), 255-261.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., & Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? arXiv preprint arXiv:1804.06323.

R Project Website. Available: https://www.r-project.org/. Retrieved on: 14.01.2023.

Rao, C. R. (1948). Tests of significance in multivariate analysis. Biometrika, 35(1/2), 58-79.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society. Series B (Methodological), 10(2), 159-203.

Rapaport, W. J. (2020). What is artificial intelligence? Journal of Artificial General Intelligence, 11(2), 52-56.

Roldan-Baluis, W. L., & Vasquez, M. S. M. (2022). The Effect of Natural Language Processing on the Analysis of Unstructured Text: A Systematic Review. International Journal of Advanced Computer Science and Applications, 13(5).

Rosen, A. (2017, November 7). Twitter made easier. Twitter Blog. Available: Available: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html. Retrieved on: 14.01.2023.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. https://doi.org/10.1037/h0042519

RStudio Desktop Website. (n.d.). Available: https://posit.co/download/rstudio-desktop/. Retrieved on: 14.01.2023.

Sathya, D., Sudha, V., & Jagadeesan, D. (2020). Application of machine learning techniques in healthcare. In Handbook of Research on Applications and Implementations of Machine Learning Techniques (pp. 289-304). IGI Global.

Schank, R. C. (1972). Conceptual dependency: A theory of natural language understanding. Cognitive psychology, 3(4), 552-631.

Schlander, M., Hernandez-Villafuerte, K., Cheng, C. Y., Mestre-Ferrandiz, J., & Baumann, M. (2021). How much does it cost to research and develop a new drug? A systematic review and assessment. PharmacoEconomics, 39(11), 1243-1269.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

Schumacher, P., Minor, M., Walter, K., & Bergmann, R. (2012, April). Extraction of procedural knowledge from the web: A comparison of two workflow extraction approaches. In Proceedings of the 21st International Conference on World Wide Web (pp. 739-747).

Sebastian, I. M., Ross, J. W., Beath, C., Mocker, M., Moloney, K. G., & Fonstad, N. O. (2020). How big old companies navigate digital transformation. In Strategic information management (pp. 133-150). Routledge.

Seddon, G., Lounnas, V., McGuire, R., van den Bergh, T., Bywater, R. P., Oliveira, L., & Vriend, G. (2012). Drug design for ever, from hype to hope. Journal of computer-aided molecular design, 26(1), 137-150.

Sehlstedt, U., Bohlin, N., de Maré, F., & Beetz, R. (2016). Embracing digital health in the pharmaceutical industry. International Journal of Healthcare Management, 9(3), 145-148.

Shao, K., Tang, Z., Zhu, Y., Li, N., & Zhao, D. (2019). A survey of deep reinforcement learning in video games. arXiv preprint arXiv:1912.10944.

Shwartz, S. (2020, March 17). Artificial Intelligence 101. AI Perspectives. Online book, companion reference to the print book “Evil Robots, Killer Computers, and Other Myths: The Truth About AI and the Future of Humanity”. Available: https://www.aiperspectives.com/introduction/. Retrieved on: 14.01.2023.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140-1144.

Singh, A., & Hess, T. (2017). How chief digital officers promote the digital transformation of their companies. Mis Q Exec 16: 1–17.

Singh, S., Kaushik, M., Gupta, A., & Malviya, A. K. (2019, March). Weather forecasting using machine learning techniques. In Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE).

Smith, S. (2021, May 25). 9 Artificial Intelligence in Drug Discovery Trends and Statistics. BenchSci. Available: https://blog.benchsci.com/artificial-intelligence-in-drug-discovery-trends-and-statistics. Retrieved on: 14.01.2023.

Stone, Z., Zickler, T., & Darrell, T. (2008, June). Autotagging facebook: Social network context improves photo annotation. In 2008 IEEE computer society conference on computer vision and pattern recognition workshops (pp. 1-8). IEEE.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

Tekin, M., Etlioğlu, M., Koyuncuoğlu, Ö., & Tekin, E. (2018, August). Data mining in digital marketing. In The International Symposium for Production Research (pp. 44-61). Springer, Cham.

Thornton, J. M., Laskowski, R. A., & Borkakoti, N. (2021). AlphaFold heralds a data-driven revolution in biology and medicine. Nature Medicine, 27(10), 1666-1669.

Trattner, C., Jannach, D., Motta, E., Costera Meijer, I., Diakopoulos, N., Elahi, M., ... & Moe, H. (2022). Responsible media technology and AI: challenges and research directions. AI and Ethics, 2(4), 585-594.

Turing, A. M. (2012). Computing machinery and intelligence (1950). The Essential Turing: the Ideas That Gave Birth to the Computer Age, 433-464.

Turing, A.M. (1950, October) I.—COMPUTING MACHINERY AND INTELLIGENCE, Mind, Volume LIX, Issue 236, Pages 433–460, https://doi.org/10.1093/mind/LIX.236.433

Ventola, C. L. (2018). Big data and pharmacovigilance: data mining for adverse drug events and interactions. Pharmacy and therapeutics, 43(6), 340.

Verhoef, P. C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Dong, J. Q., Fabian, N., & Haenlein, M. (2021). Digital transformation: A multidisciplinary reflection and research agenda. Journal of Business Research, 122, 889-901.

Violation Tracker, Pharmaceuticals. (2022). Available: https://violation-tracker.goodjobsfirst.org/industry/pharmaceuticals. Retrieved on: 14.01.2023.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.

WHO, World Health Organization. (2020, December 9). Top 10 causes of death. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Retrieved on: 14.01.2023.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686.

Yadav, A., Patel, A., & Shah, M. (2021). A comprehensive review on resolving ambiguities in natural language processing. AI Open, 2, 85-92.

Yang, Y., Ye, Z., Su, Y., Zhao, Q., Li, X., & Ouyang, D. (2019). Deep learning for in vitro prediction of pharmaceutical formulations. Acta pharmaceutica Sinica. B, 9(1), 177–185. https://doi.org/10.1016/j.apsb.2018.09.010

Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., ... & Marti, R. (2017). Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics, 22(4), 1218-1226.

Zhang, H. & Yu, T. (2020). Taxonomy of Reinforcement Learning Algorithms. In: Dong, H., Ding, Z., Zhang, S. (eds) Deep Reinforcement Learning. Springer, Singapore. https://doi.org/10.1007/978-981-15-4095-0_3

Zhu, L., Spachos, P., Pensini, E., & Plataniotis, K. N. (2021). Deep learning and machine vision for food processing: A survey. Current Research in Food Science, 4, 233-249.

APPENDICES

Appendix 1: R Code Line for Data Set Pre-Processing

```
library(textstem)
library(tm)
library(SnowballC)
library(caret)
library(tidyverse)

# avoid factorization of string
options(stringsAsFActors=FALSE)


# read text from local data folder
mr <- read_excel("Report_ws3_LDA_KNN.xlsx")

mr$classification <- trimws(mr$classification)
View(mr)
dim(mr)

seed <- 123
set.seed(seed)
mr$Type <- sample(c("train","test"), nrow(mr), replace = TRUE, prob =
c(0.8,0.2))

# check in the contingency table the distribution of the brands over
Type variable
table(mr$classification, mr$Type)

# put the Content column into a corpus to perform preprocessing opera-
tions
corpus <- VCorpus(VectorSource(mr$content))

inspect (corpus[[1]])
    <<PlainTextDocument>>
    Metadata:  7
    Content:  chars: 38
    I would like to work for your company.

 dtm <- corpus %>%
     # transform into lowercase
     tm_map(content_transformer(tolower)) %>%
     # remove the stopwords
     tm_map((removeWords),stopwords("english")) %>%
     # remove punctuation
     tm_map(removePunctuation) %>%
     # remove numbers
     tm_map(removeNumbers) %>%
     # stemming
     tm_map(stemDocument) %>%
     # white space removal
     tm_map(stripWhitespace) %>%
     # transform into a Document-Term Matrix
     DocumentTermMatrix() %>%
     # apply the Tf-Idf weighting function
     weightTfIdf() %>%
     # adjust for sparsity
     removeSparseTerms(0.98)

inspect(dtm)


dim(dtm)
# understanding the dtm: the number of dimensions changes based on
stemming or lemmatizing

# transform the dtm into a data frame, passing through an intermediary
step
# of creating a simple matrix
dtm <- as.data.frame(as.matrix(dtm))

# dtm and mr contain exactly the same documents, i.e. of rows, so we can
add to the dtm the Brand # and Type columns from dt3
dtm <- bind_cols(dtm, "Classification" = mr$classification, "Type" =
mr$type
# split the dtm into the "train" and "test" subsets and remove the Type
column from the subsets as it # became redundant
dtm_train <- dtm %>%
  filter(Type == "train") %>%
  select(-Type)

dtm_test <- dtm %>%
  filter(Type == "test") %>%
  select(-Type)

# set the validation scheme, 10-fold cross-validation
ctrl <- trainControl(method="cv", number = 10, classProbs = TRUE)
```

Appendix 2: R Code Line for Text Classification Algorithm: Training and Testing of Linear Discriminant Analysis (LDA)

```
# train the lda method on the column classification, using the dtm_train
data set
set.seed(seed)
# train the model with the train set
lda <- train(as.factor(Classification) ~ . , data = dtm_train, method =
"lda", trControl = ctrl, metric= "Accuracy", tuneLength = 5, na.action
= na.pass)

# check the model
lda

# predict on test data
lda.predict <- predict(lda, newdata = dtm_test)

# check model performance with the test set
cfm <- confusionMatrix(lda.predict, as.factor(dtm_test$Classification))

print(cfm)
```

Appendix 3: R Code Line for Text Classification Algorithm: Training and Testing of k-Nearest Neighbour (kNN)

```
# train the knn method on the column classification, using the dtm_train
data set
set.seed(seed)
# train the model with the train set
knn <- train(as.factor(Classification) ~ ., data =
data.frame(dtm_train), method = "knn", trControl = ctrl, metric="Accuracy",
tuneLength = 5)

# check the model
knn

# predict on test data
knn.predict <- predict(knn, newdata = data.frame(dtm_test))

# check performance with the test set
cfm <- confusionMatrix(knn.predict, as.factor(dtm_test$Classification))
print(cfm)
```

Appendix 4: R Code Line for Text Classification Algorithm: Data Set Pre-Processing, Training and Testing of fastText

```
library(tidyverse)
library(textstem)
library(fastText)
mr <- read_excel('Report_ws3_LDA_KNN.xlsx')

mr2 <- mr %>%
  mutate(classification = tolower(classification)) %>%
  mutate(content=gsub("[\r\n\t]", "", content)) %>%
  mutate(content=gsub("[0-9]", "", content)) %>%
  mutate(content = gsub("_x000D_", "", content)) %>%
  mutate(content = removeWords(content, stopwords())) %>%
  mutate(content = stem_words(content)) %>%
  mutate(content = gsub("  ", "", content)) %>%
  mutate(content = str_replace_all(content, "[[:punct:]]", "")) %>%

  mutate(classification = paste0("__label__",classification)) %>%

  mutate(text = paste(classification, content)) %>%
  mutate(id=row_number()) %>%
  select(id, text)

#view mr object
print(mr2)

set.seed(42)
#Create training set
train <- mr2 %>%
      sample_frac(.80)

#Create test set
valid <- anti_join(mr2, train, by = 'id') %>%
      select(text)

train <- train %>%
  select(text)

valid <- valid %>%
  select(text)

alldata <- mr2 %>%
  select(text)

write_delim(train, file="data.train", delim = "\n", col_names = FALSE,
quote = "none")
write_delim(valid, file="data.valid", delim = "\n", col_names = FALSE,
quote = "none")
write_delim(alldata,  file="alldata.txt",  delim  =  "\n",  col_names  =
FALSE, quote = "none")


# Command “supervised”
#-----------
list_params = list(command = 'supervised',
                   lr = 0.6,
                   dim = 200,
                   input = file.path("data.train"),
                   output = file.path('model_pharma'),
                   verbose = 2,
                   thread = 4)
res = fasttext_interface(list_params,
                         path_output    =    file.path('sup_logs.txt'),

                         MilliSecs = 5,
                         remove_previous_file = TRUE,
                         print_process_time = TRUE)

# 'predict' function
#-------------------

list_params = list(command = 'predict',
                   model = file.path('model_pharma.bin'),
                   test_data = file.path('data.valid'),
                   k = 1,
                   th = 0.0)
res = fasttext_interface(list_params,
                         path_output  =  file.path('predict_valid.txt'))


# 'predict-prob' function
#------------------------
list_params = list(command = 'predict-prob',
                   model = file.path('model_pharma.bin'),
                   test_data = file.path('data.valid'),
                   k = 1,
                   th = 0.0)
res = fasttext_interface(list_params,
                         path_output           =           file.path('pre-
dict_valid_prob.txt'))
# 'test' function
#----------------
```

```
list_params = list(command = 'test',
                   model = file.path('model_pharma.bin'),
                   test_data = file.path('data.valid'),
                   k = 1,
                   th = 0.0)
res = fasttext_interface(list_params)

# 'test-label' function
#----------------------
list_params = list(command = 'test-label',
                   model = file.path('model_pharma.bin'),
                   test_data = file.path('data.valid'),
                   k = 1,
                   th = 0.0)
res = fasttext_interface(list_params,
                         path_output = file.path('test_label_valid.txt'))

## verify snippet
st_dat = read.delim(file.path("alldata.txt"),
                    stringsAsFactors = FALSE)

res_pharma = unlist(lapply(1:nrow(st_dat), function(y)
  strsplit(st_dat[y, ], " ")[[1]][which(sapply(strsplit(st_dat[y, ], "

")[[1]], function(x)
    substr(x, 1, 9) == "__label__") == T)])
)

test_label_valid = read.table(file.path('test_label_valid.txt'),

                             quote="\"", comment.char="")

# number of unique labels of data equal to the rows of the 'test_label_valid.txt' file
length(unique(res_pharma)) == nrow(test_label_valid)

head(test_label_valid)
```