الجـــــــــــــــــــــمهـوريـــــة الجـــــــــــــزائريـة الديمـــــــقـراطيـة

الـشـــــــــعبيـة

DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA

وزارة التـــــــــــــــــــــــــــعليـم العـــــــالي والبـــــــــــحث العـــــلـمـــي

Ministry of Higher Education and Scientific Research

جـــــــــــــــامعة أبي بكـــر بـلـقـايد – تـلمســـــــــــــان –

Abou Bekr Belkaid University – Tlemcen –

Faculty of TECHNOLOGY

THESIS

Presented to obtain the **degree** of **DOCTORATE Third Cycle**

In : Biomedical Engineering

Speciality: Biomedical Informatics and Telemedicine

By : YOUBI Fatiha

Topic

Classification of medical documents and Opinions using machine learning

Publicly defended, on 07/11/2023, to the jury composed of

Mr. HADJ SLIMANE Zine Edine	Professor	Univ. Tlemcen	President
Mr.MESSADI Mahammed	Professor	Univ. Tlemcen	Supervisor
Mr.BECHAR Hacene	MCA	Univ. Tlemcen	Co- Supervisor
Mr. BENGANA Abdelfatih	MCA	Univ. Ain-Temouchent	Examiner 1
Mme. KHEMIS Kamila	MCA	Univ. Tlemcen	Examiner 2
Mme.SETTOUTI Nesma	MCA	Univ. Tlemcen	Invited member
Mme.LARIBI Souhila	Professor	CHU. Tlemcen	Invited member

I dedicate this work to:

My recently deceased grandfather, of whom I know how proud he was of me,even if he won't be there to see me graduate.
My loving parents, my brothers : Mohamed and Youcef and my sister : Fatima
My nephews : Amine , Abdelilah, Djawed , Ayoub and Boudis.
My husband and my wonderful daughter Israa
My brothers-in-law : abdelilah and abderaouf and sisters-in-law : Asma and Halima
My colleagues : Chaima,Hafida, Hadjer,Meriem,Asma, Imane , khadidja

Acknowledgements

My PhD experience was five years of scientific research, five years of work to reach this moment - this moment when I write these words, when I look back on the long adventure that was my thesis, and when I remember the people who helped me, supported me, taught me to keep a positive attitude, to be patient, and to work hard to achieve my goals. They simply made these five years a part of my life that I will always remember. I would like to thank all of them for their support and guidance in this delicate phase.

First and foremost, I wish to deeply thank Dr. SETTOUTI Nesma for their valuable guidance during all my phd study . I appreciate all the support and encouragement she gave me.
A very special thank you to Pro Messadi Mahammed and Dr. Bechar Hacene for accepting to supervised my work and providing me with the directions to finalize it in the best conditions.
I wish to deeply thank the members of my dissertation committee for accepting to review my framework and providing me with the directions and comments to improve it : Prof. Hadj slimane Zine Edine, Dr. Khemis Kamila from Tlemcen University and Dr . Bengana Abdelfatih from Ain Temouchent University.

Furthermore, I acknowledge my gratitude to Prof. LARIBI Souhila from CHU Tlemcen, Algeria, for her support, participation, and shared expertise during my internship in the CHU.

I would like to thank my colleagues from the Biomedical Engineering laboratory at Tlemcen University, especially the CREDOM team. These years would not have been as beautiful without the atmosphere, the cohesion, and the good mood prevailing in the laboratory. I have a particular thought for Hafida BELFILALI, Chaima CHERFI, Ilies LAHSAINI, Rima GUILAL, Hadjer ABDI, Meriem SAIM, with whom we started our research activities.

I would also like to say a heartfelt thank you to my father, who has always believed in me and encouraged me to follow my dreams. He has been by my side throughout my PhD, living every minute of it. And to my mother, who has helped me in every possible way during this difficult period, without them, I would not have had the courage to do this work. Furthermore, I would like to thank my brothers and sister who have always supported me in my work.

My deep gratitude goes to my husband who has supported me throughout my studies and encouraged me to keep going and stay strong and positive all the time, and to my dear Israa, who has been such a good little baby over the last six months and has allowed me to finish what I started.

Résumé

Ces dernières années, l'utilisation du text mining (TM) et de l'apprentissage automatique dans le domaine de la santé a suscité un intérêt croissant. La classification de documents a été une application courante, avec de nombreuses études se concentrant sur la classification de rapports médicaux à partir de données textuelles non structurées. Cependant, il est également nécessaire d'utiliser le TM et l'apprentissage automatique pour l'analyse des sentiments des données textuelles médicales dans les réseaux sociaux et les forums médicaux. Dans cette thèse, l'accent a été mis sur deux applications de TM dans le domaine médical : la classification des rapports d'autopsie pour détecter la manière de décès dans la wilaya de Tlemcen et l'analyse des opinions des patients et du public sur la santé et la pandémie de COVID-19 en utilisant des techniques d'apprentissage automatique. Les expériences menées dans les deux études ont montré que les modèles automatisés d'analyse d'opinions sont spécifiques à la tâche et que l'extraction de caractéristiques et l'architecture du classifieur d'apprentissage en profondeur jouent un rôle important dans le succès de ces modèles. Les résultats pourraient être utiles pour améliorer les stratégies liées à la surveillance des médicaments et de la COVID-19. Les orientations futures comprennent l'exploration d'autres types de techniques d'apprentissage en profondeur, l'utilisation de documents cliniques pour l'analyse des sentiments et l'analyse de l'état de santé algérien sur la base de classifieurs d'apprentissage automatique et d'apprentissage en profondeur.

Mots clés

Fouille de texte, traitement automatique de langue TAL, apprentissage automatique, apprentissage profond, analyse de sentiment médical, surveillance des médicaments, surveillance de COVID-19, rapports d'autopsie

Abstract

In recent years, there has been increasing interest in the use of text mining (TM) and machine learning in healthcare. Document classification has been a common application, with many studies focusing on classifying medical reports from unstructured text data. However, there is also a need to utilize TM and machine learning for sentiment analysis of medical textual data in social networks and medical forums. In this thesis, the focus was on two TM applications in the medical domain: classifying autopsy reports to detect the manner of death in Wilaya of Tlemcen and analyzing patient and public opinions on healthcare and the COVID-19 pandemic using machine learning techniques. The experiments conducted in both studies showed that automated models for opinion analysis are task-specific and that feature extraction and deep learning classifier architecture play important roles in the success of these models. The findings could be useful for improving strategies related to drugs monitoring and COVID-19 surveillance. Future directions include exploring other types of deep learning techniques, using clinical documents for sentiment analysis, and analyzing Algerian health status based on machine learning and deep learning classifiers.

Keywords
Text mining, NLP, machine learning, deep learning, medical sentiment analysis, drug monitoring, COVID-19 surveillance, autopsy reports.

Contents

List of Figures

List of Tables

Glossary

ADRs : adverse drug reactions
ATC : automatic text classification
AWD-LSTM : ASGD Weight-Dropped LSTM
BERT : Pretraining of deep bidirectional transformers
B-GRU : Bidirectional Gated Recurrent Units
BOW : Bag of words
CBOW :Continuous Bag-of-Words
CBR : case-based reasoning
CHU : University Hospital Center
CLSTM : Convolutional Long Short-term Memory network
CNN : Convolutional Neural Networks
CoD : cause of death
COVID-19 : Corona Virus Disease appeared in 2019
CRFs : conditional random fields
CSV : Comma-separated values
DBN : Deep Belief Network
DDI : Drug-Drug Interactions
DL : Deep Learning
DRNNs : deep recurrent neural networks
ELMo : Embeddings from Language Models
FasText : or FastText, is an embedding method for text vectorization
FN : False negatives
FP : False positives
GloVe : Global Vectors
GRUs : Gated Recurrent Units
ICMFS : improved Complete Measurement Feature Selection technique
KNN : k-Nearest Neighbors
LIME : Local Interpretable Model-agnostic Explanations
LR : Logistic Regression
LSTM : Long Short- Term-Memory
MNB : Multinomial Naive Bayes
MEDLINE : Medical Literature Analysis and Retrieval System Online
ML : Machine Learning

MLP : Multi-layer perception
MoD : manner of death
MUSE : Multilingual Universal Sentence Encoder
NLP : Natural language processing
NTFIDF : Normalized Term Frequency and Inverse Document Frequency
NB : Naive Bayes classifier
OM :Opinion Mining
PNN : Probabilistic Neural Networks
QRNN : Quasi-Recurrent Neural Network
RF : Random Forest
RBM : Restricted Boltzmann Machine
RoBERTa : A Robustly Optimized BERT Pretraining Approach
RNN : Recurrent Neural Networks
SA : Sentiment Analysis
SO : semantic orientation
SRS : simple random sampling
SVM :Support Vector Machines
TF-IDF : Term Frequency and Inverse Document Frequency
TextBlob : is a python library for Natural Language Processing
TM :Text-mining.
TN: True negatives
TP: True positives
UIMA : Unstructured Information Management Architecture
UMLS : Unified Medical Language System
Word2vec : word to vector is a well-known technique for constructing word embeddings
XGB : eXtreme Gradient Boosting

Introduction

1 The Scope of the Thesis

In the medical field, the amount of available information produced daily in medicine has increased significantly. The efficient and topical retrieval of this relevant patient and health status information is the main problem faced by any healthcare professional. Medical information systems collect huge amounts of textual and numerical information about patients, consultations, prescriptions, doctors' notes, medications, and more. This information, encapsulated in the data, can be used to make decisions about the healthcare system. This information is stocked in various file formats such as medical records, discharge reports, medical reports, etc., and may lead to improved quality of healthcare, accelerated clinical and research initiatives, reduced medical errors, and reduced costs.

In recent years, textual data has attracted increasing interest, with clinical text documents proliferating exponentially and more than 40% of data in medical records systems containing text [7]. These documents contain a wealth of valuable information about patient symptoms, diagnoses, treatments, drug use, and adverse events that can be used to follow the evolution of patients' health more accurately and therefore improve their care. A significant part of these documents exists in structured form, i.e., it can be stored and displayed in a strict and organized way, e.g., a patient's name, date of birth, height, etc. However, in computerized patient records, most of the detailed information is still stored in unstructured form: free text, where these medical data differ in complexity, length, and use of terminology. This complicates knowledge discovery and makes reuse of this data difficult, as well as doctors spend a lot of time searching for relevant patient information in these textual documents for medical decision making due to the lack of tools available to process them, also due to ethical policies regarding access to sensitive data.

In this situation, the lack of effective analytical methods and resources dedicated to clinical data significantly limits the possibilities for clinical data mining and analysis. At the same time, there is a real need for clinicians and health professionals to explore clinical data. To face these problems, Natural language processing (NLP) and text mining techniques offer a unique opportunity to extract important information from text data archives. Extracting this useful information from text using various types of statistical

algorithms is known as "text mining," "text analysis," or "machine learning from text."

In this regard, text mining (TM) offers a wide range of methods to explore this knowledge automatically using natural language processing (NLP) methods and machine learning techniques for transforming unstructured textual data into relevant structured information. A common application of text mining in healthcare domain is the classification of medical text or clinical document clustering, operate on the document level, making use of statistical and machine learning methods. Medical Text categorization has many applications, including categorizing risk factors and clinical alerts, classifying adverse drug reactions, categorizing electronic medical records, exploring patient symptoms, analyzing patient feelings, and classifying medical reports such as autopsy reports.

However, in many advanced applications, the main challenge in clinical medicine today is to use text mining tools to analyze and extract information from medical text data in social networks and medical forums, which provide valuable information for improving the quality of healthcare, monitoring drugs and tracking disease. This field is called medical sentiment analysis.

In recent years, the extraction of patients' opinions from medical forums on the web has seen different aspects that can be related to the following: the analysis of medical sentiment in the content tweets to assess people's opinions on patients' health status and their emotions about their medicines, as well as the identification of diseases in online communities for disease surveillance, to detect outbreaks and anomalies in blogs.

Throughout this thesis, all the brief notions presented above about text mining in medicine and especially related to report classification and medical sentiment analysis will be detailed by reviewing the main works in the literature. In the following section, we present our research goals and the scope of each chapter of our manuscript.

2 Summary of Research Goals and Contributions

In healthcare, the application of text mining with machine learning has become increasingly essential. Text mining techniques in medicine are very broad and can benefit patients and healthcare professionals in different areas of medicine. Many research works have addressed the approach of classifying medical reports from unstructured text data using NLP and machine learning methods, while others have focused on medical sentiment analysis. In this thesis, we focused on these two text mining applications in the medical domain.

Our first research objective was to study the benefits of using text mining and machine learning, including deep learning techniques, in the classification of medical reports. We collected a set of medical reports and studied the classification of autopsy reports or the detection of manner of death from autopsy reports. The objective of this work was to determine the manner of death automatically from these reports.

The second part of our thesis was the analysis of text mining and machine learning, including deep learning techniques, for opinion mining (OM) in the healthcare domain. Our goal was to demonstrate the effectiveness of these methods in sentiment analysis

and find the best model for the specific task. In recent years, there has been a growing focus on medical sentiment analysis, which offers a unique perspective as it can highlight diagnostic support systems and propose a new approach to improving the quality of medical devices.

Two major contributions were proposed in this part of our thesis. The first one was drug monitoring by analyzing machine learning and deep learning frameworks for opinion mining on drug reviews. The second contribution was the analysis of public emotions towards the COVID-19 pandemic in order to understand public reactions and aid epidemiologists in monitoring the spread of the virus. Our study proposes models for sentiment analysis on COVID-19-related tweets using traditional and advanced deep learning techniques and introduces a new hybrid model based on CNN and RNN for sentiment analysis.

Overall, our thesis explores the benefits and effectiveness of using text mining with machine learning, including deep learning techniques, in the medical domain for the classification of medical reports and medical sentiment analysis. More details and explanations about our research goals and contributions can be found in Chapter 2 and 3 of our manuscript.

3 Thesis Organization

The main aim of this thesis is to introduce text mining and sentiment analysis approaches in healthcare by providing simplified real-world examples. The intended audience of this paper is text mining researchers and medical professionals interested in understanding the underlying mechanisms and including them in future research work.

The manuscript is organized as follows:

- Chapter 1 provides a comprehensive overview of existing techniques and resources for performing text mining tasks in medicine. It describes the main practical applications, terminology resources, tools, and open challenges of this approach in medicine, with the goal of providing readers with the necessary knowledge of text mining in the medical field so that they can use it in real-world applications.

- Chapter 2 examines one application of text mining, document classification, and demonstrates the use of text classification techniques to predict the mode of death (MoD) from free-text forensic autopsy reports at Wilaya of Tlemcen using traditional and deep learning algorithms. The work is divided into two parts: part 1 discusses data collection and organization, and part 2 involves a comparative study of machine learning and deep learning algorithms for autopsy report classification.

- Chapter 3 represents the second part of the manuscript, focusing on medical sentiment analysis using NLP and machine learning techniques. This chapter provides a detailed overview of the basic knowledge and standard methods used to create sentiment analysis models. Proposed contributions in this area are detailed, including analysis of machine learning and deep learning frameworks for opinion mining on

drug critiques and analysis of public opinion on the COVID-19 pandemic. A comparative study was conducted in the two works between different machine learning and deep learning methods using well-known text vectorization techniques for exploring opinions, with the goal of finding the best model of OM.

- Finally, the conclusion and future directions section concludes the study and indicates future directions of research.

CHAPTER 1

Text-mining in medical health

1 Introduction

Text mining is the use of artificial intelligence techniques to extract useful information from large amounts of unstructured text. It is based on natural language processing methods that help to extract a more comprehensive understanding of meaning from the text. Text mining involves different aspects of linguistics, such as syntax (lemmatization, morphosyntactic labeling, parsing, etc.) and grammatical structure (noun or prepositional phrases, subject or object, ...), and can utilize knowledge representations like entity ontologies or synonym thesauruses.

The field of text mining encompasses various sub-domains, such as information retrieval, document classification, sentiment analysis, and more. In recent years, there has been growing interest in the application of text mining in healthcare, particularly in the field of medicine (medicine3.0), as it has the potential to improve diagnosis, treatment, and prevention of diseases. With advancements in technology, natural language processing and text mining methods are becoming increasingly important in automatically extracting valuable information from medical text data.

In the medical field, text mining techniques have been applied in various research studies such as clinical text summarization, disease prediction, and others. Clinical text summarization is an NLP task aimed at creating a summary from large amounts of clinical reports. Disease prediction using NLP involves mining unstructured patient health records for insights and information, which can aid in early detection, slow disease progression, and improve risk-adjustment procedures.

NLP-based models for disease prediction offer the potential to save money for healthcare insurers, as early treatments are typically less complicated and expensive than those given at later stages of a disease. In this context, the application of NLP and text mining in the healthcare domain is becoming increasingly important, as described in this chapter.

2 The problem of extracting medical information from text

In recent years, the daily production of information by healthcare professionals in medicine has increased rapidly. This information is stored in various sources and formats, mostly in textual form, and can be found in various documents such as discharge summaries, clinical monitoring sheets, clinical records, and medical reports. This textual information contains valuable knowledge that is essential for supporting medical decision-making.

Textual data stored in electronic medical records, clinical reports, and summaries has the potential to revolutionize health-related research and can be used for various purposes such as disease registers, epidemiological studies, monitoring pharmaceutical safety, clinical trials, and healthcare service audits. In most biomedical records, clinicians have the option of structuring their information or capturing it in free text.

However, the huge amount of mainly unstructured and non-standardized textual information presents a challenge for computer medicine. Extracting, discovering, and reusing the knowledge hidden in this data is one of the main challenges. In this regard, text mining (TM) offers a wide range of tools to extract this knowledge automatically using natural language processing (NLP) methods and machine learning techniques for transforming unstructured textual data into relevant structured information [1]. In the following sections, we will discuss the different applications of text mining in medical health.

2.1 Text summarization

The vast amount of textual information present in various medical sources is continuously growing. To help health professionals and researchers gain a better understanding of this vast amount of information, text summarization techniques have been developed. These techniques aim to analyze large amounts of information and generate a concise summary of the most important points in a document, by identifying its main themes.

There are different types of summarization methods, including extractive and abstractive summarization [8], that can be single or multi-document, general or domain-specific, and multimedia. These techniques help health researchers quickly and easily access essential information from multiple medical documents, reducing the time and effort required to wade through vast amounts of information.

2.2 Hypotheses generation and knowledge discovery

The use of text mining techniques for extracting knowledge hidden in medical text data is crucial for healthcare professionals in making informed decisions, including identifying risk factors for diseases, adverse drug events, symptoms, and important patient events.

For instance, in a study by Baron et al. [9], text mining was applied to a meta-analysis of 119310 articles to identify the adverse effects of aspirin use, simplifying the process compared to manual analysis. Similarly, Tafti et al. [10] developed a big data neural network system using NLP and text-mining with the word2vec algorithm to identify

adverse drug reactions from scientific articles on health-related social networks. This research led to the discovery of rare adverse effects, such as lactic acidosis caused by metformin use.

2.3 Medical Text Classification

Approximately 40% of medical information is stored in text form [7]. This information comes from various sources including computerized medical records, databases, articles, social networks, patient interviews, and biomedical literature. Technological advancements have enabled the computerization of medical records, interviews, and articles, and the exchange of information over the internet. As a result, the automatic classification of text has become an important task in medicine.

Text classification has many applications in the medical field, including categorizing risk factors and clinical alerts, classifying adverse drug reactions, categorizing electronic medical records, exploring patient symptoms, analyzing patient opinions and feelings, and classifying medical reports.

Most of the research in this area is based on three main applications [1]: automatic diagnosis, which uses text mining and machine learning techniques to classify diseases; patient stratification, which involves the analysis and classification of patient clinical characteristics from free-text medical reports using NLP methods and machine learning algorithms; and the classification of medical literature, which involves the creation of medical corpora through the automated collection and labeling of thousands of articles. An example of this is the classification of medical literature in the MEDLINE database.

2.4 Advanced systems for text-mining techniques

Advanced systems have integrated text mining (TM) tasks to identify concepts in large collections of biomedical text. Two examples of these systems are MetaMap ([1] and UIMA [2]. MetaMap uses the semantic base UMLS and UIMA is used to analyze unstructured information. UIMA's main role is to perform grammatical and multilingual analysis and document classification. These new TM tasks have great relevance in the healthcare domain, including improving the quality of care services, reducing time and cost associated with health management, and reducing medical errors.

2.5 Text-mining techniques on Health social media

The integration of new technologies in the medical field is crucial for advancing and improving medicine. Currently, one of the main challenges in clinical medicine is to utilize text-mining (TM) tools to analyze and extract information from medical textual data in social networks and medical forums, which provide valuable insights for improving the quality of healthcare, monitoring drugs, and tracking disease progression. The use of text-mining and artificial intelligence has allowed for a comprehensive analysis of this data. Some recent applications of TM include analyzing medical sentiment in the content of tweets to gauge people's opinions about patients' health status and emotions toward

[1] https://metamap.nlm.nih.gov/
[2] https://uima.apache.org/external-resources.html

their drugs, as well as identifying diseases in online communities for disease surveillance to detect outbreaks and anomalies from blogs.

3 Natural language processing and text mining

Natural Language Processing (NLP) is a field of artificial intelligence that enables computers to process and understand human language. This is accomplished by utilizing machine learning algorithms to analyze text, speech, and grammatical syntax. NLP aims to replicate natural human communication and can handle various forms of speech, including misspellings.

Text Mining, on the other hand, is a subfield of data mining that specifically deals with the extraction of information from text files. It encompasses various data mining and machine learning methods applied to textual information, including both structured and unstructured data. The main focus of text mining is on the structure of the data, rather than the meaning of the content. It is mainly used for the analysis of qualitative data.

4 Text-mining process

Generally, the majority of the most studied and applied methods and applications in the field of text mining go through the steps of the overall text mining process described in Figure 1.

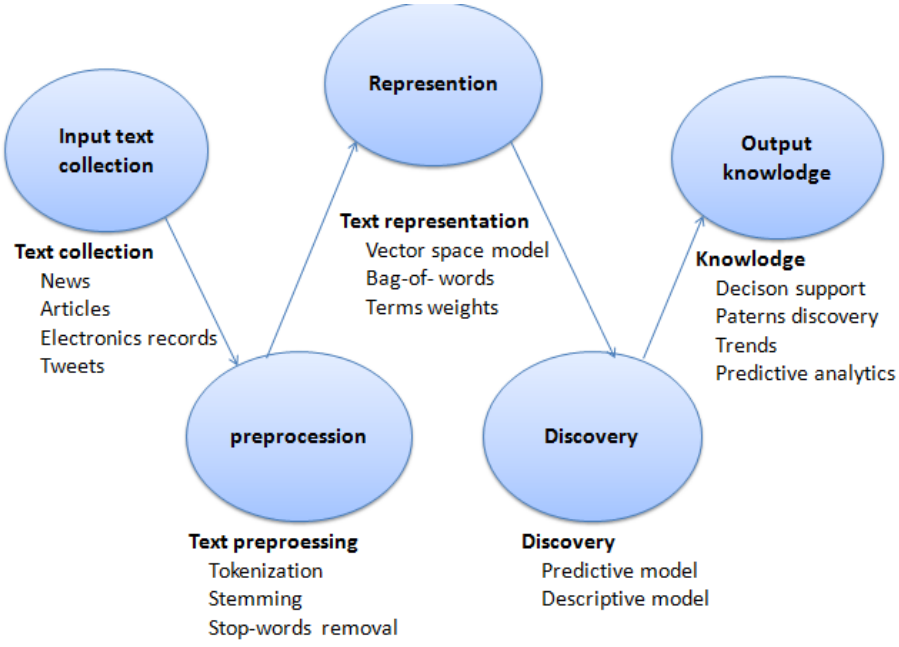

Figure 1: Text-mining process [1]

It is a process in which relevant and hidden information is extracted from textual data based on a series of steps:

1. **Data Collection and Preprocessing**, where unstructured textual input is cleaned and normalized using various Natural Language Processing (NLP) methods and converted into structured data.
2. **Text Representation or Vectorization**, where the preprocessed data is transformed into a vector representation model for identifying and analyzing patterns.
3. **Knowledge Discovery**, where important knowledge is extracted and discovered from the data using machine learning techniques such as classification (text/opinions), clustering, etc. The valuable information is then stored in a database.

In literature, there are several applications in the field of text mining, such as information extraction, text summarization, named entity recognition, etc. More recently, with the advancement of web technologies, text classification, which includes sentiment analysis, has become the most frequently applied in many fields, such as medicine. In the following section, we provide a brief overview of the basic components for the text classification task.

5 Basic building blocks for text classification

Natural language processing (NLP) refers to the intelligent processing of textual data through the use of linguistic computing tools to interpret documents written in natural language. In recent years, the term "text mining" has become more commonly used in reference to NLP, and refers to the use of machine learning tools and statistical, supervised and unsupervised classification algorithms for text analysis [7].

Text classification, including opinion classification, is a subfield of text mining. It involves automatically analyzing an incoming text and determining its category.

In general, the steps used in text classification can also be applied to sentiment analysis. This section will provide a detailed explanation of how to build a strong baseline for a text classification task.

5.1 Text-mining from structured/unstructured data

Structured data is information that is formatted according to a predefined structure, allowing it to be organized and analyzed. This type of data can be numerical or text-based, such as lists of standardized occupations and skills. On the other hand, unstructured data is the more commonly encountered form of information in organizations, which is stored in its original format without any specific processing. This data usually takes the form of textual documents, such as descriptions of radiology reports in various formats. [2].

Figure 2 provides an example of the world of structured data in medical health.

Typically, data mining applications utilize structured information, meaning unstructured data must be transformed into a structured format, such as a table, for information extraction tasks. For instance, in document classification, the initial collection of documents must be transformed into a spreadsheet by utilizing specific formats for data preparation, such as binary representation or TF-IDF transformations (as shown in Figure

	Systolic		disease
Gender	BP	Weight	Code
M	175	65	3
F	141	72	1
...	...	...	...
F	160	59	2

Figure 2: A spreadsheet example of structured data in medical health [2].

3). This structured representation allows the application of artificial learning methods for the classification task. These methods were designed to work only with sparse data in the cells of the spreadsheet.

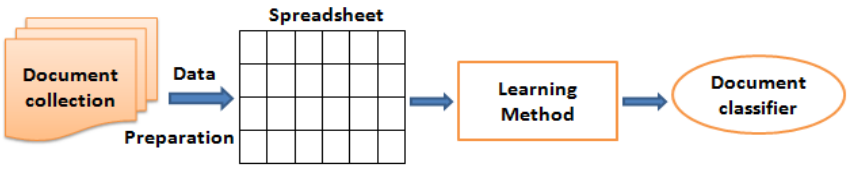

Figure 3: Learning to Predict from Text [3].

5.2 Segmentation and Tokenisation

In any text-mining framework, the first step is data segmentation and tokenization. The basis of the text is a string of characters which form words, sentences, and documents. To efficiently process natural language, it is necessary to represent the text formally. Hence, this step is based on the principle that the smallest unit of information in a sentence is a word, not a character. It involves representing the given text as a sequence of words by separating the terms, taking into account the other alphanumeric characters, white spaces, punctuation marks, and carriage returns contained in the text.

The second step in text-mining is tokenization, which involves breaking down an expression or text into its smallest units, called tokens. This step involves decisions on how to parse the input stream of characters, including whether sentence delimiters should be included, and how constructs like "let's" should be treated. For instance, in the case of a dosage expression like "400 mg/day", a standard tokenizer can be used and adapted to the specific domain, or a new tokenizer can be created from scratch.

As an example, the sentence "The patient has signs of COVID-19 in his left lung, let's try a new treatment with anakinra, 100 mg/day for 10 days." is tokenized as follows:

‘‘The’’ ‘‘patient’’ ‘‘has’’ ‘‘signs’’ ‘‘of’’ ‘‘covid-19’’ ‘‘in’’ ‘‘his’’ ‘‘left’’
‘‘lung’’ ‘‘,’’ ‘‘let’’ ‘‘"’’ ‘‘s’’ ‘‘try’’ ‘‘a’’ ‘‘new’’ ‘‘treatment’’ ‘‘with’’ ’‘‘anakinra’’
‘‘,’’ ‘‘100’’ ‘‘mg’’ ‘‘/’’ ‘‘day’’ ‘‘for’’ ‘‘10’’ ‘‘days’’ ‘‘.’’

5.3 Morphological Processing

Morphology is the linguistic processing of each token that focuses on analyzing word morphemes. There are several types of morphemes such as inflections, prefixes, infixes, and suffixes. The operations of morphological combination, including derivation, inflection, composition, and mixing, are also included. The purpose of morphological processing is to reduce the derivation and inflection forms of a term to a common base form, which facilitates the processing of terms and their meaning by many NLP systems.

5.3.1 Lemmatization

Lemmatization is a process that morphologically analyzes words in a vocabulary by removing inflectional endings and restoring the basic form of a word, which is known as the lemma. For example, "am," "are," "is" become "be." This process is useful for inflected languages, such as Swedish, German, Polish, etc. On the other hand, English has a simple morphology and does not usually require advanced lemmatization.

5.3.2 Stemming

The process of stemming involves removing the suffix of a word to obtain its root form. Different stemming algorithms exist that determine the number of characters to remove, but these algorithms lack the understanding of the meaning of the word in a language. As a result, the root form may not always be the actual word. For instance, the root form of the words "boat", "boater", and "boats" would be "boat". The Snowball system on GitHub includes several stemmers and also lists of stop words, which are meaningless and account for around 40% of the terms in a document [7]. Thus, it is advisable to remove stop words during the preprocessing phase.

5.4 Feature Engineering

A computer and machine learning algorithms do not operate similarly to a human brain. Unstructured text is simply a sequence of characters with grammar, which is meaningless to algorithms. To enable efficient processing of natural language by computers, it's necessary to represent the text in a format that can be processed by machine learning models, known as feature vectors. This process of representing the text is called feature engineering [11].

It is a crucial step in text classification algorithms and all NLP projects, and there are two main methods: *statistical and neural methods.*

5.4.1 Statistical Methods

Before the advent of word embeddings, statistical-based approaches were utilized for extracting features from text. Popular statistical techniques included word count matches and weight matrices, which were later used as inputs to machine learning algorithms.

Count Vectors as features

The aim is to represent the frequency of a particular term in a specific document. The most commonly used method for this is Bag of Words/Bag of n-grams [12]. This is a straightforward neural network that transforms a sentence or text based on the frequency

of defined keywords (tokens or n-grams of tokens) in a predefined vocabulary [13]. This method has the benefit of being simple and efficient from a computational perspective, but it's crucial to carefully select the vocabulary representation based on context and the learning program's vocabulary.

TF-IDF Vectors as features

TF (Term Frequency) and IDF (Inverse Document Frequency) [14]. The TF represents the frequency of a term in a given document and the IDF measures the significance of the term in the entire corpus or database. The final score is obtained by multiplying the TF and IDF scores. The TF-IDF algorithm is widely used in text classification and information retrieval tasks, as it provides a way to represent the importance of a term in a given text, while considering its significance in the entire corpus. The principle of TF-IDF assumes that if a word is important for a text, it must repeatedly appear in that document, whereas, it should rarely appear in other documents. Its concept is composed of two terms:

- Term Frequency (TF), representing the frequency of occurrence of a feature term in the text set,(computing the normalized Term Frequency)
- Inverse Document Frequency (IDF), is a measurement of the general importance of a term, which is offset by the frequency a term appears in the data set.

Equation 1.1 calculates the Term Frequency-Inverse Document Frequency, where $tf_{i,j}$ represents the number of times word i appears in document j. A higher value of $tf_{i,j}$ indicates that the word is important in document j. The parameter df_i is the number of documents in which word i appears at least once. The IDF value is calculated as the logarithm of the ratio of the total number of documents in the corpus N divided by the document frequency of word i. The output of the TF-IDF score ranges from 0 to 1, with words having a high score representing the important words in the corpus.

$$TF - IDF_{(i,j)} = tf_{(i,j)} * \log \frac{N}{df_i + 1} \tag{1.1}$$

Vector representation using TF-IDF can be performed at various levels of input tokens, including words, characters, and n-grams. The following outlines the different types of TF-IDF:

- Word Level TF-IDF: A matrix that gives the TF-IDF scores of each word in various documents.
- N-gram Level TF-IDF: N-grams represent combinations of N words together. The TF-IDF matrix therefore represents the TF-IDF scores of the N-grams.
- Character Level TF-IDF: The matrix gives the TF-IDF scores of the character level n-grams in the corpus.

5.4.2 Popular techniques of Neural Methods

include the use of word embedding techniques, which are forms of word and document representation using dense vector representations. This approach disregards the distribution of words and focuses on representing the meaning of words based on their context. It can be trained using pre-trained word embeddings such as GloVe, FastText or by using the input corpus itself.

These tools are neural network models that convert each term in the text into a vector, typically with 50 to 300 floating point numbers representing the input layer. The dense vectors capture and represent semantic similarity and syntactic information between terms. The algorithm performs word counting using simple distance measures such as the work of the Word2vec and GloVe methods [6].

Word2vec

W2V is a well-known technique for constructing word embeddings and is one of the most successful algorithms in this field. It was introduced by Mikolov et al. [4] in 2013. It is a neural network that processes text data and builds a vocabulary from the training corpus. The algorithm generates distributed word vectors in a high-dimensional vector space. These vectors are then used to form a matrix of sentences, and are ultimately utilized as features for machine learning models, including both traditional and deep learning methods.

There are two variants of the W2V algorithm: Skip-gram and Continuous Bag-of-Words (CBOW). Both methods use a neural network architecture consisting of three layers: an input layer, a hidden layer, and an output layer. The output layer is composed of neurons with a softmax activation function, and these architectures have been shown to result in high-quality term embedding vectors [15].

- **CBOW** architecture is a method for predicting a target word based on its context, as defined by a sliding window that encompasses the word and the words surrounding it. The projection layer is shared by all words in the context. If the vocabulary size is W terms of dimension w, a one-hot encoding is used to represent each word. Each term can have a binary vector of w dimensions. Given a vector of k hot-encoded context terms, the CBOW algorithm computes the sum of the embeddings of the context words to predict the target word. A log-linear classifier with a hidden layer of N dimensions is used to predict the target word vector of dimension w based on the context words, which are connected and have vocabulary dimension w. After training the classifier, each term has a vector in the $N \times W$ weight matrix, which is connected to a softmax layer to predict the target term. This method aims to maximize the likelihood of the prediction, but works best for frequent terms, as more training data is available for words that appear more often.

- **Skip-gram** architecture is a log-linear neural network consisting of three layers, unlike CBOW. It predicts the context window by using the word at the center of the context. The word at the center, of dimension w, serves as the input of the network (hidden layer of size N), generating a layer of continuous projections. These

projections are then connected to a softmax function, each with a dimension of w, to predict context words k, which are the output of the network (see Figure 4). The softmax function is modeled as a k multi-class classifier. The Skip-gram model assumes that the embedding of each term will be close to the terms in the same context. To train the classifier, the size of the window before and after the term must be defined by giving high weight to nearby words and lower weight to distant words.

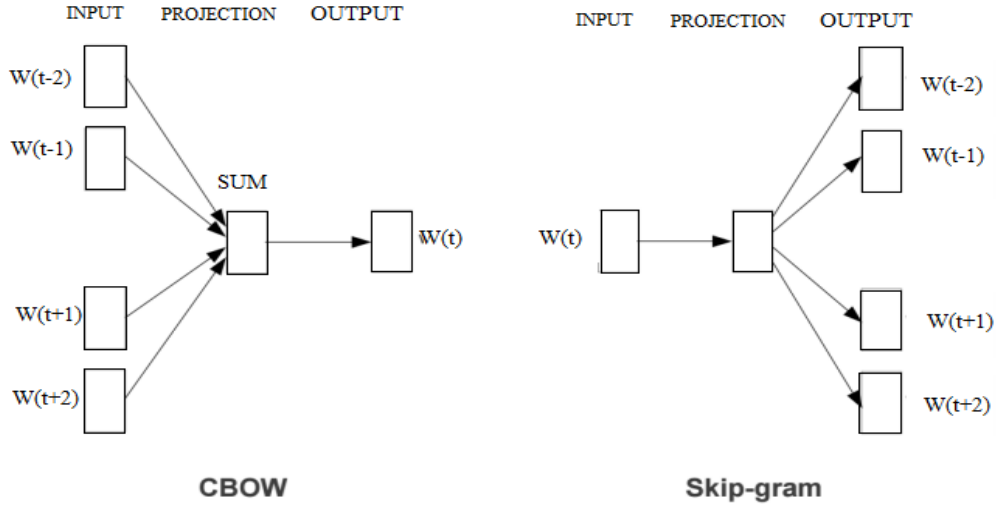

Figure 4: CBOW and skip-gram model as presented by Mikolov et al. [4].

Mathematically, Skip-gram architecture aim to maximize the following formula (eq. 1.2) given a set sequence of terms w_1, w_2, w_3,..., w_T

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq cj\neq 0}\log p(w_{t+j}|w_t) \tag{1.2}$$

Where: T represents the total number of training words, j is the index of a word in the context window, w_t refers to the target word, and c is the size of the context window. It is worth noting that a larger c value results in more training samples, which can lead to higher accuracy, but also longer training time. To calculate the probability $p(w_{t+j}|w_t)$, the softmax function as defined in equation 1.3 is utilized.

$$p(w_0|w_1) = \frac{\exp(\acute{v}_{w0}v_{w1}^T)}{\sum_{w=1}^{W}\exp(\acute{v}_w v_{w1}^T)} \tag{1.3}$$

where v_w and v'_w are the vector representations of word w for the input and output layers, respectively, and W is the size of the vocabulary.

The Skip-gram model is more suitable for smaller datasets and performs well with infrequent words.

GloVe GloVe (Global Vectors) is a recent approach developed by researchers at Stanford University in 2014 [16]. It involves constructing a global word co-occurrence matrix

by processing the corpus with a sliding context window. Each element in the matrix represents the number of times word i appears in the context of word j. GloVe is an unsupervised learning model that considers all the information in the corpus, not just the information in a window of words, hence the name Global Vectors. After the matrix is constructed, a least squares regression model is trained to generate vector representations. The developers of GloVe suggested pre-incorporating millions of English tokens from Wikipedia and common crawl data.

Mathematically, GloVe trains word vectors from the co-occurrence matrix using the following objective function (eq. 1.7):

$$J(\theta) = (u_i^T v_j - \log P_{ij})^2 \tag{1.4}$$

Where $J(\theta)$ is the objective function that depends on the parameter θ, which are the word vectors. u_i and v_j represent the input and output word vectors, respectively, that correspond to a word's row and column in the co-occurrence matrix. P_{ij} is the count of the number of times that the words i and j appear together.

The goal of GloVe is to optimize the word vectors by minimizing the difference between the dot product of the vectors for words i and j and the logarithm of their co-occurrence count squared.

FasText FT, or FastText, is an embedding method introduced in [17] and is an extension of the word2vec method [4]. Unlike word2vec, which considers words as unbreakable atomic units, FastText considers words as bags of character n-grams. FastText represents words by summing the vectors associated with their character n-grams, thus allowing for extraction of more semantic relationships between words that share common n-character grams. FastText can also generate embeddings for rare words that have never been seen before by summing its known character n-gram vectors. Character n-gram embeddings have been shown to perform better than word2vec and GloVe on smaller datasets [18].

Deep contextualized models In recent years, researchers in the field of text mining have shifted their focus towards a new model of feature representation that considers the context in which words are used [19]. One such model is the Bidirectional Long Short-Term-Memory (Bi-LSTM) network [5], which generates an Embeddings from Language Models (ELMo) representation by taking an entire sentence as input and training a coupled language model. This approach has proven to be effective in incorporating information from the sentence into the ELMo representation, leading to improved results in deep contextual models.

Currently, there are several pre-trained neural network models that have been proposed and shown to perform well on a variety of linguistic tasks, such as BERT [20] and OpenAI GPT [21].

5.5 Use of machine learning Models

Text categorization is a task in which text data is classified into different categories. It is often modeled as a classification problem in which a classifier is fed with text data and returns the corresponding category. The text data is preprocessed and vectorized before being fed into the classifier. The machine learning algorithm then predicts the class of the given text. There are two main types of text classification:

1. **Binary text classification:** In this case, each document d_i in D, where $D = d_1, d_2, ..., d_n$, is classified into one of two categories. For example, the categories could be "politics text" and "medical text," or "negative" and "positive" in the case of opinion classification.

2. **Multi class text classification:** In this case, each document d_i is classified into one of several categories in C, where C represents the set of categories. The level of the corresponding category for each document is represented by C.

The most commonly used statistical methods in text categorization are supervised learning techniques. This type of method involves representing each document as a set of variables generated through text vectorization tools. A model is built using examples of text with known labels (such as the polarity or category of each text), and this model is then used to assign the corresponding polarity/class to a new, unlabeled document.

These supervised learning techniques can be divided into two groups: traditional or classical methods, and deep learning-based models [22]. The traditional classifiers used in this field include the Naive Bayes classifier (NB) [23], Support Vector Machines (SVM) [24], k-Nearest Neighbors (KNN) [25], Random Forest (RF) [26], Logistic Regression (LR) [27], and Maximum Entropy (ME) [28]. On the other hand, deep learning methods include Convolutional Neural Networks [6] (CNN) and Recurrent Neural Networks (RNN) models [29]. Figure 5 shows the sentiment polarity classification process using traditional machine learning methods.

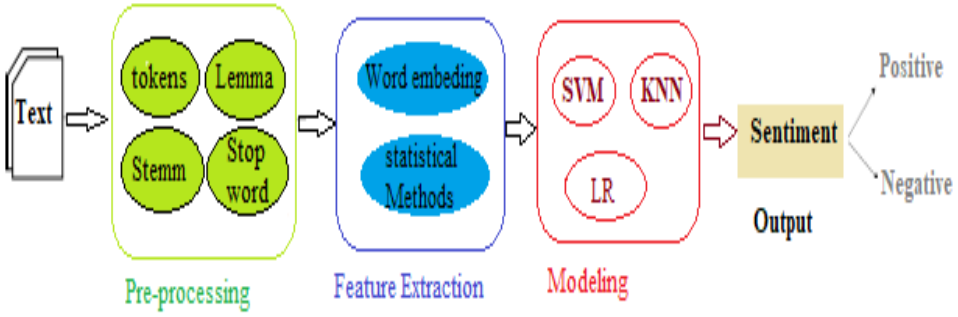

Figure 5: sentiment polarity classification using machine learning

Text clustering can also be performed using an unsupervised machine learning approach, utilizing opinion lexicons and grammatical analysis [6]. In this case, the documents are not labeled and the clustering process doesn't require supervised information to categorize the text. After transforming the text into numerical vectors or matrices, an unsupervised machine learning tool groups reviews with similar properties into a cluster. These methods can be classified into two types: partition clustering, which divides the database into non-overlapping subclusters, and hierarchical clustering, which involves the hierarchical decomposition of the database [30].

In situations where a small amount of labeled data is available, and a large volume of data is unlabeled, especially if the labeling is done manually, a semi-supervised machine learning approach can be used. In this approach, the unlabeled examples are transformed into labeled points for further analysis. The classifier trains on the small labeled dataset,

and based on these values, it predicts the classes of the unlabeled dataset. These data are then added to the training set until all the data is classified.

Classical machine learning techniques have been widely used in various text classification studies in different domains. The performance of results depends on the classifier's capacity and the best representation of the data, which is influenced by the choice of feature extraction method [31]. Creating an effective representation of the data requires a high level of expertise in the domain. However, deep learning models have emerged as a promising solution in the automatic extraction of complex semantic features of text data without the need for a feature engineering step. These models have been widely adopted by researchers to solve text-mining problems, including opinion classification, text classification, and information extraction.

5.6 Deep learning methods for text classification

Recently, with advancements in technology, deep learning (DL) has become a widely studied field in various areas, particularly in image and speech recognition and text mining. DL techniques have particularly been effective in text classification tasks, where they play a crucial role. Unlike traditional machine learning methods, which rely solely on text vectorization techniques for feature extraction, deep learning techniques leverage multi-layer approaches to extract features in the hidden layers of a neural network architecture, resulting in improved performance and accuracy. Figure 6 illustrates opinion polarity classification using deep learning methods. This section discusses various deep learning-based text classification techniques, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs).

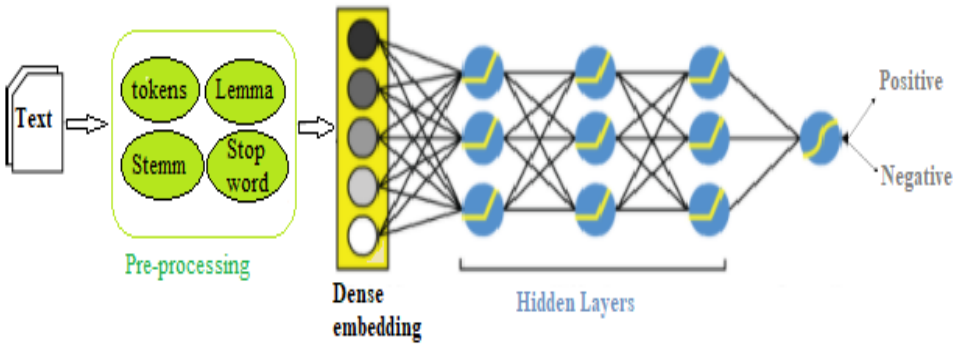

Figure 6: opinion polarity classification using deep learning approaches

5.6.1 Convolutional Neural Networks (CNN)

A convolutional neural network (CNN) is a type of feed-forward neural network that is widely used in the field of deep learning, particularly in image processing. CNNs have proven to be effective in sentiment analysis tasks as text data is treated as an image and used as input for classification.

The CNN architecture consists of three main layers: the convolution layer, the pooling layer, and the fully connected layer. The convolution layer extracts features from the

input using multiple filters, the pooling layer reduces the dimensionality of the feature maps by keeping a simple probability score that reflects the likelihood corresponding to a label, and the fully connected layer uses back-propagation to make the final classification decision.

The figure 7 shows the process of how an input integration matrix is processed by a CNN consisting of three convolution layers and two pooling layers. The first convolution layer uses 64 filters to extract different features, followed by a pooling layer to prevent overfitting. The second and third convolution layers use 32 and 16 filters, respectively, followed by another pooling layer. The final fully connected layer predicts the classes by reducing the height vector of 16 to an output vector of one.

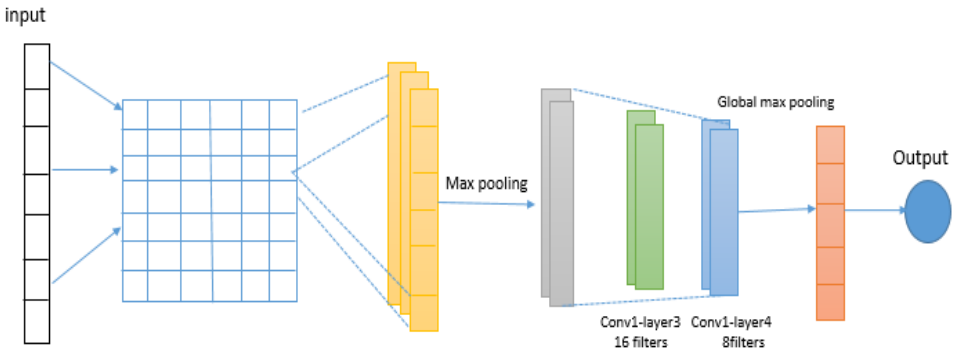

Figure 7: Convolutional Neural Network Algorithm.

5.6.2 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNN) are a type of deep neural network designed for processing sequential data. They are capable of capturing the correlation between current and previous time steps by using a memory mechanism created by feedback loops within the network. At each time step t, the RNN has a hidden state ht that represents the state of the input processing system. The use of time t allows the network to convert the input sequence into the final output sequence (eq. 1.5). RNNs have been widely used in various NLP tasks, including sentiment analysis ([6]).

$$ht = f(h_{t-1}, X_t) \tag{1.5}$$

With: h_{t-1}: the output at $t-1$ and X_t: the current input at t. For each input X_t at time t, a non-linear function f helps to predict the system status at time t using the status at time $t-1$, f is meant as a linear transformation function added to a non-linear activation function (eq. 1.6) which has the following form:

$$ht = tanh(W[h_{t-1}, X_t] + b) \tag{1.6}$$

With: W: represent the weight, b: the bias, $tanh$: the hyperbolic tangent activation function and ht: the output. This type of network is necessary in solving time series tasks such as speech recognition, natural language processing (NLP), machine translation, video sequence processing.

5.6.3 Long Short-Term-Memory (LSTM) and Bidirectional Long Short-Term-Memory (Bi-LSTM)

Long-term memory (LSTM) is a popular type of recurrent neural network and is considered more effective than GRU for processing longer sequences. It was first introduced by Hochreiter and Schmidhuber in 1997 [29]. LSTM extends the basic RNN architecture by adding memory cells in the form of a collection of interconnected subarrays.

The LSTM architecture consists of four fundamental components: the input gate (it), the output gate (ot), the forget gate (ft), and a memory cell that controls the memory length and the write, read, and reset operations for the cell. Each gate is represented by a sigmoid function.

For each sequence of input vectors (xt) at time t, the forget gate (ft) decides which information from the previous output should be deleted by considering the current inputs (xt) and the previous hidden state ($ht-1$). The input gate (it) decides which information should be entered into the cell state (ct), and the output gate (ot) decides which part of the cell state should be passed to the next hidden state. These three sigmoid functions enable the memory cells to store, update, and access information from multiple training examples. A visual representation of the LSTM architecture is shown in Figure 8.

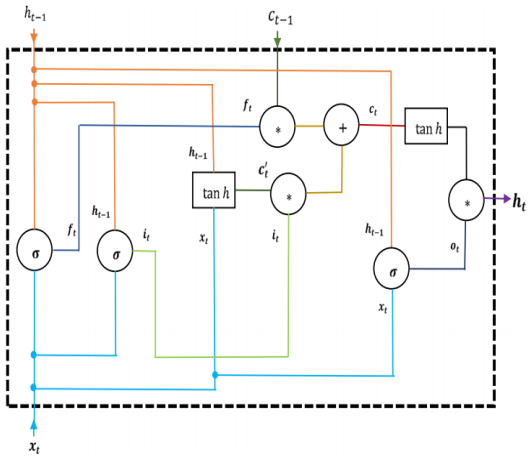

Figure 8: LSTM Architecture presented by A. Graves [5]

Long Short-Term-Memory (LSTM) is a powerful technique for natural language processing, particularly for sequential prediction tasks such as text analysis and classification. It uses a special mechanism to selectively process its memory, deciding which parts of the information to retain, forget, and update based on the memory vector and partial output. This allows the network to learn from longer input sequences and mitigate the risk of exploding gradients.

However, one limitation of the LSTM model is that it only has access to information from the past, not the future. To address this issue, the Bidirectional LSTM (Bi-LSTM) was introduced. This model uses information from both the past and future by connecting hidden layers in opposite directions. The Bi-LSTM consists of two parallel LSTMs, each processing information in one direction, and the results from each network are combined to make the final prediction.

5.6.4 Gated Recurrent Units (GRUs)

The Gated Recurrent Unit (GRU) is a simplified version of the Long Short-Term Memory (LSTM) network, introduced in 2014 by [32]. It handles sequential data and is designed to avoid the exploding gradient problem [6]. The GRU architecture consists of only two gates, the update gate, which captures dependencies in the input sequence, and the reset gate, which captures short-term dependencies. This reduced complexity compared to LSTM leads to a smaller number of parameters and allows for an increase in the number of hidden states, making the GRU better suited for handling large data sets, such as sentiment analysis.

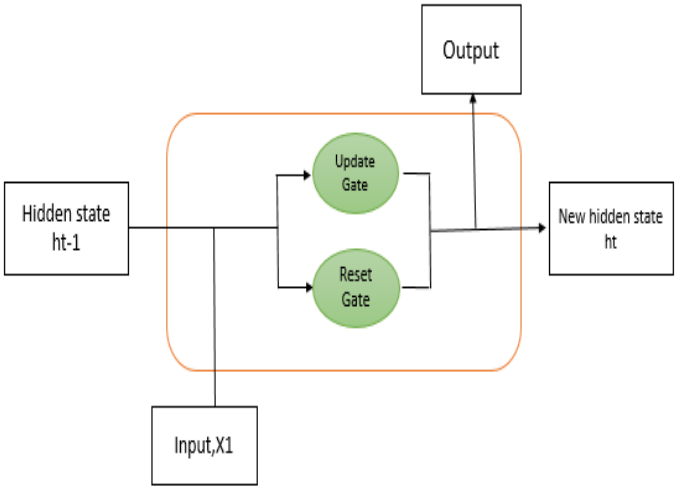

Figure 9: Gated recurrent cell architecture [6]

Figure 9 illustrates a simple architecture of the GRU cell, which has two gates. The update gate uses a sigmoid layer (eq. 1.7) to determine which new information should be added or dropped, by considering both the inputs xt and $ht-1$. The reset gate, on the other hand, determines the amount of memory to be discarded, using a sigmoid layer (eq. 1.8) and taking into account both the inputs xt and $ht-1$. The Tanh layer (eq. 1.9) computes the proposed $\tilde{h}^{<t>}$, while the final memory (eq. 1.10) is calculated by evaluating what to keep from the current memory and previous steps.

$$\Gamma_u^{(t)} = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u) \tag{1.7}$$

$$\Gamma_r^{(t)} = \sigma(W_r[a^{<t-1>}, x^{<t>}] + b_r) \tag{1.8}$$

$$\tilde{h}^{<t>} = \tanh(x^{<t>}W_h + (\Gamma_r^{(t)})W_h + b_h) \tag{1.9}$$

$$h^{<t>} = \Gamma_r^{(t)}.c^{<t-1>} + (1 - \Gamma_r^{(t)}).h^{<t>} \tag{1.10}$$

5.6.5 Attention model

In recent years, various modifications have been made to the basic recurrent neural network structure, resulting in significant advancements in state-of-the-art performance. The

attention mechanism, introduced by [33], has become particularly popular in natural language processing tasks. This model uses a single hidden layer to determine the importance of each hidden state, and it combines the input features into a weighted sum. The attention mechanism operates using the encoder-decoder principle. Figure 10 illustrates a simple sequence-to-sequence model that employs the attention mechanism in an automatic translation task. The model consists of a Bi-LSTM layer for encoding, an LSTM layer for decoding, and a single attention layer between them. The feature vectors $a^{(t)}$ from the encoding Bi-LSTM are transmitted to the attention mechanism, which calculates the context vector $c^{(t)}$ as a weighted sum of the BLSTM features. The context vector is then passed to the LSTM layer for decoding, and the feature vector is computed using the softmax function.

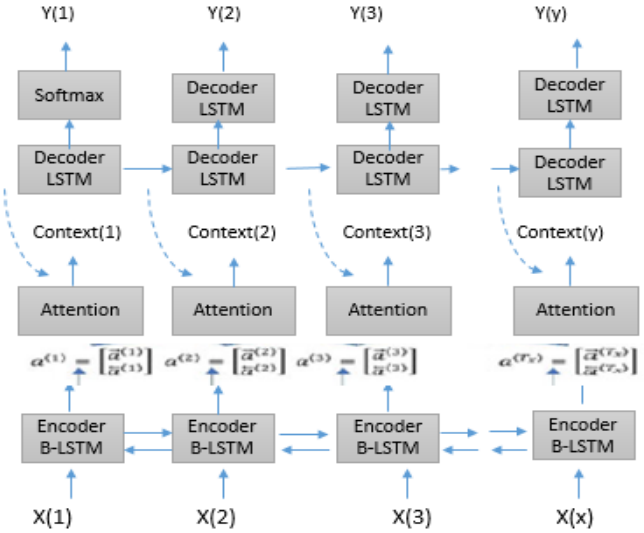

Figure 10: Attention model [6]

5.6.6 Other Deep Neural Networks

Recent developments in deep neural networks have played a key role in a variety of natural language processing tasks, such as sentiment analysis and text classification. One such network is the transformer model, which expands upon the attention mechanism models [30]. The transformer still uses the encoder-decoder principle along with the attention mechanism, but is based on pre-trained contextualized representations, such as BERT [20] and RoBERTa [34]. Another type of deep neural network is the Deep Belief Network (DBN) [35], which is a multilayer model with directed and undirected edges. Hybrid deep learning [36], which combines multiple deep learning techniques in a single architecture, has also achieved remarkable results in several opinion mining tasks. For example, combining RNNs (such as LSTMs [29]) with CNNs can take advantage of the strengths of both models: RNNs can learn dependencies of long sequences, while CNNs can extract relevant, high-level features. Another example of hybridization is combining Probabilistic Neural Networks (PNN) and a two-layer Restricted Boltzmann Machine (RBM) [37].

6 Conclusion

In this chapter, we focus on text mining techniques for medical diagnostic support, specifically text classification tools such as medical report classification and medical sentiment analysis. These automated methods can greatly aid medical professionals in their decision-making processes, such as monitoring patients' emotions towards their medications through drug monitoring or discovering new knowledge through the automated classification of medical reports like autopsy reports. The aim of this chapter is to provide readers with a comprehensive understanding of medical text mining using machine learning methods, in order to help them effectively apply these tools to real-world data.

CHAPTER 2

Autopsy reports classification : study of mortality and causes of death in Wilaya of Tlemcen

1 Introduction

In recent years, with the advancements in technology and artificial intelligence, bioinformatics researchers are highly interested in studying the actual needs of doctors by collecting real medical data, such as medical images, medical reports, and numerical data, to uncover hidden knowledge in this large volume of data, known as big data, as well as correlations between these data.

One of the most widely studied population-based studies is the epidemiology of forensic deaths in a given population, achieved through analyzing death certificates. These certificates are forensic documents prepared by pathologists that contain external and internal examinations of the deceased person's body, as well as other information about the person. The objective of these reports is to determine the primary cause of death (CoD).

Forensic autopsy reports serve as a valuable source of information in criminal and civil investigations, and analyzing multiple reports can provide timely warnings about disease activity. However, this is only useful if the reports contain accurate and quantitative data. Preparing an autopsy report can be challenging and time-consuming, with an initial version ready in two to three days and a final report taking up to three months if it's complex. This can result in inconsistencies in the detection of CoD in a short time.

Recently, text-mining solutions have helped minimize time consumption, inconsistencies, and irregularities. Text mining and artificial intelligence techniques have provided the capability to automatically predict the cause of death from free-text medical autopsy reports.

In this study, text classification techniques were used to predict the manner of death (MoD) from free-text forensic autopsy reports in Wilaya of Tlemcen using traditional and deep learning algorithms. Firstly, the basic concepts of forensic medicine are briefly introduced, followed by the proposed work, which is divided into two parts: the first part

involves data collection and organization, and the second part involves the application of machine learning and deep learning algorithms for the classification of autopsy reports using the proposed TF-IDF feature extraction method. The goal is to improve the performance of autopsy report classification by finding the best classification model in terms of accuracy and a combination of methods.

2 Forensic Medicine

Forensic medicine is a branch of medicine that serves as a convergence point between medicine (health) and law (justice). It is carried out by forensic doctors and aims to fulfill a social and legal need for discovering the truth. Unlike therapeutic or preventive services, it does not provide a cure or prevention.

One of the primary questions in forensic medicine is how to differentiate between natural death and violent death by conducting a clinical description of death and an investigation into the mechanisms and causes leading to death, through autopsies of the body of individuals who have died under unclear circumstances, to enlighten the judge.

Autopsies or post-mortem examinations make a significant contribution to health education and enhance the quality of healthcare. During the autopsy examination, pathologists examine the external and internal parts of the deceased body and gather information about the organs. Additionally, pathologists gather information about the deceased's personal details, injuries, histopathology reports, and medical history to distinguish between different forms of death: natural or violent.

2.1 The forensic autopsy

The forensic or medico-legal autopsy is a critical component of the criminal justice system and the medical community. Its main goal is to determine the cause of death (CoD) by thoroughly examining the physical evidence present in the deceased person's body. The autopsy serves several important functions, including:

1. **Establishing the medical cause of death:** The autopsy provides detailed information about the deceased person's medical conditions, helping to determine the cause of death and contributing to the improvement of medical practice.
2. **Determining the form of death:** The autopsy helps to categorize the form of death as either a homicide, suicide, accidental, or natural death, which is critical for the criminal justice system to pursue appropriate legal action.
3. **Determining the date of death:** The autopsy provides information about the estimated time of death, which is crucial in criminal investigations and legal proceedings.
4. **Identification of the deceased:** In some cases, the deceased person may be unidentified. The autopsy can provide important information that can assist with identification, such as dental records, medical history, and physical characteristics.
5. **Determining the method used in cases of violence:** In the case of a violent death, the autopsy helps to determine the instrument used by the aggressor, contributing to the criminal investigation and prosecution of the perpetrator.

Overall, the medico-legal autopsy serves as a vital tool in uncovering the truth behind a person's death and improving the functioning of both the medical and criminal justice systems.

2.2 The medical autopsy

Scientific autopsies, also known as research autopsies, are performed with the aim of advancing medical knowledge and improving patient care. They involve the removal of organs for transplantation or for further study, and they provide families with a detailed understanding of the cause of death. The comparison of clinical observations made during the person's lifetime with the findings of the autopsy helps to reveal any underlying medical conditions or abnormalities that may have contributed to the person's death. The results of scientific autopsies can aid physicians in resolving medical problems and improving patient care.

2.3 The causes of death

he death can be classified into various categories: death due to illness, death as a result of an accident, suicide, and murder. These are grouped into two broad categories of cause of death: *natural and violent death*.

2.3.1 Natural death

Natural death refers to the death caused by a physiological condition such as old age, illness, or other natural causes. However, sometimes a seemingly healthy individual may die suddenly, and in such cases, an autopsy is often the only way to determine the true cause of death. In the forensic context, natural deaths are mainly caused by cardiac issues, leading to sudden death, but it can also occur due to other causes, including:

- Nervous system-related causes: accounting for 27% of natural deaths, including various hemorrhages, tumors, cancer, and aneurysm ruptures.
- Digestive causes: digestive hemorrhage and
- Respiratory diseases: accounting for 7% of natural deaths, including pneumonia, bronchopneumonia, and pulmonary embolism. Additionally, death can occur due to cancers and metabolic disorders such as diabetes.

2.3.2 Violent death

Violent death is defined as an unnatural death resulting from a deliberate external intervention (physical or toxic, by a person, machine, or product) in circumstances that can be criminal (homicide), accidental (a brutal external cause), or self-inflicted (suicide). The diagnosis in this case is not always straightforward [4]. In the context of violent death, the causes usually refer to an injury or wound.

A. Injuries: Injuries in the context of forensic medicine are classified based on the nature of the cause, the location on the body, and the type of tissue affected. These classifications help to understand the mechanism of injury and to determine if the injury was inflicted by a criminal act or was accidental. The different types of injuries include:

- Contusions: bruises or superficial injuries caused by blunt trauma,
- Dermabrasions: injuries caused by abrasion of the skin,
- Wounds: cuts, lacerations, stab wounds, and puncture wounds,
- Burns: injuries caused by thermal, chemical or electrical agents,
- Fractures: breaks in the bones caused by traumatic force.

The examination of the pattern, shape and location of the injury can provide valuable information in the investigation of a violent death, such as the type of weapon used or the position of the victim at the time of injury (see Figure 11).

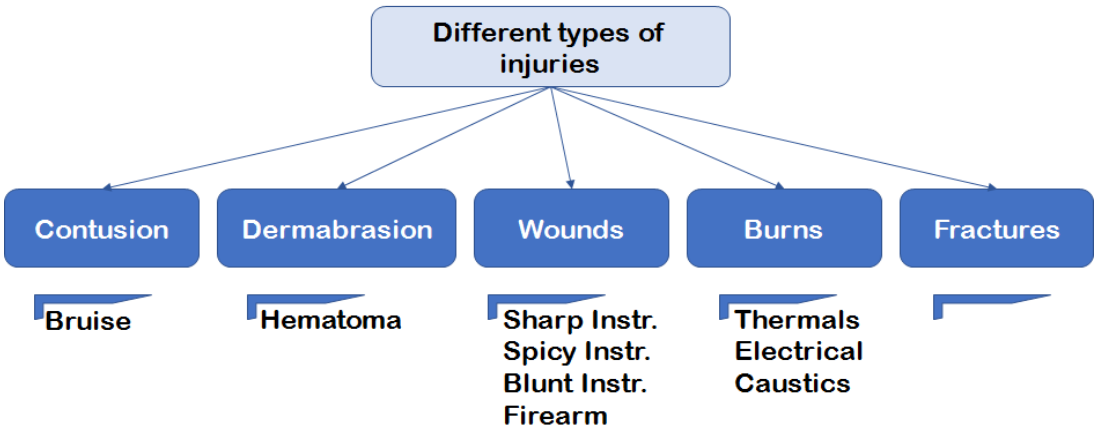

Figure 11: Different types of Injuries

- **Contusion :**
  is a type of injury in which the tissues are crushed by a forceful impact that does not cause a break in the skin or surface of the organ, often caused by a strong punch or intense pressure. This leads to a rupture of small blood vessels, resulting in the formation of bruises or hematomas along with crushed cells.

  Ecchymosis is a type of intramuscular and subcutaneous bleeding that doesn't fade under pressure. Over time, bruises change color and appearance, providing valuable information to the forensic scientist in determining the approximate time of the injury on the body.

- **Dermabrasions :**
  also known as abrasions, excoriations, or scratches, which are important factors in the description of the external body during an autopsy. During an autopsy, the medical examiner will perform punctures at the site of bruises to look for underlying hematomas.

  Hematoma is the later stage of a bruise and refers to collections of encysted tissue blood. It's the intensity and repetition of the trauma that results in increased blood flow, such as in a cerebral contusion.

- **Wounds:** is a disruption of the skin or mucous membranes caused by a mechanical agent. It results in a bleeding from the escape of blood components and the flow of blood outside the tissues. Wounds can be classified as follows:

  - Simple wound: characterized by no loss of tissue, clean and well-defined edges, typically caused by sharp cutting instruments.
  - Contused wound: characterized by bruised, ragged or irregular edges, caused by blunt and sharp instruments.
  - Complicated wound: characterized by additional factors, such as involvement of deeper structures or infection.

- **Burns :** Extensive burns can lead to high mortality rates and potential severe and long-lasting consequences. Burns can be classified into two types: thermal burns and electrical burns.

  Thermal Burns: In cases where burns cause death, it raises suspicions. Thermal burns can be classified into five different depths:

  - 1st Degree: Superficial burns, resulting in a visible redness on the skin.
  - 2nd Degree Superficial: Blisters on the skin, destruction of the epidermis.
  - 2nd Degree Deep: Destruction of the superficial layer of the epidermis.
  - 3rd Degree: Total destruction of the epidermis and the dermis.
  - Carbonization: A unique appearance of the body, often referred to as "boxer's posture" which includes flexion of the thighs on the pelvis and forearms on the arms.

  Electrical Burns: These burns occur at points of contact between the body and an electrical conductor and also from sparks produced by an electric arc. Electrical burns can be characterized by two types of lesions:

  - Entry wound: small, round, pale yellow lesions with raised edges and a slightly depressed center, hard to the touch and appears to be encrusted in the skin.
  - Exit wound: more circumscribed, presenting as small necrotic or ulcerated areas on the soles of the feet.

B. Toxic death Toxicology, the science of poisons, has been a long-standing specialty in forensic medicine. In cases of suspected death by poisoning, the forensic scientist must request a toxicological analysis and collect toxicological samples to obtain a comprehensive toxicological profile of the deceased.

In medicine, a poison is defined as a substance that temporarily or permanently disrupts the vital functions of an organism, leading to death. Toxicology considers several anatomo-clinical characteristics, including:

- Hepatic syndrome: sub-icterus and hepatic insufficiency.
- Toxic hepato-nephritis syndrome: digestive disorders, jaundice, oliguria or anuria.
- Ocular syndrome: helps in diagnosing three toxins: botulism, alcohol, atropine.
- Acute encephalopathy: characterized by delirious, convulsive, or comatose forms.
- Respiratory form: characterized by symptoms such as coughing, dyspnea, and pulmonary edema.

C. Mechanical asphyxias: (Asphyxia syndrome) It is the interruption of breathing due to a mechanical obstruction of the airways. It presents with respiratory distress caused by the blockage of air flow. The following are its types: suffocation, hanging, strangulation, and drowning.

Necropsy Findings of Mechanical Asphyxia:

- External Examination: Blue discoloration of the face, lips, ear lobes, and extremities, as well as subconjunctival hemorrhages, and sometimes hemorrhagic marks.
- Internal Examination: Air-filled foam in the trachea, larynx, bronchi, congested red mucosa, and Tardieu spots (a characteristic appearance of the lungs).

3 Related works

In recent years, unstructured text classification has gained attention in the medical field as a means to categorize clinical reports into predetermined categories [38]. One important area in need of classification is that of medical autopsy reports, which provide valuable information about a deceased individual. Despite its importance, there is limited research on the classification of autopsy reports. Researchers have attempted to tackle this challenge by exploring different approaches, including novel feature extraction methods and text classification techniques, in an effort to identify the most efficient and accurate model.

Danso et al. [39] carried out a comparative study to evaluate the performance of various text classification algorithms and text vectoring techniques in determining the CoD from verbal autopsy reports. The results of their experiment showed that the SVM decision model combined with inverse document frequency (IDF) and normalized IDF (NTFiDF) for feature representation produced the best results in terms of accuracy and speed.

Similarly, Yeow et al. [40] used case-based reasoning (CBR) in combination with a naive Bayes classifier to determine the CoD from forensic autopsy reports, with an accuracy of 80%. However, the study was limited by two major factors: first, the researchers only used a summary of the autopsy reports, instead of the full detailed reports, in the training set; second, the work did not consider the issue of synonyms at the word level to extract similar concepts from the core features.

Shahid et al. [41] developed an automatic text classification system for categorizing autopsy reports. In their comparative study, they evaluated different automatic text classification (ATC) techniques by integrating feature extraction and feature reduction methods. The results of their case study showed that feature reduction approaches improved accuracy.

Mujtaba et al. [42] used various machine learning-based text classification methods to determine the cause of death (CoD) from a dataset of complete forensic autopsy reports. The researchers compared the performance of the different classifiers, using several feature extraction models, to find the most optimal representation. The results showed that the best accuracy was 78%, with the SVM classifier outperforming the other algorithms. In terms of feature representation, the unigram text vectorization model was found to be more effective than the bigram and trigram models. However, this study had a low

accuracy due to its reliance on traditional text classification algorithms to categorize autopsy reports. In comparison, contemporary studies tend to use deep learning approaches, which have been found to produce more accurate results for clinical report classification.

Duarte et al. [43] introduced a deep neural model to code the ICD-10 text that describes causes of death. The model integrates word embeddings, recurrent units, and neural attention for feature representation, and analyzes unstructured text descriptions from death certificates, autopsy reports, and clinical bulletins. The results of the experiments showed that the proposed model was highly accurate and outperformed simpler baselines. The model achieved an accuracy of 89%, 81%, and 76% for ICD chapters, blocks, and full codes, respectively. These results have important implications for public health surveillance.

Esteban et al. [44] proposed a model for automatic determination of manner of death, using transfer learning for text classification on low-resource, domain-specific data. The dataset used in the study consisted of textual descriptions of injuries, demographic information, scene descriptions, and microscopy results found in autopsy reports. The results of the experiments showed that the proposed model significantly improved the classification performance compared to other algorithms such as AWD-LSTM (ASGD Weight-Dropped LSTM) and QRNN (Quasi-Recurrent Neural Network).

Yan et al. [45] introduced an automatic classification model based on character integration to improve the classification of causes of death from VA narratives. The model used distributed word vectors (Sentence2vec, GLoVE, and Word2vec) for feature extraction and combined these with character embedding methods (CNN and GRN) for the classification task. The results showed that the proposed model significantly improved the performance of determining causes of death.

In the literature, there have been numerous efforts to extract the cause of death (CoD) from autopsy reports using both traditional machine learning and cutting-edge deep learning techniques. The cited works all involve the use of machine learning and text classification methods to determine causes of death (CoD) from various sources such as forensic autopsy reports, death certificates, clinical bulletins, and VA narratives. Some of these studies use classical text vectorization and supervised machine learning algorithms for reports classification, while others employ transfer models and deep neural algorithms for both text vectorization and classification. All the studies showed that the use of machine learning and text classification methods can be useful for accurately determining the causes of death. Although most of these frameworks deliver good results, the best performance is contingent on the dataset used in the study. Thus, a comparative study is required to determine the most suitable approach.

In this study, we aim to evaluate the impact of using classical machine learning methods and state-of-the-art deep learning algorithms, such as the convolutional neural network (CNN) and the recurrent neural network, along with various NLP tools, to extract the manner of death (MoD) from autopsy reports. Our objective is to find an efficient and accurate model for the task.

4 Proposed framework

This section presents the framework proposed for predicting the CoD from autopsy reports. The complete procedure of this study is illustrated in Figure 12. The subsequent sections delve into the details of each step.

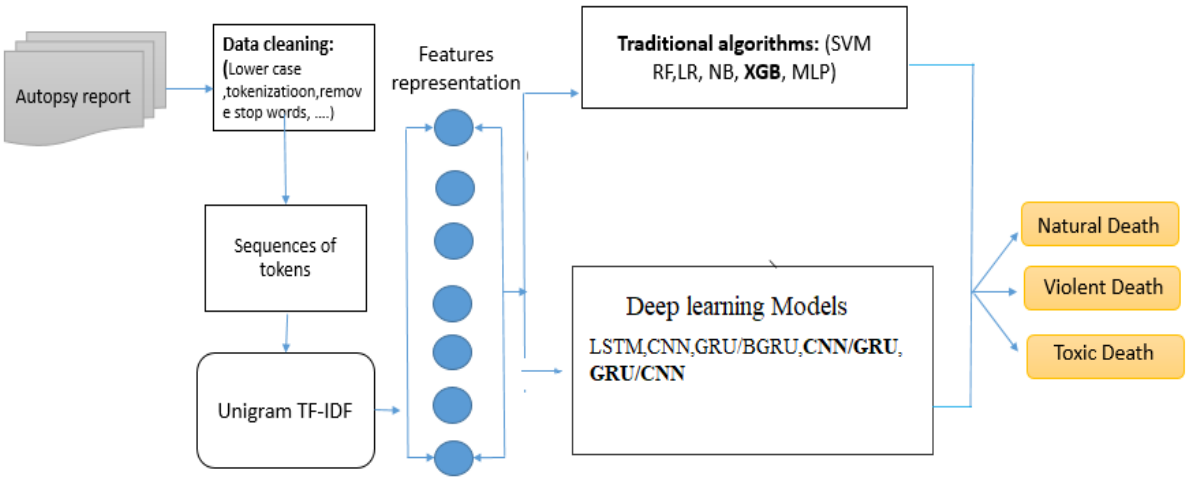

Figure 12: Workflow process

4.1 Data collection

This study analyzed 200 autopsy reports gathered from the University Hospital of Tlemcen in Algeria, within the forensic medicine department. The reports were classified into three Manner of Death (MoD) categories: natural death, violent death (penetrating, suicide, homicide), and toxic death. The personal information of the deceased, such as name, first name, and address, was removed from the reports, retaining only demographic information, details about external and internal examinations of the body (including the nervous, cardiovascular, respiratory, and digestive systems), and the conclusions about the cause of death.

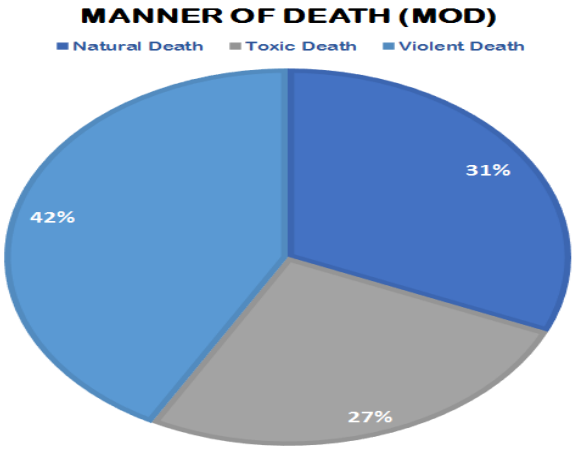

Figure 13: Distribution of manner of death.

The collected dataset is in French and consist of 63 reports for natural death, 53 for toxic death, and 84 for violent death, as shown in Figure 13. Each type of manner of death encompasses reports with multiple unique causes of death (CoD), resulting in a collected dataset with 12 different causes of death across the three categories. The distribution of these causes of death can be seen in Table 1.

Manner of Death (MoD)	Cause of Death (CoD)
Natural Death	Myocardial infarction
	Decompensated cardiomyopathy
	Cerebral palsy
	Pulmonary neoplasia
	Infectious myelopathy
	Rupture of a cerebral aneurysm
Violent Death	Head trauma
	A vital hanging
	Electrical burns
	Thermal burns
Toxic Death	Acute lung edema
	Asphyxia syndrome

Table 1: CoDs categorization

The Figure 14 represents the 5 most frequent causes of death in the dataset.

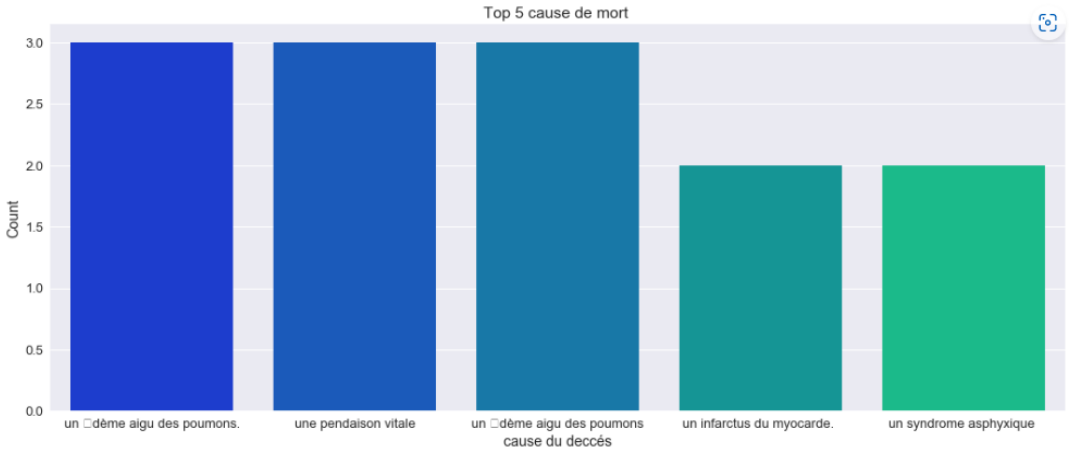

Figure 14: Top 5 causes of death.

4.2 Data preprocessing

In this step, the autopsy reports in Word format were first transformed into CVS format, labeled by the manner of death, for further processing. Then, a data cleaning step was carried out, including converting capital letters to lowercase, segmenting sentences

into tokens using the tokenizer function, removing empty French words using the Spacy library, and eliminating common sentences that may not be useful in the classification task. Finally, the dataset was divided into two parts using the simple random sampling (SRS) method with the help of the "train-test-split" function of the Sklearn Python library.

4.3 Features extraction

In this phase, the textual data from the autopsy reports had to be transformed into numerical vectors for further processing by machine learning and classification methods. To accomplish this, we applied the statistical vectorization method of TF-IDF (term frequency with inverse document frequency) to extract term-based features from the autopsy reports.

The majority of the works cited in the literature have utilized unigram TF-IDF for data vectorization, and the study by Mujtaba et al. [42] found this method to be effective for extracting relevant features for CoD detection. Therefore, we chose to use unigram TF-IDF in our study as well.

The result of the feature extraction phase was a main numerical feature vector where each row represented an autopsy report and each column represented the values generated by TF-IDF for each concept. This main feature vector was then used as input for the text classifier.

4.4 Classification

In this study, the autopsy reports were categorized into three labels: natural death, violent death, and toxic death. The TF-IDF feature representation was used to create the main feature vector for each report, which served as the input for text classifiers to predict the manner of death (MoD). A comparison was made between classical machine learning methods and deep learning algorithms in this classification phase. The traditional classifiers utilized in the study include: Support Vector Machine (SVM) [24], Random Forest (RF) [26], Logistic Regression (LR) [27], Naive Bayes (NB) [23], eXtreme Gradient Boosting (XGB) [46], and Multi-layer Perceptron (MLP) [47], for building predictive models.

Support vector machine (SVM) : is an application of complexity theory developed by Vapnik [24]. In [48], Joachims proposed SVM for text categorization. It is a supervised machine learning method that has been applied effectively in text classification. They are robust in high dimensional spaces; it is also strong when the data is sparse. In addition, SVM has attained good results in the opinion mining domain.

Random forest (RF) : is an ensemble learning method for classification tasks. This algorithm is a grouping of trees. It can be defined as a learning ensemble founded on bagging of unpruned decision trees apprentices with a randomized selection of features on each splits [26].

Logistic Regression (LR) : is a regression technique to predict a dichotomous reliant on the variable. In creating the LR equation, the maximum-likelihood ratio was utilized to find the statistical significance of the features [27].

Multinomial Naive Bayes (MNB) : is a variation of the Naive Bayes intended to resolve the text classification tasks. It uses multinomial distribution and is based on the use of the number of occurrences of a word or the weight of the word as a feature classification [23].

eXtreme Gradient Boosting (XGB) : is a supervised machine learning algorithm that works with all types of dataset [46], it is one of the implementations of gradient boosting, it produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. XGBoost is used for regression and classification problems, it uses a more normalized model formalization to control over-fitting which gives it better performance.

Multi-layer perception (MLP) : is a neural network trained to the standard back-propagation algorithm. It studies how to convert input data into the desired response, they are extensively used for pattern classification [47].

In this study, we utilized deep learning algorithms, including Convolutional Neural Network (CNN) [6] and three types of Recurrent Neural Networks (RNN): Long-Short-Term Memory (LSTM), Bidirectional Long-Short-Term Memory (BLSTM) [5], and Gated Recurrent Unit (GRU) [32]. To further improve the performance, we also explored the hybridization of these networks. Through various experimentation and evaluation, we were able to select the best deep learning architecture that provided optimal results.

Deep learning models : The autopsy reports are transformed into numerical representations using the TF-IDF feature extraction method, which converts words (tokens) into vectors. These resulting features are then fed into various deep learning algorithms to predict the manner of death (MoD) of each report. The principle behind each model is explained in detail in Chapter 1, Section 5. The next models we will build in this study are the Bidirectional GRU (B-GRU) and the hybrid B-GRU/Convolutional Neural Network (CNN) models.

BI-GRU Architecture : The Bi-GRU is a state-of-the-art Recurrent Neural Network, similar to an LSTM. Unlike the LSTM, the Bi-GRU only uses the hidden state to transfer information. This model considers two separate sequences, one processed from right to left and the other in the reverse order. The parameters used in the proposed model are detailed in Figure 15. The implementation of the model was done in Python using the Tensorflow library.

```
Model: "sequential_10"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_10 (Embedding)     (None, 50, 100)           1300000
_________________________________________________________________
spatial_dropout1d_2 (Spatial (None, 50, 100)           0
_________________________________________________________________
bidirectional_5 (Bidirection (None, 50, 500)           526500
_________________________________________________________________
bidirectional_6 (Bidirection (None, 64)                102336
_________________________________________________________________
dense_8 (Dense)              (None, 3)                 195
=================================================================
Total params: 1,929,031
Trainable params: 1,929,031
Non-trainable params: 0
_________________________________________________________________
```

Figure 15: BI-GRU model parameters

BI-GRU /CNN Hybrid architecture : The Convolutional Neural Network (CNN) is commonly used for image processing. However, text data has a temporal dimension, meaning it contains rich contextual information that needs to be maintained throughout processing, making it different from image data. While CNN can extract high-level features, it may not be able to effectively analyze sequences. Traditional Recurrent Neural Networks (RNNs) may face gradient disappearance or explosion when processing long sequences. The Gated Recurrent Unit (GRU) is a better solution, but Bidirectional GRU (B-GRU) is even more effective as it employs two RNN directions to better extract long-term dependencies. The B-GRU structure is simpler and requires less time to converge compared to traditional RNNs. By combining B-GRU with CNN, the ability to process long sequences is improved, resulting in a highly accurate classifier. The parameters used in the proposed hybrid model are shown in Figure 16.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_16 (Embedding)     (None, 50, 100)           1300000
_________________________________________________________________
dropout_10 (Dropout)         (None, 50, 100)           0
_________________________________________________________________
bidirectional_16 (Bidirectio (None, 50, 500)           526500
_________________________________________________________________
conv1d_10 (Conv1D)           (None, 50, 32)            48032
_________________________________________________________________
global_max_pooling1d_1 (Glob (None, 32)                0
_________________________________________________________________
dense_9 (Dense)              (None, 64)                2112
_________________________________________________________________
dropout_11 (Dropout)         (None, 64)                0
_________________________________________________________________
dense_10 (Dense)             (None, 3)                 195
=================================================================
Total params: 1,876,839
Trainable params: 1,876,839
Non-trainable params: 0
```

Figure 16: BI-GRU /CNN model parameters

5 Results and Discussion

For the experiments conducted in this study, we utilized Python along with the NLTK library for data preprocessing and the sklearn and keras libraries for feature extraction

and classification. The traditional classifiers were run using the default parameters with a multi-class classification approach.

For the deep learning methods, we utilized various architectures of CNN and RNN networks, with the parameters as specified in Table 2.

Table 2: parameters of deep learning algorithms

max-features	max-words	batch-size	num-classes	epochs	activation function
13000	50	128	3	40	softmax

To evaluate the performance of each model, we utilized the Accuracy evaluation metric with LIME (Local Interpretable Model-agnostic Explanations) [49] an explainability technique used to understand and interpret the predictions of a black box model, such as deep learning models. It is model-agnostic, meaning it can be applied to any machine learning model, and its aim is to provide a human-readable explanation for individual predictions. LIME does this by creating a simplified, local explanation of the model's predictions for a particular instance. This allows understanding of why the model made the predictions it did, and can also help identify potential biases in the model's behavior.

The accuracy formula is defined as:

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

The performance results of traditional and deep learning algorithms are presented in Tables 9 and 8 respectively.

Table 3: Performance results of traditional methods

Methods	Representation	Accuracy
XGBClassifier	TF-IDF	0.9545
MLPClassifier	TF-IDF	0.8181
RF Classifier	TF-IDF	0.9090
MultinomialNB	TF-IDF	0.6363
LogisticRegression	TF-IDF	0.6818
SVM classifier	TF-IDF	0.7727

From Table 9, it is evident that the combination of the TF-IDF feature representation and the XGB classifier produces the best results, with an accuracy of 95.45%. The RF classifier also performs well compared to the other traditional algorithms, however, the NB and LR classifiers produce subpar results.

Table 4: Performance results of deep learning algorithms

Methods	Representation	Accuracy
GRU	TF-IDF	89.39
Bi-GRU	TF-IDF	93.94
LSTM	TF-IDF	86.36
CNN	TF-IDF	89.393
CNN/B-GRU	TF-IDF	85.84
B-GRU/CNN	TF-IDF	90.91

The results in Table 8 indicate that the highest accuracy was achieved by the B-GRU classifier with 90.90% and the GRU classifier with 89.39%. The CNN classifier also had a good accuracy of 89.39%. However, the hybridation of CNN and GRU showed the lowest accuracy of 84.84%.

Comparing the results of text vectorization using TF-IDF and the different classification methods, it can be seen that the best performance was obtained by using TF-IDF with the XGB classifier, with an accuracy of 95.45%. On the other hand, the results of using TF-IDF with deep learning classifiers were not as good.

These findings suggest that the XGB classifier is a strong choice among supervised machine learning algorithms for autopsy report classification. Additionally, deep learning methods perform better when combined with embedding methods, rather than features generated by classical methods. This is because the TF-IDF technique is considered a powerful feature engineering method when used with shallow, lightweight text processing models like XGB classifiers, whereas deep learning tools incorporate structural feature learning through methods like word embeddings, avoiding the rigid and fragile approach of feature engineering.

5.1 Explainability analysis

The use of machine learning models in the medical field can be challenging due to the difficulty in interpreting their predictions, especially when dealing with high-dimensional medical concepts. This can make it challenging for medical practitioners to understand the underlying mechanics of the model and the potential pitfalls associated with it. In order to find the right balance between a powerful model and an interpretable model, LIME (Local Interpretable Model-agnostic Explanations) was used in this study.

LIME is a model-agnostic Explainability tool that can be applied to any machine learning model. By manipulating the input data samples and observing the changes in predictions, LIME aims to provide insights into how the model works. The technique requires an interpretable representation of the input data that is understandable to humans. In this study, the TF-IDF vector was used to explain the predictions made by the machine learning classifier, XGB, which achieved the best results.

LIME generates a list of explanations that reflect the contribution of each concept word to the prediction of a data sample. This allows to determine which terms have the most impact on the XGB prediction. An example of this is shown in Figures 17 and 18, which present the probabilities of features towards two labels for a randomly selected report in the test set (Document with idx=9). This document was labeled as "Natural" and predicted as "Natural" by the model. The figures show the contribution of different features towards the prediction of the two labels, "Natural" and "Violent.

```
Explanation for class Naturel
('naturelle', 0.48497677734199984)
('mort', 0.06595123352383259)
('colon', 0.06233220773170181)
('cardiaque', -0.05683388194734076)
('intestins', 0.04121500914923464)
('gauche', -0.03794097594389867)
```

Figure 17: Relevant concepts probabilities for class Natural.

```
Explanation for class Violent
('naturelle', -0.1805633179466105)
('colon', -0.11430821348738314)
('gauche', 0.05170605121545568)
('intestins', -0.049048349840441145)
('cartilage', 0.03979802170370399)
('cardiaque', -0.025420202384178997)
```

Figure 18: Relevant concepts probabilities for class Violent.

From Figures 17 and 18, it is evident that the document in question has the highest explanation for the "Natural" label. The positive and negative signs in the explanation refer to the contribution of the features to a particular label. For example, the word "naturelle" is positively contributing to the "Natural" class, but negatively contributing to the "Violent" class. Hence, by generating explanations for the top two classes, "Natural" and "Toxic" are obtained as seen in Figure 19.

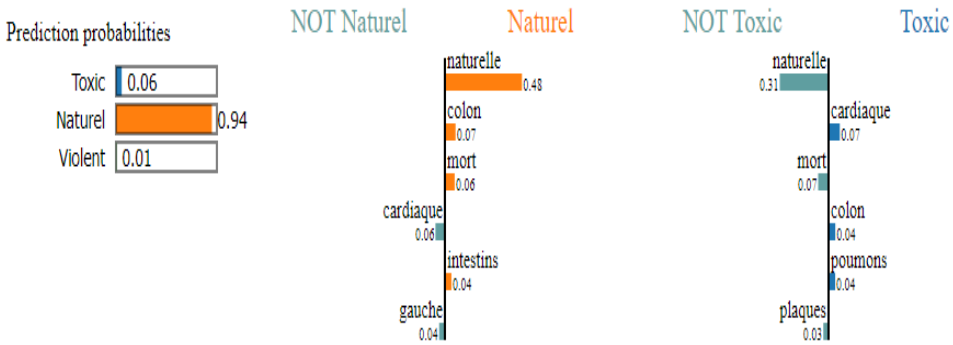

Figure 19: Top Two labels for the report

As shown in Figure 19, the word "naturelle" has the highest positive score for class Natural in this report. The XGB model predicts the label of this document as Natural with a probability of 94%. If the word "naturelle" were removed from the report, the probability of the XGB model predicting label Natural would decrease from 94% to 46% (94% - 48%). On the other hand, the word "naturelle" has a negative score for class Toxic and the words "poumon" and "cardiaque" have a small positive score for the same class. These words are related to acute lung edema, which is a significant cause of toxic death.

6 Conclusion

Autopsy report classification using NLP and machine learning techniques plays a crucial role in the medical field, as autopsy reports contain vital information about the cause of death (CoD) that can aid doctors in improving their understanding of this area. Our study consisted of two main parts: data collection and algorithm comparison. The data collection phase involved collecting 200 autopsy reports from the CHU Tlemcen Forensic Medicine department, which contained information about external and internal examinations of the deceased, personal information, and corresponding labels for the manner of death (natural, violent, or toxic).

The second part of this study compared traditional machine learning and deep learning algorithms for autopsy report classification, using the TF-IDF method for feature extraction and analyzing the performance of each algorithm when combined with this tool. Our experiment results showed that the XGB classifier combined with TF-IDF performed best in terms of accuracy, while deep learning methods performed poorly when combined with TF-IDF. This can be attributed to the fact that TF-IDF is a form of feature engineering that is most effective when used with shallow, lightweight text-processing models such as traditional classifiers, whereas deep learning methods prefer structural feature learning through word embedding methods.

To interpret the output of our approach, we used the evaluation metric LIME, which is a great tool for making complex models interpretable. Our conclusion from this study encourages us to continue our work by collecting more autopsy reports and using word embedding methods for feature extraction in the future, with the aim of improving our results.

CHAPTER 3

Sentiment analysis from social networks for medical decision support

1 Introduction

Since 2003, the evolution of the Web has been vast, due to the growing use of the Internet [50]. Social media networks, blogs, and emotional websites generate a vast amount of unstructured textual data in the form of emotions and opinions on various topics across various domains, such as products, education, and health communities. These emotions have a significant impact on end-users and are important for experts to understand the strengths and weaknesses of products and to improve the quality of production. As a result, public opinion will become increasingly important. In this context, the ability to extract, capture, and analyze the feelings of the general public from large amounts of unstructured textual data is of growing interest to data mining researchers. This task requires an automated process known as sentiment analysis or opinion mining [6].

In recent years, Opinion Mining (OM) based on machine learning (ML) and deep learning algorithms has become a powerful computational technique for solving complex sentiment analysis tasks from different perspectives [51]. ML tools have been used to process large amounts of free-text comments due to the inherent capabilities and hierarchy of machine learning and deep learning models.

The opinion mining process is performed at various granularities such as sentence, document, and aspect-based levels. The automated process is based on basic building blocks that include various traits considered to represent free text for sentiment analysis and classification. In subsequent sections, we will discuss different levels of sentiments and each trait, providing an overview of the basics and standard methods used in building OM models.

2 Sentiment Analysis in Medical health

Information search and retrieval involves processing information to answer a question or solve a problem. In the medical domain, this activity is carried out with the aim of finding practical information, treatment information, disease evolution information, or connecting with other patients to gain new medical knowledge [52].

The health domain is a field that showcases the evolution of the information retrieval process. With the growth of the internet, medical forums have seen a surge in health communities, used by millions of users, many of whom are patients sharing their medical problems and experiences with others facing similar situations.

Sentiment analysis has been applied in various medical domains such as clinical record evaluation, drug control, and more recently, the COVID-19 pandemic. Researchers have been interested in analyzing people's feelings towards the pandemic to track its evolution and its impact on the general population. A study by the Pew Internet and American Life Project [1] showed that nearly 80% of internet user forums in the United States discuss medical topics, with 63% focusing on exploring information and knowledge about a specific medical problem and 47% of internet users seeking effective medical treatment or protocol by asking specific questions and seeking feedback from other patients [53].

It is important to understand the criteria used by patients to validate the medical information found in forums and to extract these valuable medical emotions. This has led health and data mining researchers to be interested in these emotions in health forums and to consider them as a criterion for evaluating search results on medical devices.

Medical sentiment analysis offers a unique perspective as it can highlight diagnostic support systems and propose a new approach to improving the quality of medical devices. In recent years, medical opinion analysis has become a vital topic in medical research, focusing on extracting medical emotions from a vast amount of clinical narratives available on the web and medical social media sources [54].

There are different facets of medical opinions in medical forums that can be linked to various entities and events in the medical domain (as shown in Figure 20). Patients' sentiments can pertain to the following topics:

- **Health status:** This refers to patients' opinions about how their health condition has changed over a given period and the extent to which it affects their life. For example, a change in medical condition could have a direct impact on the severity of the disease and the patient's circumstances. In this case, a patient's primary concern can carry significant weight in determining the health status.
- **Treatment effects:** This aspect of patients' opinions in the medical context refers to their views on a specific treatment as expressed in social media forums.
- **Sentiment towards a drug:** This refers to patients' sentiments towards a particular drug, such as their positive or negative reactions to its consumption. For example, the outcome of medication could be positive, negative, or neutral.

[1] http://www.pewinternet.org/

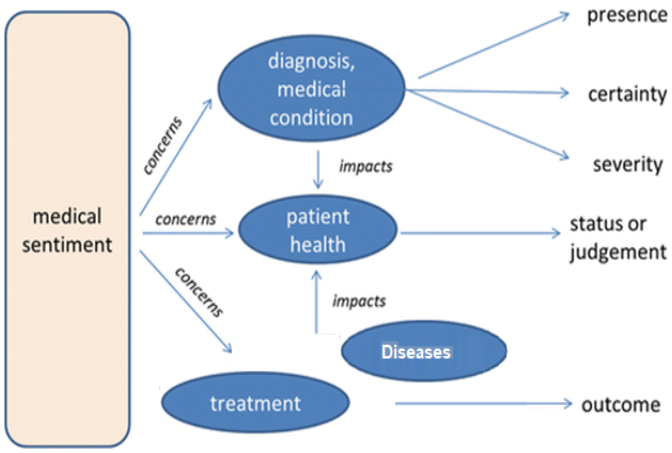

Figure 20: Facts of sentiment in health-care domain.

- **Certainty of a diagnosis:** When health professionals are uncertain about a patient's diagnosis, they seek opinions from other patients and health professionals. This emotion about the certainty of a diagnosis greatly impacts the treatment protocol decision. For example, if the diagnosis is uncertain, a final treatment decision can be made; if not, further medical tests may be required.
- **Opinion towards disease:** For instance, the impact of COVID-19 on a patient's life. Analyzing public opinion on COVID-19 can help epidemiological experts understand people's reactions and monitor the spread of the virus.

Thus, the concept of medical opinion is complex, making it interesting for automatic analysis. To achieve this, biomedical researchers have created domain-specific corpora containing clinical texts, such as drug reviews and comments from medical blogs like WebMD and DrugRating. They use machine learning and deep learning techniques to build automatic sentiment extraction processes from these corpora, contributing to the development of an effective healthcare system.

3 Opinion mining categorization and levels

The aim of Opinion Mining (OM) is to detect, identify, and classify sentiments expressed in user-generated reviews on social media or product reviews as positive, negative, or neutral. There has been some debate in the literature about the difference between sentiment and opinion, leading to confusion about whether the field should be referred to as Sentiment Analysis or Opinion Mining.

Merriam-Webster's Collegiate Dictionary defines a feeling or sentiment as an attitude, thought, or judgment fostered by a sensation, and opinion as a view, judgment, or evaluation formed in the mind about a particular subject [55]. The distinction between the two is subtle and each contains elements of the other.

3.1 Sentiment categorization

In sentiment analysis, sentences written in natural language can be classified as objective or subjective. When a sentence is objective, no further classification is required. When

a sentence is subjective, its polarity (positive, negative, or neutral) must be estimated through a process called Polarity Classification (as shown in Figure 21). Fine-grained sentiment analysis may further categorize polarity into very positive, positive, neutral, negative, and very negative, with each category mapped to an evaluation score such as 5 stars for "very positive" and 1 star for "very negative". The polarity scores assigned to individual sentences are then aggregated to give a global score for multiple sentences.

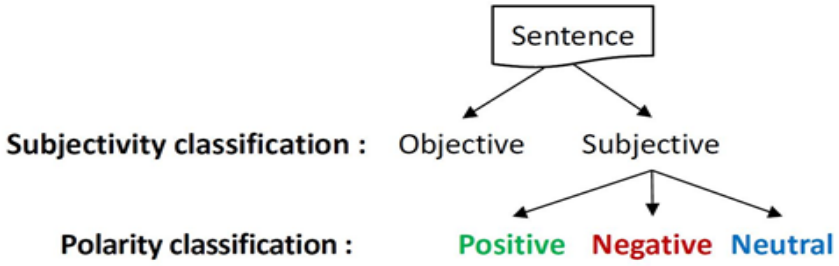

Figure 21: Workflow of sentiment analysis process.

3.2 Sentiment levels

The application of sentiment analysis begins with defining the text that will be analyzed in a study. Sentiment mining can be performed at three levels of granularity [56]: document level, sentence level, and aspect-based level.

Document level: This level characterizes the polarity of a complete text, such as a document or paragraph, assuming that the text expresses a single opinion about a single entity. The model is usually modeled as a binary classification problem, for example, to determine if a product review is positive or negative, or as a regression problem, for example, to assign a rating score from 1-5 stars to a story review.

Sentence level: This level determines the polarity of each sentence in a text, assuming that each sentence expresses a single opinion about a single entity. The sentiment analysis process first identifies whether a sentence is subjective or objective, then, in case of a subjective sentence, uses polarity classification to determine whether it expresses a positive, negative, or neutral emotion.

Aspect-based sentiment analysis: This level performs a more detailed and refined analysis by examining the sentiment for each aspect of a text. The process involves identifying the aspect terms in the text, determining their polarity, and detecting their categories. For example, a sentence such as "The iPhone is very good, but still needs work on battery life and safety issues," would evaluate three aspects - iPhone (positive), battery life (negative), and security (negative).

3.3 Sentiment analysis Approaches

In the literature, there are several methods and algorithms for implementing sentiment analysis systems, which can be broadly categorized into three groups:

- **Automatic approach:** The automatic approach is based on machine learning techniques. In this approach, the task of sentiment analysis is typically modeled as a classification problem, in which a machine learning classifier is fed with text and returns the corresponding sentiment category, such as positive, negative, or neutral in the case of polarity analysis (as shown in Figure 22).

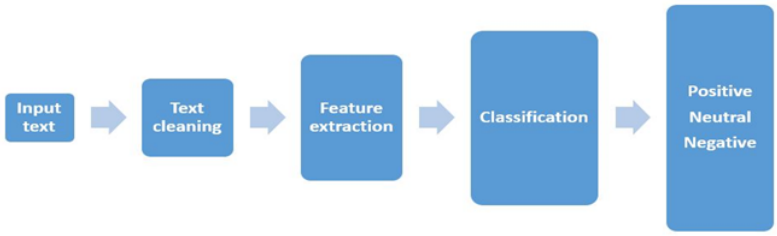

Figure 22: Workflow of automatic approach.

- **Rule-based approach:** The rule-based approach, also known as the lexicon approach, uses a set of rules defined in a programming language (script) to identify the subjectivity, polarity, or subject of an opinion. This approach can utilize various inputs, including classical NLP techniques like tokenization, POS-tagging, and chunking, or a sentiment dictionary containing opinion words to match with the data to determine its polarity. See Figure 23.

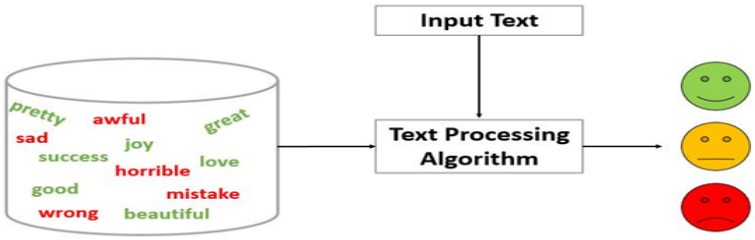

Figure 23: Workflow of Rule-based approach.

- **Hybrid approach :** combines the strengths of both rule-based and automatic approaches to achieve improved accuracy in sentiment analysis. By leveraging the rule-based approach's ability to capture patterns and relationships and the automatic approach's ability to learn from data, hybrid methods offer a promising solution for sentiment analysis.

4 Aim of study

The sentiment in medical texts reflects a patient's health status, as captured by observations and objective information about their medical conditions and treatments. The extraction of this sentiment from medical texts has numerous applications in the medical field and can provide valuable information for clinical decision-making. To address the diversity and importance of medical sentiment analysis, a comprehensive approach is required.

Text mining and machine learning techniques have been instrumental in developing robust models for sentiment analysis in medical texts. Research in this field has covered medical social media and biomedical literature, utilizing deep learning methods. In our study, we focus on sentiment analysis in medical texts using NLP and machine learning approaches. Our goal is to demonstrate the effectiveness of these methods in sentiment analysis.

To that end, we have two main objectives :

- First, we aim to highlight the value of NLP and machine learning in processing opinions and feedback about drugs from medical forums. This can improve the quality of drugs and provide insight into their effects and effectiveness, which can benefit both patients and doctors.
- Second, we analyze public sentiment towards the COVID-19 pandemic to understand public reactions and aid epidemiologists in monitoring the spread of the virus. Our work proposes models for sentiment analysis on COVID-19-related tweets using traditional and advanced deep learning techniques and introduces a new hybrid model based on CNN and RNN for sentiment analysis.See Figure 24.

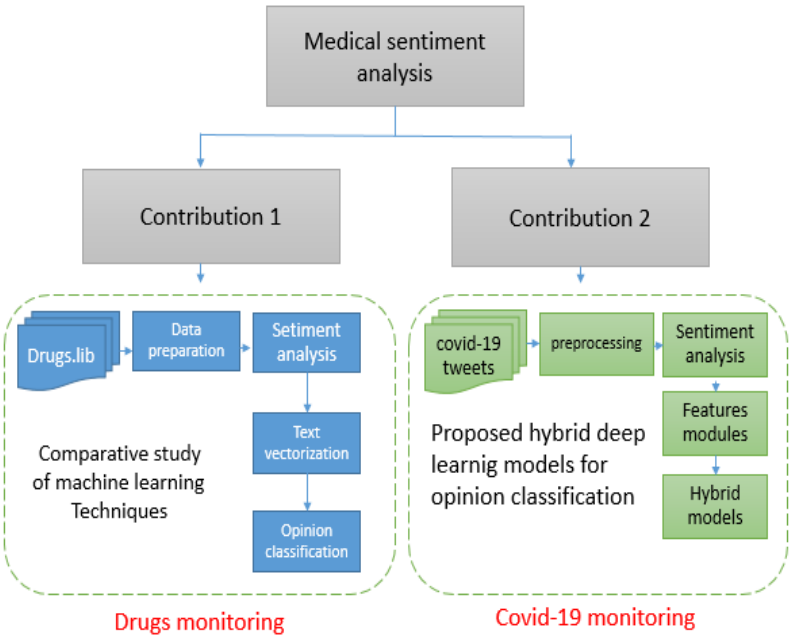

Figure 24: schema desriptive of the two contributions

5 Drug monitoring: Analysis of machine learning and deep learning frameworks for opinion mining on drug reviews

Abstract

The extraction of emotions from medical social media, such as medical forums discussing patients' treatments, is a significant challenge. Opinion mining (OM) of these sources provides insight into the behaviors of both doctors and patients, attracting widespread attention globally. The goal is to improve healthcare services and drug production efficiency through a better understanding of patients' health status and treatment feedback. This study seeks to develop an effective model for analyzing patients' drug-related emotions through the use of machine learning and deep learning methods. Traditional feature extraction techniques (BOW, TF-IDF) and word integration methods (Word2vec, GloVe) are applied alongside advanced machine learning methods such as convolutional neural networks (CNN), long-term memory (LSTM) and bidirectional long-term memory (BLSTM) recurrent neural networks. The best model was found to be a CNN combined with the Skip-gram model. The results show that the performance of a model is dependent on not only the algorithm efficiency but also the dataset type, feature extraction method, and collaboration between the classifiers and feature extraction methods. This section provides an in-depth presentation of this work, including a medical background, a review of related studies, a description of the methods used, experimental results, and a discussion of these results. The results of this study were published in the Oxford - Computer Journal and an extended results was presented at the International Conference on Intelligent Systems and Pattern Recognition (ISPR '20) in October 16 - 18, 2020, Tunisia, grabbing the Best Paper Award.

5.1 Context of the study

Communication through social networks has become a huge part of daily life, where users express their honest opinions on various topics from different domains. These emotions are crucial in decision-making, and researchers in text-mining and data-mining have taken a great interest in utilizing this source of subjective information by using NLP tools and machine learning methods [57]. However, not much work has been done in the field of health to explore opinions on medicine and healthcare [58].

On the other hand, online review sites have a significant impact on patient communication. Patients share their experiences and interact with other patients in social networks, influencing their beliefs and decisions. It is important for health professionals, such as doctors, pharmacists, and medical equipment manufacturers, and patients to know patient opinions and stories through their reviews to gain new insights to make better decisions.

The opinions of patients towards adverse drug reactions are a critical public health issue. The goal of pharmacovigilance is to analyze these opinions, explain them, and implement measures to prevent adverse drug reactions. The pharmaceutical industry must also take these opinions into account to improve the quality of medicines.

With the advent of new technology, the doctor-patient relationship has changed, giving patients the ability to report adverse reactions directly on health authority websites. Patients look for information online and share their experiences, especially about pre-

scribed medicines and adverse reactions. Some patients may not be able to communicate easily with their doctors about the effects of these medicines, but they can express their emotions on social networks by interacting with other patients. This creates an important source of data for pharmacovigilance that has yet to be exploited.

Therefore, data mining researchers are focusing their work on extracting emotions from drug reviews using NLP methods and machine learning tools to develop an automated opinion mining model that can help doctors and patients minimize the effects of drugs. However, it is challenging to analyze complex textual data and find an efficient model that works for all types of data.

Deep learning techniques, such as word embedding tools and convolutional neural networks for classification, have made it easier to build computational and efficient models that maintain the reliability and accuracy of opinion data without the need for manual feature selection. Advanced techniques such as long-term short-term memory (LSTM) and bidirectional long-term short-term memory (BLSTM) are also effective for sentiment analysis by considering opinion extraction as a sequence problem.

In the fields of NLP and data mining, it is challenging to find an effective emotion extraction model from unstructured databases. The best framework for sentiment analysis of drug reviews depends on not only the capability of the chosen techniques, but also the integration of all NLP and ML methods in the process, and the type, size, and preprocessing of the data.

In this work, we propose a comparative study between traditional and deep learning techniques for opinion extraction in drug reviews. The study includes pre-processing the textual data to keep only relevant concepts, feature extraction using classical tools and embedding methods, and classifying opinions using various algorithms. The results and analysis of the experiment are reported, followed by the discussion and conclusions.

5.2 Related Works

Initial research in sentiment-based classification was partially knowledge-based, with some works focusing on classifying the semantic orientation of individual words using linguistic heuristics [59–61]. Subsequently, several methods were introduced in the literature to analyze opinions from online reviews (e.g., blogs, forums, commercial websites) based on semantic orientation (SO) and machine learning (ML) approaches [62].

Sentiment analysis using machine learning algorithms has received considerable research attention, as it enables fast processing of large amounts of unstructured comment data [51]. The majority of this work focuses on classical supervised machine learning algorithms for extracting and classifying opinions from reviews [63–65]. For example, in the work of [66], the researchers used the unigram text vectorization method with the TF-IDF count vectorization technique for feature extraction from a Twitter dataset and then applied Support Vector Machines (SVM), Naive Bayes (NB), and K-nearest neighbors (K-nn) algorithms in the classification phase.

Deep learning (DL) has also shown effectiveness in opinion mining, with its architecture-based algorithms allowing for fast processing and improved performance on large amounts of free text data [6, 57]. This approach has been widely studied in the context of social networking and web text data, with most studies relying on deep learning tools to construct features from text using word embedding algorithms such as Word2vec, and then classify them using convolutional neural networks or recurrent neural networks [67].

Irsoy and Cardie [68] applied deep recurrent neural networks (DRNNs) to analyze emotional expressions in critics. The results of the experiments showed that DRNNs performed better than conditional random fields (CRFs), which were constructed by stacking Elman-like RNNs. Each layer utilized the memory sequence of the previous layer as an input and calculated its memory representation.

Soujanya et al. [69] employed a 7-layer deep convolutional neural network (DCN) to extract specific aspects of a product from the emotional text and classify each aspect of the opinion sentence. The proposed method yielded better results in terms of accuracy compared to other techniques used.

In the field of medical health, sentiment analysis has been widely studied to analyze emotions related to various entities and events. The aim of medical sentiment analysis is to extract patients' emotions from medical documents [58]. Most of the studies in this field apply deep learning algorithms [70] for emotion analysis and make use of word embedding tools such as word2vec and glove [71] [4] to extract and represent the feature vector of medical concepts. Drug monitoring is a crucial area where sentiment analysis of drug reviews can provide valuable insights into the side effects of medical treatments for both health professionals and patients. However, there is limited work in this field using both traditional machine learning methods and deep learning techniques [54].

The researchers Uysal et al. [72] proposed a comparative study of six textual feature selection techniques for extracting features from drug reviews. They then used two classical classifiers, Support Vector Machines (SVM) and Naïve Bayes (NB), in the classification step. The results demonstrate the effectiveness of feature selection tools in the opinion mining process, with the Improved Complete Measurement Feature Selection (ICMFS) technique performing better than the other methods.

Another study by Tianhua et al. [73] focuses on the application of fuzzy rough feature selection for automatic sentiment analysis of drug reviews to reduce noisy data in the process. The researchers apply the TF-IDF method for feature extraction and use four popular supervised algorithms for sentiment classification.

Shweta et al. [74] developed a deep Convolutional Neural Network (CNN) to analyze patients' opinions regarding their health status, medical condition, medication, and treatment. They collected these opinions by scraping the medical forum website 'patient.info.' In another study, Leaman et al. [75] investigated a large dataset from online health-related websites to uncover correlations between medical treatments and their side effects.

Gopalakrishnan et al. [76] used a neural network approach to examine emotions and assess the performance of classification algorithms on reviews of two different medications. The results of the experiment indicate that the neural network-based approach outperforms the statistical approach in terms of precision, recall, and F-score.

Liu et al. [77] proposed a Convolutional Neural Network (CNN) architecture for extracting Drug-Drug Interactions (DDIs). The model first takes a DDI instance generated by word embedding as input, then feeds it into a convolutional layer for feature extrac-

tion using filters of different sizes. The pooling layer then generates feature vectors, and finally, a fully connected SoftMax layer classifies the DDI type.

Sisi and Ickjai [78] conducted a comparative study of sentiment analysis on drug review datasets using medical word integration and sequence representation techniques. They evaluated and compared the results of sentiment classification using word embedding tools such as Word2vec and GloVe with a CNN model, as well as several sentiment lexicon-based vector representation methods with various machine learning algorithms. The experiments showed the effectiveness of word embedding methods in sentiment classification for drug reviews.

Liu et al. [77] proposed a CNN architecture for extracting drug interactions (DDIs) by utilizing word embedding. The input is first transformed into a DDI instance, then fed into a convolutional layer for feature extraction, followed by a pooling layer to generate feature vectors. Finally, a fully connected SoftMax layer is used to classify the DDI type.

In another study, Sisi and Ickjai [78] conducted a comparative analysis of sentiment analysis techniques for drug review datasets, using medical word integration and sequence representation techniques. The researchers evaluated and compared the performance of the Word2vec and GloVe word embedding tools with a CNN model and several sentiment lexicon-based vector representation methods with different machine learning algorithms. The results indicated that word embedding methods were effective for sentiment classification in drug reviews.

Ru et al. [79] proposed deep neural network models including CNN, LSTM, and CLSTM (Convolutional Long Short-term Memory network) for extracting drug serendipity usage from social media data. The challenge was to use contextual information obtained from patients' comments on medicines, medical anthologies, and medical knowledge. The results of the experiment with the different models showed that the combination of deep neural networks and word integration techniques is an advantageous area for further research.

Min et al. [80] proposed a combination of a CNN model and a bi-directional long-term memory (Bi-LSTM) for classifying adverse drug reactions (ADRs). The proposed model outperformed other deep learning (DL) architectures in terms of accuracy. Cocos et al. [81] considered text as a sequence of words and labeled the input words with ADR membership using a recurrent neural network (RNN) model. Three different architectures, BiLSTM-M1, BiLSTM-M2, and BiLSTM-M3, were used, each incorporating pre-processed embedded words from a large, non-domain-specific Twitter dataset. The proposed RNN model demonstrated its ability to address the limitations of CNN in sequential modeling of text across sentences.

Kadam et al. [82] developed a hybrid RNN stacked with a bi-directional LSTM model for a drug recommendation system that recommends drugs based on reviews from users.

In the field of NLP and machine learning, various studies have been conducted for extracting opinions from drug reviews. Most of these studies focus on adverse drug reaction reviews and employ traditional text vectorization and supervised machine learning techniques for opinion classification. Others utilize convolutional neural networks (CNNs) for both text vectorization and opinion classification.

More recent research is exploring the use of recurrent neural networks for this task, and

many studies have reported promising results. This study aims to investigate the impact of applying classical machine learning methods and popular deep learning algorithms, such as CNNs and recurrent neural networks, along with various NLP tools, on opinion extraction from drug reviews. The goal is to find an efficient and accurate framework for opinion mining from the social web in the field of drug monitoring. In the following section, more details on the proposed methods will be presented, followed by a discussion of the results obtained using four evaluation measures.

5.2.1 Proposed Approach

This study aims to analyze and compare the most efficient models of NLP and machine learning tools for sentiment analysis in drug reviews. In this framework, we propose a comparative study between the most efficient NLP and machine learning tools for sentiment analysis in drug reviews. The proposed process is divided into four steps as shown in Figure25.

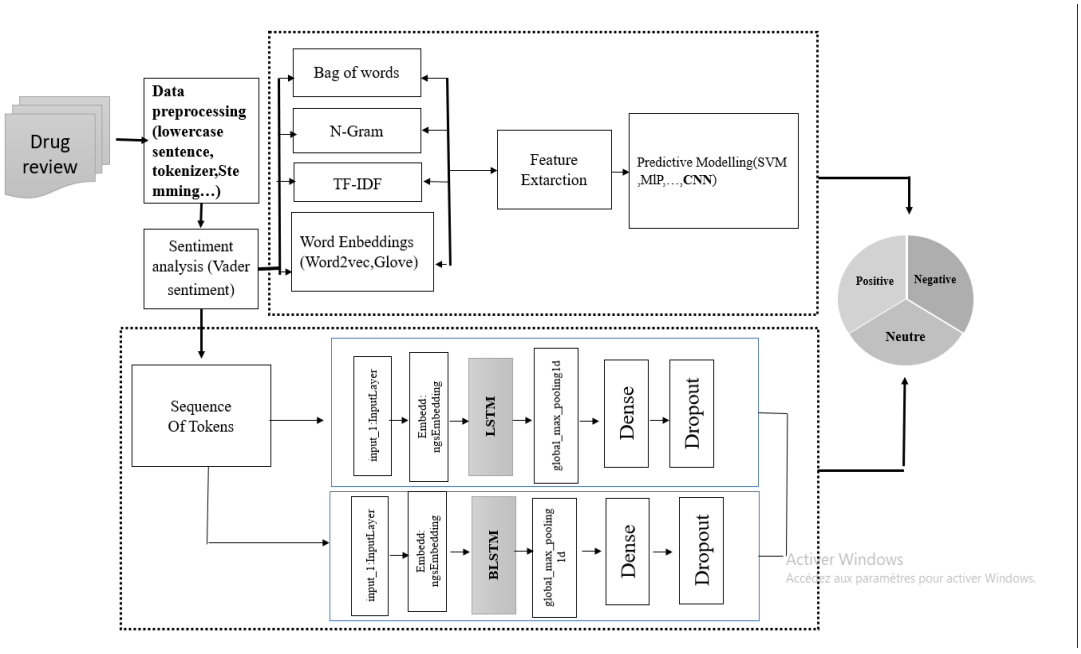

Figure 25: Diagram representing the overall operating process for opinion mining in drugsreviews

- Data Exploration and Preprocessing
- Unsupervised Sentiment Analysis using the Vader Algorithm
- Text Vectorization using Classical and Embedded Feature Extraction Methods
- Comparative Study between Machine Learning and Deep Learning Networks for Sentiment Analysis

5.3 Data Pre-processing

The data preprocessing step is crucial for the overall outcome of the study. It ensures that the data used in the subsequent steps is clean and ready for analysis. The chosen

preprocessing methods depend on the dataset used. In this study, the Drug.lib dataset [83] is used, which contains redundant words that could lead to incorrect analysis. Therefore, these words must be removed. The data also contains words that are contextually similar, such as "improve," "improvement," and "improved" or "take," "taking," and "taken." To address this issue, each review in the data is first converted to lower-case, and then tokenization [84], stop-word removal, and stemming are carried out. These steps are detailed in the first chapter of the study.

5.4 Sentiment Analysis

The main objective of this phase is to comprehend patients' emotions regarding their medication from social media by using rule-based and lexical methods to generate compound feelings' polarity scores from patient reviews. These opinion polarity scores are then grouped into three categories (positive, negative, and neutral). In this study, two sentiment analysis libraries, TextBlob [85] and Vader Sentiment [86], are used to analyze opinions and extract sentiments. The polarity labels generated by the Vader algorithm are chosen for the next steps, as it can analyze more factors such as exclamation marks and emotions, and these punctuation marks are included in the computation of polarity scores.

5.5 Feature Modules

In this study, six text vectorization techniques were utilized: Bag of Words, N-grams, Term Frequency-Inverse Document Frequency (TF-IDF), Word2vec, and GloVe, which were detailed in Chapter 1. These methods were applied to generate vector representations from the dataset.

5.6 Predictive Modeling

5.6.1 Deep learning models

CNN: The architecture of the CNN used in this work is depicted in Figure 27. The input reviews are transformed into small-scale representations known as word embeddings or terms through a feature extraction process. These word features are then fed into the convolution layer. The convolution layer computes the weighted sum of two words at a time as a filter slides over a sentence and generates a per-element product. The results from the convolution layer are then pooled to a representative number and passed into a fully-connected layer for classification.

In the convolution step, the comments are arranged in a matrix, with each row representing a word or word embedding. This layer scans the reviews like an image, decomposes the image into elements, and judges which element fits or does not fit the label. We used three Conv1D convolutional layers in this work, with 512 filters for the first two layers and 256 filters for the last one, as shown in Figure 2.

In the pooling step, the sum of the products generated by the convolution layer is the actual textual feature. This layer reduces the dimensionality of the word features and only retains a simple probability score that reflects the likelihood of the label. We also

used a dropout rate of 0.2 to reduce overfitting and computational cost.

In the fully-connected layer, the scores are used as inputs, and a back-propagation process is applied to determine the most accurate weights. Each neuron in the layer receives weights that prioritize the most appropriate class (positive, negative, or neutral sentiment). The neurons then "vote" on each of the labels, and the winning vote becomes the final classification decision. We used 256 and 512 neurons with ReLU activation in the fully-connected layer to add non-linearity to the network.

(LSTM) and (BLSTM): In this work, we utilized Recurrent Neural Networks (RNNs) to identify concepts related to positive, negative, and neutral emotions in long user reviews, as the semantic meaning of a term can be influenced by the words before and after it. We employed two types of RNNs: Long Short-Term Memory (LSTM) [87] and Bidirectional LSTM.

LSTM is a promising method in natural language processing and sequential prediction tasks due to its ability to handle arbitrary spatiotemporal dimensions. It reads the entire sentence from beginning to end, incrementally updating its understanding of the sentiment after each step by considering both its memory and partial output.

For our study, we utilized the LSTM architecture to analyze the sentiment of a single review by treating it as a sequence of 7484 unique tokens learned during the convolution operation. We employed One-Directional LSTM and Bidirectional LSTM networks, along with a SpatialDropout1D, two dropout layers, and three fully connected layers with a softmax activation function.

The first layer is the Embedding layer, which converts tokens into embeddings with a size of 200. The LSTM layer consists of three LSTM layers with 256 hidden units each. The Dropout layer helps prevent overfitting by randomly and periodically removing some of the network's neurons and their connections. Finally, three fully connected layers are added to map the LSTM output to the desired output size of three, provided by the softmax activation function.

5.6.2 Machine learning Models

For comparison with the deep learning models, we used six supervised machine learning algorithms selected from the literature to build predictive models. All the features were generated through text vectorization methods. The selected classification algorithms include Support Vector Machine (SVM) [24], Random Forest (RF) [26], Logistic Regression (LR) [27], Naive Bayes (NB) [23], eXtreme Gradient Boosting (XGB)

5.6.3 Data description

The study utilized a dataset from the well-known UCI machine learning repository [83]. The dataset was the drug review dataset, which comprised of 4132 patient reviews on specific drugs along with their associated conditions and a 10-star rating system indicating overall patient satisfaction. The data was collected by crawling online pharmaceutical review sites. The sentiment polarity was determined using the Vader sentiment function, resulting in three labels: positive, neutral, and negative.

	Bigram TF-IDF		TF-IDF		BOW	
	Accuracy	**F-score**	**Accuracy**	**F-score**	**Accuracy**	**F-score**
SVM	0.8070	0.81	0.8288	0.83	**0.8531**	**0.85**
RF	0.5339	0.45	0.5728	0.51	0.5691	0.51
LR	0.7439	0.74	0.7682	0.77	0.8240	0.82
MNB	0.6893	0.69	0.7075	0.71	0.6771	0.68
XGB	0.8216	0.82	0.8313	0.83	0.8398	0.84
MLP	0.7572	0.76	0.7936	0.79	0.8410	0.84
CNN	0.7607	0.76	0.7794	0.78	0.7383	0.74

Table 5: Performance results using classical methods

	CBOW		GloVe		Skip gram	
	Accuracy	**F-score**	**Accuracy**	**F-score**	**Accuracy**	**F-score**
SVM	0.8218	0.74	0.4280	0.39	0.8209	0.74
RF	0.8218	0.74	0.5163	0.50	0.8227	0.74
LR	0.8218	0.74	0.4957	0.47	0.8218	0.74
MNB	0.4917	0.55	0.4316	0.35	0.4840	0.54
XGB	0.8180	0.75	0.5453	0.54	0.81897	0.76
MLP	0.8209	0.74	0.5230	0.47	0.8180	0.75
CNN	0.8219	0.82	0.8343	0.83	**0.8595**	**0.86**

Table 6: Performance results using word embeddings methods

5.7 Results

The implementation was on the open-source framework TensorFlow and Sklearn running on Python 3.6.4 environment. For all experiments, the Accuracy and F-Score metrics are applied to assess the performance.

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

$$F - score = 2 \ \frac{Precision \ . \ Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP} \ ; \ Recall = \frac{TP}{TP + FN}$$

With: TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative.

The following tables (Table 5, 8 and 9) display the performance results obtained through the use of different classifiers and feature extraction techniques. According to Table 5, the SVM classifier outperforms other supervised classifiers and CNNs in terms of accuracy and F-score across all three feature extraction techniques. The MLP and XGB classifiers also deliver satisfactory results.

For feature extraction, the bag of words (BOW) method was the most effective for the drug reviews dataset, as the sentences were short and the BOW method is well-suited for small datasets. The best results were achieved through the combination of the BOW and SVM classifiers, which produced an accuracy of 85.31% and an F-score of 85%. For text vectorization methods and the CNN classifier, the best performance was achieved using the TF-IDF method with an accuracy of 79.94% and an F-score of 78%. These results suggest that the combination of the bag of words and SVM improves classification

performance, making it a good choice for opinion mining tasks on small textual datasets.

Afterward, we evaluated the effectiveness of different embedding methods for the feature extraction step when using different classifiers. The results are presented in Table 8, which shows that the CNN classifier performed better than the other classifiers. Among the feature extraction techniques, both Word2vec models performed better than GloVe, which produced low accuracy when using supervised classifiers.

The results from Table 5 and 8 also reveal that the CNN classifier performs well when using embedding methods as inputs, rather than features generated by classical methods. This can be attributed to the fact that classical methods involve feature engineering, which results in a loss of sentence structure by treating tokens as a set rather than a sequence. Although this tokenization approach is powerful when using shallow and lightweight text-processing models like SVM and MLP classifiers, deep learning avoids this rigid and fragile approach, instead opting for structural feature learning through word embedding techniques [12]. These techniques map human language into geometric space, with the relationships between word vectors reflecting semantic relationships between the words.

Word embedding techniques also have several advantages over classical methods. Embedding vectors can be loaded from a highly structured embedding space, which captures the generic aspects of language structure, and the vectors are dense, low-dimensional, and learned from data, allowing for more information to be contained in fewer dimensions. On the other hand, vectors generated by classical methods are high-dimensional and scattered, with the same dimensionality as the number of words in the corpus. The findings of Ru et al. [79] also support these results, as they concluded that deep neural models performed better with word embedding techniques than with n-grams and contextual information.

It should be noted that the use of pre-trained word embeddings in NLP with CNN is similar to image classification, where the availability of data is crucial for learning effective features. The difference lies in the type of features learned, which are generic visual features in image classification and semantic features in NLP.

The highest performance was achieved by combining the skip-gram model and CNN, resulting in an accuracy of 85.95% and an F-score of 86%. Meanwhile, using the GloVe method produced a satisfactory result of 83.43%. These results highlight the suitability of deep learning methods for drug review opinion mining tasks. The simple CNN architecture used by WE was able to achieve optimal accuracy, demonstrating their ability to handle more advanced features compared to traditional machine learning techniques.

The LSTM and BLSTM recurrent neural networks are evaluated for drug review classification without feature extraction. The results of these algorithms can be seen in Table 9. The BLSTM outperforms the LSTM, as compared to the results achieved by the CNN and the skip-gram model. The performance of the CNN shows that the convolution layer can effectively represent the content and extract informative features when combined with the skip-gram word embedding. On the other hand, the LSTM vectorization transforms text into a sequence of integers or a vector to extract key words. The word2vec method produces the best representation of features, allowing for the acquisition of richer, more

Model	Accuracy(%)	F-score(%)
LSTM	0.8174	0.80
BLSTM	0.8296	0.83
Skipgram/CNN	**0.8595**	**0.86**

Table 7: Performance results using LSTM and BLSTM

complex, and more important features. However, the memory principle of LSTM has limitations as the input information about the features cannot be obtained from the current state of the dataset.

Furthermore, the performance of a sentiment analysis model depends not only on the task, but also on the type of textual dataset being used, according to Liu [77]. In this case, the dataset consists of reviews with relatively short sentences. As a result, word predictions can be made based on a limited context around concepts. For this reason, it is more likely that the CNN model based on Skip-gram will perform better for review classification rather than the LSTM model. The ability of CNN to effectively capture the characteristics of short sentences makes it a suitable choice, while the advantages of LSTM in handling longer sentences are not relevant in this scenario. This is also why the Bag of Words approach, which also supports small datasets, yields good results

6 Conclusion

Opinion mining based on machine learning is a fascinating and demanding field in literature. However, there are limited studies in the healthcare domain, particularly in the field of drug monitoring. Despite the fact that patient feedback on drugs, as stated by doctors and medical staff, improves their use and facilitates critical medical decisions. Hence, we present a framework for classifying opinions of drug reviews as a solution to this issue.

Our study aims to determine the best model for analyzing patient's perspectives on drug reviews to gain a better understanding of their overall opinions. In this context, we conducted a comparative study between different machine learning and deep learning methods cited in literature using well-known text vectorization techniques for exploring opinions about drugs. The results show that the combination of the skip-gram model and CNN achieved the best performance.

Our experiments lead us to the conclusion that automated models for sentiment analysis are task-specific, and to find the most efficient model, a comparative study is necessary. We also conclude that the best feature representation leads to better results and improved performance can be achieved through better collaboration between classifiers and feature extraction methods. The performance of drug reviews opinion mining can be significantly enhanced using appropriate deep learning architectures.

7 COVID-19 monitoring: A comparative analysis of public opinion mining on Social Media using machine learning and deep learning approaches

Abstract

The COVID-19 virus and its variants are currently one of the most pressing problems affecting both mental and physical health, causing stress and anxiety. News about the epidemic has rapidly spread through social media, presenting different opinions and emotions as people interact with the events. Public sentiment analysis has gained great interest in understanding people's feelings towards the virus, providing guidance for making appropriate health decisions. In this context, we analyzed COVID-19-specific tweets collected during the first few months of the crisis using deep learning (DL) approaches. Our aim was to find the best DL model for sentiment analysis about COVID-19. To achieve this, we proposed a new model that combines CNN and LSTM into one architecture. We conducted a comparative study between classical deep learning and our proposed framework to validate the model. This work was defended in the International Conference on Advances in Communication Technology, Computing and Engineering ICACTCE'21, March 24-26, 2021 CyberSpace (Virtually from Morocco).

7.1 Context of the study

At the end of December 2019, the world faced an outbreak of a new coronavirus. This infectious disease, known as COVID-19, rapidly spread around the world and had a significant impact and severe consequences on health systems, leading the World Health Organization to declare a state of emergency [88]. Currently, COVID-19 is one of the most pressing issues facing the world; it affects not only public health but also people's mental well-being. Psychologists and sociologists agree that more people are suffering psychologically from the pandemic than those who are infected with the virus.

In fact, the COVID-19 outbreak has become a source of depression, worry, stress, and anxiety. With the current lockdown and social distancing measures in place, people have become heavily dependent on the internet and social media [89]. Online forums and platforms such as Twitter have become a major part of our daily lives, providing an accessible outlet for users to express their emotions and share information during this global crisis. However, by analyzing tweets and opinions, it was found that many people are spreading false information about COVID-19, which is dangerous and concerning. It is important to report and address this misleading information.

These tweet data can also provide valuable insights for epidemiology experts to track rapidly changing public sentiments, gauge public concerns and interest, and estimate real-time COVID-19 activity and trends to monitor the transmission of the virus. As a result, researchers have focused their studies on analyzing opinions derived from tweets using Natural Language Processing (NLP) tools and machine learning (ML) methods.

Recent advancements in technology have made deep learning (DL) a popular and promising field for sentiment analysis of COVID-19 tweets. These methods use a max pooling layer approach to identify relevant and significant features of terms, resulting in improved accuracy and efficiency. The use of DL algorithms not only improves the accuracy and speed of predictions, but also reduces the prediction time.

However, for both CNN and RNN algorithms, there is no set method for dealing with significant features, as the max pooling layer selects the most important characteristics based on the highest activation value. Additionally, automated models for sentiment analysis are task-specific, and the best DL model depends not only on the capabilities of the chosen method but also on the integration of all selected NLP and ML methods in the process, as well as the type, size, and preparation of the data.

Inspired by the individual success of CNN and RNN in opinion mining tasks, we propose a new deep learning (DL) model based on a combination of RNN and CNN. This model leverages the strengths of both architectures to improve the performance of opinion classification tasks. First, we use CNN to learn relevant context features from the input representation. Then, these context features are used as input for the RNN models. Finally, the feature map from RNN is used to perform the opinion classification task.

To uncover the most effective model for examining public emotions and psychological responses during the COVID-19 crisis, in this part section, we provide a summary of related studies on COVID-19 opinion mining monitoring, along with a comprehensive explanation of our proposed approach. This includes the proposed architecture combining CNN and RNN and the steps involved in comparing various techniques. We then evaluate and delve into the results of our comparison experiments. Finally, we draw conclusions and offer future directions.

7.2 Related Works

The present study examines the capability of emotional analysis in Twitter data, particularly in the healthcare field. Due to the recent success of CNN and RNN in opinion classification, some researchers have explored combining these two networks into hybrid models. One such example is the study by Wang et al. [90] who proposed a combination of CNN and RNN for performing opinion mining on short texts. The CNN was employed to extract high-quality linguistic features, while the RNN learned long-word dependencies. The hybrid architecture was tested on three benchmark sentiment analysis datasets.

Another study, [91], proposed a framework that utilized LSTM, CNN, and a highway network for language modeling. [92] combined CNN and LSTM for sentence classification. [93] used a multi-architecture-based feature fusion approach of CNN and RNN for sentiment analysis of social media comments. [94] proposed a new RNN model with a CNN-based attention mechanism. This model used Glove word embedding techniques for text vectorization, followed by CNN filters for feature extraction and RNN for sequential processing of the features.

With the ongoing COVID-19 pandemic, numerous machine learning studies have been conducted to examine its impact on people's emotions. [95] proposed a neural network to analyze Twitter comments in Europe and tested various pre-trained word-embedding methods such as word2vec, multilingual BERT, MUSE, and the Vader algorithm. The results showed that MUSE and BERT provided better results compared to the word2vec model. [96] analyzed public sentiment towards COVID-19 using RNN to predict opinions on Twitter data. The model first identified connections between terms and then labeled them with positive or negative sentiment. [97] aimed to show how popularity affects accuracy on social media and proposed sentiment analysis of COVID-19 tweets using various deep learning classifiers, with doc2vec models and Gaussian membership function-based fuzzy rules providing successful results in sentiment identification. [98]

proposed a transformer-based model pre-trained on a large dataset of COVID-19-related tweets, which extends the deep BERT model and was improved for use in COVID-19 tweets. [99] evaluated two CNN models to analyze Twitter user behavior in response to religious misinformation during the COVID-19 pandemic in India. [100] proposed a CNN deep learning model supported by pre-trained BERT for converting tweets into mathematical vectors, with results showing that the proposed model outperformed state-of-the-art models.

In recent years, several works have investigated this issue, such as [101], [102], and [103]. These studies have utilized feature extraction methods such as TF-IDF and bag of words, and employed both supervised and ensemble machine learning classifiers. The results of these studies indicate that the best performance is typically achieved using a supervised Naive Bayes algorithm.

Through literature review, several deep learning-based methods have been proposed for opinion mining in medical domains. Most of these studies concentrate on advanced deep learning models that incorporate both CNN and RNN, and they have demonstrated good performance. For sentiment analysis of COVID-19, researchers are trying to analyze public opinion regarding the pandemic using machine learning methods. Some studies adopt classical methods for text vectorization and sentiment classification, while others use deep learning for both sentiment analysis and feature extraction. There are also a few studies that only employ CNN or RNN for sentiment classification.

Given the promising results of hybrid deep learning approaches, conducting a comparative analysis would be valuable. Hence, this study aims to determine the most efficient and accurate framework for opinion mining from social media for COVID-19 monitoring. To achieve this goal, the study proposes a comparison between classical and advanced deep learning methods combined with NLP techniques, data preprocessing, and feature extraction methods (such as Glove, word2vec, TF-IDF, and bag of words) for categorizing COVID-19-specific tweets as negative, positive, or neutral.

7.3 Proposed framework

This section outlines the approach used in this work and is divided into four parts. The first part involves preprocessing the tweets by data cleaning, including removal of links, punctuation, stop words, and tokenization, to prepare them for sentiment analysis. The second part uses the Vader library to calculate the sentiment of the cleaned data. In the third part, classical and embedded methods such as Bag of Words, N-Gram, word2vec, and Glove are used for feature extraction through text vectorization. Finally, the last part employs Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) classifiers, such as LSTM and Bi-LSTM, to classify the data into three categories: negative, positive, and neutral. The entire process is depicted in Figure 26.

7.3.1 Step1 : data pre-processing

First, we tokenizes each tweet into smaller units called tokens, as described in [78]. The process continues with the removal of user names, hyperlinks, and conversion of all tweets to lowercase. Terms with two or fewer characters, white-space, etc. are also removed. However, punctuation is retained as it is important for sentiment analysis. Stop

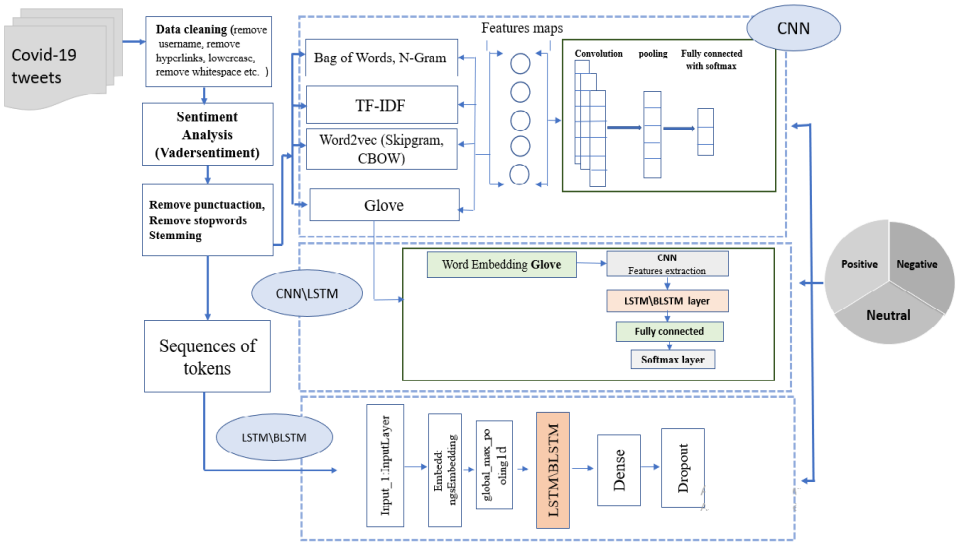

Figure 26: Diagram representing the overall operating process for COVID-19 opinion mining.

words, which carry little significant information [104], are also removed and stemming is applied to shorten the lookup and normalize sentences.

7.3.2 Step2: Sentiment Analysis

This step aims to determine the public sentiment towards the COVID-19 pandemic from social media by calculating the compound sentiment polarity scores of each tweet and categorizing them into three labels: positive, neutral, and negative. We use the Vader sentiment library [105] for sentiment analysis as it considers important factors such as exclamation marks and emotions, and takes punctuation into account when calculating polarity scores, making it the most suitable choice.

7.3.3 Step3 : features modules

This step involves transforming the cleaned tweets into numerical vector representations, which are useful for the classification step. In this study, we used several traditional methods such as N-grams [12], TF-IDF [14], Bag of Words [13], and word embedding methods Word2Vec [60] and Glove [16] to generate vector representations from the dataset.

7.3.4 Step4: Deep Learning Methods

In this step, the feature maps are processed through multiple and successive layers of deep learning classifiers. The number of layers, data representation, and advancement in deep learning architectures are crucial factors for the network's operation and the efficiency of the deep learning method. Hence, the deep learning techniques applied in the framework are briefly described.

Convolutional Neural Network (CNN) The input tweets are transformed into sequences of tokens or word embedding vectors in the feature module. These features

are then fed into a Convolutional Neural Network (CNN) for sentiment analysis. Our study explores several CNN architectures, and the best-performing architecture is shown in Figure 27. When combined with the Glove method, this architecture yields particularly good results.

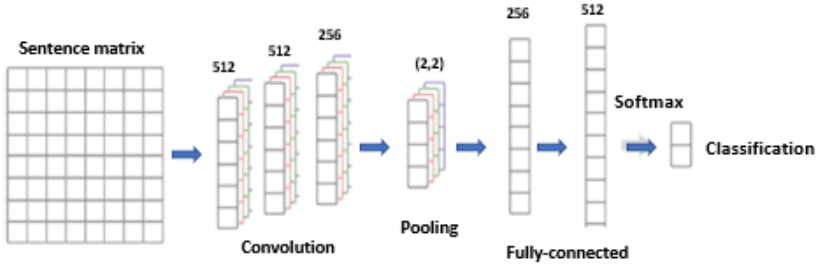

Figure 27: Architecture of the proposed CNN

The CNN architecture consists of three main components: Convolution Layer, Pooling Layer, and Fully Connected Layer. The Convolution Layer extracts features from the tweets by sliding filters over the tweets, computing the product of the weights of each term and the filter's weights. The Pooling Layer reduces the dimensionality of the feature maps and retains a simple probability score that indicates the likelihood of the corresponding label. This layer also helps to reduce computational cost, prevent overfitting, and display features more efficiently. In our study, researchers used dropout with a rate of 0.2, removing a portion of the network to promote distributed learning.

The Fully Connected Layer takes the results from the Convolution and Pooling Layers, computes a representative number, and feeds it into a network of backpropagation processes. The network then makes a classification decision (positive, negative, or neutral sentiment) based on the weights assigned to each feature of the tweets. In our study, the number of neurons with the ReLU activation function was set to 64, adding non-linearity to the network.

Long Short-Term-Memory (LSTM) and Bidirectional Long Short-Term-Memory (Bi-LSTM) Figure 28 illustrates the proposed LSTM model, which demonstrates how the LSTM utilizes the token sequences of each tweet to classify its sentiment.

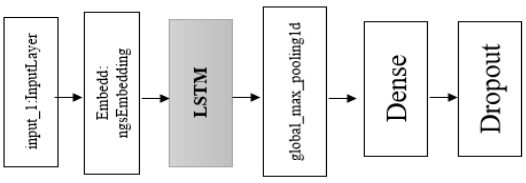

Figure 28: Architecture of the proposed LSTM.

In this approach, a single cleaned tweet is interpreted as a sequence of 80000 unique words learned through the use of convolutional filters. The relevant terms, along with their corresponding sentiments, are then processed by LSTM and Bi-LSTM, including:

- The Embedding layer transforms the tokens in the dataset into 200-dimensional embeddings.
- The LSTM layer comprises one LSTM layer with 256 hidden units.

- The Dropout layer aids in avoiding over-fitting, and one dropout layer was employed for regularization purposes.
- The Fully Connected layer maps the output from the LSTM to the desired classes using the softmax activation function. For the Bi-LSTM model, two fully connected layers were used. The first one consisted of 30 neurons with the 'ReLU' activation function, while the second one included three fully connected layers with a softmax activation function.

7.3.5 Hybrid deep learning model

The proposed model aims to leverage the advantages of both CNN and RNN (LSTM/Bi-LSTM) architectures to effectively classify tweets based on both contextual and temporal features.

In the proposed architecture, the CNN first extracts high-level features from the pre-trained Glove embeddings, which were generated from Wikipedia. The context features generated by the convolutional layers are then passed on to the RNN models for sequential processing and classification, as depicted in Figure 29.

This combination of CNN and RNN models allows the proposed architecture to capture both the relevant and high-level features from the Glove embeddings and the contextual and temporal features from the RNN models, resulting in improved performance for tweet classification tasks.

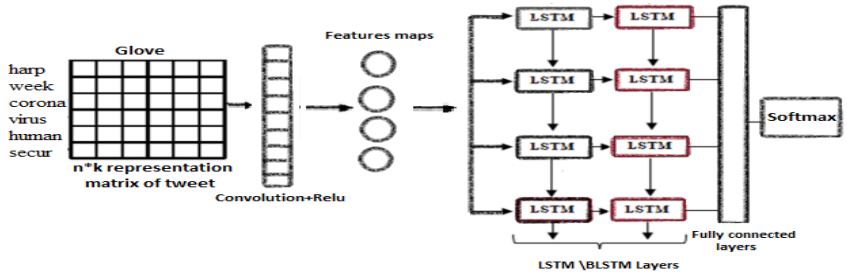

Figure 29: Combine architecture of Hybrid deep learning model.

7.4 Experimental results

7.4.1 Data description

The database used in this study is a set of tweets (COVID-19 Tweets Dataset) [106] collected as part of an ongoing published project [2]. It consists of a group of CSV files that contain identifiers and scores for various feelings related to the COVID-19 pandemic expressed in English. The dataset was updated until March 20, 2020, and for our study, a CSV file containing 1578957 tweets was selected and the dataset was fixed at 76399 tweets.

7.4.2 Results

The results of our experiments, which compared the performance of various algorithms including both classical and deep learning models, are presented in this part of the study.

[2] https://live.rlamsal.com.np

To determine the best model, the Accuracy and F-Score metrics were applied to assess their performance. The performance values obtained using different classifiers and feature extraction techniques are shown in Tables 8 and 9.

Table 8: Performance results

Methods	**Representation**	**Accuracy**	**F-score**
LSTM	sequence-level	0.9761	0.98
Bi-LSTM	sequence-level	0.9744	0.97
CNN	Bag of words	0.9679	0.97
CNN	TF-IDF	0.9601	0.96
CNN	Bigram TFIDF	0.9607	0.96
CNN	Glove	0.9743	0.97
LSTM	Glove	0.9431	0.94
Bi-LSTM	Glove	0.9558	0.96
CNN	CBOW	0.9235	0.93
CNN	SkipGram	0.9475	0.94
CNN/LSTM	Glove	**0.9771**	**0.98**
CNN/Bi-LSTM	Glove	0.9764	0.98

Table 9: Performance results

	TF-IDF		TF-IDF- 2 gramms		Bag of words		Glove	
Classifiers	**Acc**	**F-score**	**Acc**	**F-score**	**Acc**	**F-score**	**Acc**	**F-score**
SVM	0.9713	0.97	0.9691	0.97	0.9731	0.97	0.4362	0.45
RF	0.4970	0.41	0.5320	0.46	0.4680	0.36	0.5397	0.49
LR	0.9506	0.95	0.9496	0.95	0.9666	0.97	0.4499	0.35
MNB	0.8545	0.86	0.8547	0.86	0.8388	0.84	0.4294	0.29
XGB	0.9168	0.92	0.9176	0.92	0.9160	0.92	0.7899	0.54
MLP	0.9637	0.96	0.9613	0.96	0.9709	0.97	0.4135	0.24

Table 8 presents the comparative experimental results obtained using different traditional and deep learning models on the same training and test data sets. The performance of the classifiers is evaluated using various pre-trained word representation models.

In the experiments, we used the gensim Python library, which includes Word2Vec models. We applied two Word2Vec models: one using the Continuous Bag Of Words (CBOW) model, and the other using the skip-gram model to extract word vectors. For the Glove method, we used a specific set of Glove vectors trained on tweets, which includes four different versions of tweet vectors, each with different dimensions (25, 50, 100, 200), trained on 2 billion tweets. In our study, we used the 50-dimension pre-trained Glove vectors available at [3].

From Table 9, it can be observed that the SVM classifier performs better in terms of accuracy and F-score compared to other supervised classifiers for all three feature extraction techniques. It achieved the highest accuracy of 97.31%. The MLP algorithm also shows a significantly better performance with an accuracy of 97.09% and an F-score

[3]https://nlp.stanford.edu/projects/Glove/

of 97%. Experiments indicate that LR and XGB classifiers provide satisfactory results. However, the RF classifier performs poorly for opinion classification for all three text vectorization techniques. Additionally, the GloVe embedded method results in poor accuracy when using supervised classifiers.

From Table 9, we can observe that the hybrid models of CNN/LSTM and CNN/Bi-LSTM, in combination with GloVe, outperform all other models for all feature extraction techniques. These models achieved the highest accuracy of 97.71% and 97.64%, respectively. This performance is nearly 1% better than the results obtained from the LSTM, Bi-LSTM models when using sequence-level representations and the CNN model supported by GloVe.

In our study, we first tokenized the tweets into words and removed stop words, resulting in the loss of sequential form of the data. In this case, the use of CNN for feature extraction was more appropriate, as it is commonly used to solve problems related to non-sequential inputs that can be treated as images and used as the input for CNN. Additionally, the use of pre-trained word embedding in NLP with CNN is similar to image classification, where the availability of data plays a crucial role in learning powerful features. This is demonstrated by the use of the GloVe method, which can produce high-quality vector representations using word co-occurrence statistics in a matrix form and can also cover out-of-vocabulary terms. The convolutional CNN layer can effectively represent these vectors and acquire more informative features, which are then used in the CNN max pooling layer to find relevant, significant features of terms.

However, opinion mining of COVID-19 specific tweets requires learning of implicit long-distance dependencies through the sequences of tweets. Hence, we used the LSTM layer, which contains an internal memory capable of learning from imperative experiences with long-term conditions. Unlike the fully connected nodes in the CNN layer, the nodes in the LSTM layer are connected in a directed graph along a time sequence.

Based on the results, it can be observed that the CNN architecture has the advantage of focusing on the most important and relevant features and being capable of efficiently capturing local, window-based compositions. Meanwhile, LSTM is effective in learning long-term dependencies. The hybrid model, which combines the strengths of both CNN and LSTM, proves to be effective in handling sequence prediction problems with spatial inputs. Furthermore, LSTM and Bi-LSTM models with sequence-level representation also contribute to improving the representation and classification of tweets specific to COVID-19.

7.5 Conclusion

Public sentiment analysis of the pandemic using COVID-19 specific tweets has garnered great interest among researchers in the fields of health and data mining. This study investigates the impact of using deep learning algorithms in sentiment analysis of these tweets with the goal of finding the best model to analyze people's feelings towards the pandemic and gain insight into their overall opinions. A new hybrid CNN/LSTM model is proposed that combines the strengths of both architectures. To validate the model, a comparison between traditional DL techniques and the proposed framework

is made based on accuracy and F1-score metrics, and the results show that the hybrid CNN/LSTM model outperforms others with higher accuracy. The experiments suggest that a comparative study is necessary to determine the most efficient model for opinion mining. Features extraction is a crucial factor in the success of these deep learning models, and the appropriate deep learning classifier architecture can significantly improve performance. Our findings may aid in the improvement of practical strategies for health services related to COVID-19.

8 Synthesis analysis

In this chapter, we discuss the importance of medical sentiment analysis in the healthcare domain and how text mining and machine learning techniques can be used to develop robust models for opinion mining in medical texts. Our primary goal is to demonstrate the effectiveness of these methods in sentiment analysis, and to that end, we have selected two areas where the patient's opinion is crucial for the medical service. The first area concerns patient feedback about their treatment, while the second focuses on analyzing public sentiment towards the COVID-19 pandemic.

In this synthesis analysis, we can further elaborate on the technical arguments and comparative analysis of models' potential. Firstly, it's worth noting that sentiment analysis in the medical domain is a challenging task due to the complexity and specificity of medical language. The utilization of text mining and machine learning techniques can aid in developing robust models for opinion mining in medical texts, but it's crucial to select the appropriate techniques and models to achieve the best results.

To conduct a comparative study between machine learning techniques, we utilized traditional and advanced deep learning methods to determine the best model that would provide the most accurate results. The pre-processing techniques used in each database differed from one another, depending on the type of data.

Our findings indicate that deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), were effective in the first study. Thus, we proposed a hybridization of these techniques in the second contribution, combining the advantages of both in features extraction and classification. The experiments were conducted based on accuracy and F1-score metrics.

Despite the promising results, there were limitations to our study. For instance, our dataset was small and specific to the autopsy reports, patient feedback, and COVID-19 tweets, which could limit the generalizability of our findings. Additionally, some aspects of our study, such as feature selection, could have been improved. However, our work demonstrates the potential of text mining and machine learning techniques in the medical domain and highlights the need for further research in this field.

9 Conclusion

Patients often have strong feelings about their healthcare experiences, and these emotions can have a significant impact on the patient's overall satisfaction with the care they receive. As a result, sentiment analysis has become increasingly important in the healthcare industry. This technology can be used to help identify patient needs, improve marketing efforts, and ultimately enhance the patient experience.

In this chapter, we have explored the use of natural language processing and machine learning methods for medical sentiment analysis. Our work has focused on two key contributions: patient feedback regarding their treatment and the analysis of public sentiment surrounding the COVID-19 pandemic. Throughout our research, we have investigated the effectiveness of various deep learning algorithms in sentiment analysis and

have conducted a comparative study to identify the most efficient model.

Our experiments have led us to several key findings. First, we have discovered that automated models for opinion analysis must be task-specific, and a comparative study is necessary to determine the most efficient model. Additionally, we have determined that feature extraction plays a crucial role in the success of deep learning models. Finally, we have found that the appropriate deep learning classifier architecture can significantly improve performance through better collaboration between classifiers and feature extraction methods.

Overall, our research has significant implications for the healthcare industry. By improving sentiment analysis techniques, we can enhance patient experiences and ultimately improve business outcomes on a larger scale. Additionally, our findings could be used to develop more effective strategies for drug monitoring and COVID-19 surveillance.

Conclusion and future directions

This thesis focuses on text mining and its main applications in healthcare, particularly document classification and sentiment analysis. Many real-world problems in the medical field require text mining approaches to be solved. This manuscript first provides a detailed overview of the medical text mining frameworks, including medical document classification and medical sentiment analysis, applications, and common tools and strategies used for each approach, and finally presents the current challenges recently investigated by researchers in this field.

These automated methods significantly support the decision-making process of healthcare professionals, such as monitoring patients' emotions towards their medications through drug monitoring or discovering new knowledge through automatic classification of medical reports, such as autopsy report classification.

Then, we investigated the benefits of using text mining and machine learning techniques for medical document classification, focusing on the classification of autopsy reports. This plays a crucial role in the medical field, as autopsy reports contain vital information about the cause of death (CoD) that can aid doctors in improving their understanding of this area. Our study consisted of two main parts: data collection and algorithm comparison. The data collection phase involved collecting 200 autopsy reports from the CHU Tlemcen Forensic Medicine department, which contained information about external and internal examinations of the deceased, personal information, and corresponding labels for the manner of death (natural, violent, or toxic).

The first part of this work compared traditional machine learning and deep learning algorithms for autopsy report classification, using the TF-IDF method for feature extraction and analyzing the performance of each algorithm when combined with this tool. The goal was to find the best model of classification in terms of accuracy and medical interpretation. The experimental results showed that the XGB classifier combined with TF-IDF performed best in terms of accuracy, while deep learning methods performed poorly when combined with TF-IDF. To interpret the output of our approach, we used the evaluation metric LIME, which is a great tool for making complex models interpretable.

The conclusion of this study suggests that further work should be done to build upon this initial contribution. Specifically, there are several areas that could be addressed in future research:

- The current study analyzed a limited set of only 200 autopsy reports obtained from the University Hospital of Tlemcen. To enhance the generalizability and reliability of the results, it is recommended to collect more autopsy reports to expand the training data set.
- The study used a traditional feature extraction method, TF-IDF, for text vectorization, which yielded the best results when combined with XGB classifier. However, deep learning methods did not perform as well with this method due to the fact that they prefer to learn structural features through word embedding techniques. To address this issue, future work may explore the use of contextualized language models, such as BERT (Bidirectional Encoder Representations from Transformers) [20], CamemBERT, or RoBERT [107], which have shown promising performance in various language processing tasks, particularly in French language processing.
- The collected reports were classified into three categories of Manner of Death (MoD): natural death, violent death (including penetrating, suicide, and homicide), and toxic death. Each category contains reports with multiple unique causes of death (CoD), resulting in a dataset with 12 different CoD across the three categories. To further expand the application of this study, future work may consider using a multi-label classification approach to detect the specific cause of death from these reports.

The second part of this thesis focuses on medical sentiment analysis using NLP and machine learning methods. Opinion mining is a research technique that measures patient satisfaction with the healthcare services they have received, aiming to improve and optimize the quality of healthcare services in various healthcare facilities.

The framework addresses two major contributions where the patient's opinion is crucial to the medical service:

- The first contribution concerns patients' feedback about their treatment. The study aims to determine the best model for analyzing patient perspectives on drug reviews to gain a better understanding of their overall emotion. In this context, a comparative study between different machine learning and deep learning methods was conducted, using well-known text vectorization techniques to explore opinions about drugs. The results show that the combination of the skip-gram model and CNN achieved the best performance.
- The second contribution analyzes public opinion on the COVID-19 pandemic using COVID-19-specific tweets. The study demonstrates the impact of using deep learning algorithms for sentiment analysis of these tweets, with the aim of analyzing people's sentiments towards the pandemic and finding the best models to gain insight into their overall emotions. Given the promising results of deep learning approaches in the first contribution, a new hybrid CNN/LSTM model is proposed in this work that combines the strengths of both architectures. To validate the model,

a comparison between traditional DL techniques and the proposed framework is performed. The results show that the hybrid CNN/LSTM model outperforms the other models with higher accuracy.

From our experiments in both studies, we conclude that automated models of opinion analysis are task-specific, and comparative studies are necessary to find the most efficient models. We also conclude that feature extraction is a key factor in the success of these deep learning models and that a proper deep learning classifier architecture can significantly improve performance through better collaboration between classifiers and feature extraction methods. Our findings will help improve practical strategies for drug surveillance and health services related to COVID-19 surveillance.

In further work, we plan to:

- Include additional datasets covering other drugs and diseases in the future. We also aim to explore other types of deep learning techniques such as Transformer models, other hybrid approaches, and new vocabulary resources for better results.
- Focus on analyzing clinical documents for sentiment analysis, as existing studies of sentiment analysis from medicinal texts have mainly focused on medical social media and biomedical literature. This analysis can be used for various purposes, including therapeutic decision making.
- Focus on Algerian health monitoring through the analysis of sentiment in biomedical texts to address analysis of Algerian sentiment and Algerian health status based on machine learning and deep learning classifiers. This requires creating an Arabic sentiment dataset consisting of health-related tweets from professional Algerian Twitter accounts.

Scientific outputs

Numerous methodological contributions have resulted from this thesis, these productions are parallel or simultaneously integrated into some major parts or chapters of the thesis. Several collaborative works with colleagues have served to address several direct and indirect reflections on the thesis.

Chapter Book

- N. SETTOUTI and **F. YOUBI**. *Convolutional and Recurrent Neural Networks for opinion mining on Drug Reviews*. In book : Deep Learning for Social Media Data Analytics in Springer Book Series "Studies in Big Data". February 2022.

Peer-reviewed Journals

- **F. YOUBI** and N. SETTOUTI. *Analysis of Machine Learning and Deep Learning frameworks for opinion mining on drug reviews*. In the Computer Journal, 2021; bxab084, DOI: 10.1093/comjnl/bxab084

International communication

- **F. YOUBI** and N. SETTOUTI. *COVID-19 tweets sentiment analysis using machine learning approaches and divers document representations*. In the International Conference on Advances in Communication Technology, Computing and Engineering ICACTCE'21, March 24-26, 2021 CyberSpace (Virtually from Morocco).
- **F. YOUBI** and N. SETTOUTI. *Convolutional Neural Networks for opinion mining on Drug reviews*. In Proceedings of the 1st International Conference on Intelligent Systems and Pattern Recognition (ISPR '20). Association for Computing Machinery, New York, NY, USA, 17–21. DOI: 10.1145/3432867.3432888 (**Best Paper Award**)
- **F. YOUBI** and N. SETTOUTI and M. SAIDI. *Machine learning methods to predict chemotherapy treatment of Multiple Myeloma patients using clinical data*. International Congress on Health Science and Medical Technologies 2021, ICHSMT'21, 27-29 June 2021, Tlemcen, Algeria.
- M. SAIDI, **F. YOUBI** and N. SETTOUTI. *Instance selection algorithms for a Cost sensitive medical diagnosis*. In the Third International Conference on Biotechnology and Cancer (ICBC'19), December 07-08, 2019 in Oran, Algeria.

National communication

- **F. YOUBI** and N. SETTOUTI. *Le réseau neuronal récurrent à portes pour la prédiction des émotions à partir du signal EEG*. Le Colloque National sur l'Apport des Neurosciences dans le Marketing-Management (Neuromarketing- Neuromanagement). Online Conference, Juin 09, 2022, Tlemcen, Algeria.

Research Report

- **F. YOUBI**, Souhila Lairibi, Nesma Settouti. Autopsie Médicale : étude de la mortalité et causes de décès au niveau CHU Tlemcen, Algérie. Biomedical Engineering Laboratory, Tlemcen University Algeria. 2023. `hal-04055998`

Collected Dataset

- **Youbi, Fatiha**; LARIBI, Souhila; SETTOUTI, Nesma (2023), ''Autopsy reports Datasets'', Mendeley Data, V1, doi:`10.17632/n9z3v2k8wv.1`

Bibliography

[1] Carmen Luque, José M Luna, Maria Luque, and Sebastian Ventura, ''An advanced review on text mining in medicine,'' *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, pp. e1302, 2019.

[2] Sholom M Weiss, Nitin Indurkhya, and Tong Zhang, ''Overview of text mining,'' in *Fundamentals of Predictive Text Mining*, pp. 1–12. Springer, 2010.

[3] Sholom M Weiss, Nitin Indurkhya, and Tong Zhang, ''Using text for prediction,'' in *Fundamentals of Predictive Text Mining*, pp. 41–79. Springer, 2015.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, ''Efficient estimation of word representations in vector space,'' *arXiv preprint arXiv:*, vol. 1301.3781, 2013.

[5] Alex Graves, ''Generating sequences with recurrent neural networks,'' *arXiv preprint arXiv:*, vol. 1308.0850, 2013.

[6] Basant Agarwal, Richi Nayak, Namita Mittal, and Srikanta Patnaik, ''Deep learning-based approaches for sentiment analysis,'' 2020.

[7] Hercules Dalianis, *Clinical text mining: Secondary use of electronic patient records*, Springer Nature, 2018.

[8] Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol, ''Text summarization in the biomedical domain: a systematic review of recent research,'' *Journal of biomedical informatics*, vol. 52, pp. 457–467, 2014.

[9] John A Baron, Stephen Senn, Michael Voelker, Angel Lanas, Irene Laurora, Wolfgang Thielemann, Andreas Brückner, and Denis McCarthy, ''Gastrointestinal adverse effects of short-term aspirin use: a meta-analysis of published randomized controlled trials,'' *Drugs in R&D*, vol. 13, no. 1, pp. 9–16, 2013.

[10] Ahmad P Tafti, Jonathan Badger, Eric LaRose, Ehsan Shirzadi, Andrea Mahnke, John Mayer, Zhan Ye, David Page, Peggy Peissig, et al., ''Adverse drug event discovery using biomedical literature: a big data neural network adventure,'' *JMIR medical informatics*, vol. 5, no. 4, pp. e9170, 2017.

[11] Zhang J. Zong C., Xia R., "Text representation. in: Text data mining," *Springer, Singapore*, (2021).

[12] Francois Chollet, *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek*, MITP-Verlags GmbH & Co. KG, 2018.

[13] Yin Zhang, Rong Jin, and Zhi-Hua Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[14] Claude Sammut and Geoffrey I. Webb, Eds., *TF–IDF*, pp. 986–987, Springer US, Boston, MA, 2010.

[15] Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi, "Sentiment analysis based on improved pre-trained word embeddings," *Expert Systems with Applications*, vol. 117, pp. 139–147, 2019.

[16] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[17] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov, "Fasttext. zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[18] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[19] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo, "Evaluating word embedding models: Methods and experimental results," *APSIPA transactions on signal and information processing*, vol. 8, 2019.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," 2018.

[22] Abinash Tripathy, *Sentiment Analysis Using Machine Learning Techniques*, Ph.D. thesis, 2017.

[23] Ashraf M. Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes, "Multinomial naive bayes for text categorization revisited," in *Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence*, Berlin, Heidelberg, 2004, AI04, p. 488499, Springer-Verlag.

[24] Vladimir Vapnik, Steven E. Golowich, and Alexander J. Smola, "Support vector method for function approximation, regression estimation and signal processing," in *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, 1996, pp. 281–287.

[25] Shengyi Jiang, Guansong Pang, Meiling Wu, and Limin Kuang, "An improved k-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012.

[26] Leo Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 532, Oct. 2001.

[27] David W. Hosmer and Stanley Lemeshow, *Applied logistic regression*, John Wiley and Sons, 2000.

[28] Kamal Nigam, John Lafferty, and Andrew McCallum, "Using maximum entropy for text classification," in *IJCAI-99 workshop on machine learning for information filtering*. Stockholom, Sweden, 1999, vol. 1, pp. 61–67.

[29] Sepp Hochreiter and Jürgen Schmidhuber, "Lstm can solve hard long time lag problems," *Advances in neural information processing systems*, pp. 473–479, 1997.

[30] Nhan Cach Dang, María N Moreno-García, and Fernando De la Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electronics*, vol. 9, no. 3, pp. 483, 2020.

[31] Rajkumar S Jagdale, Vishal S Shirsat, and Sachin N Deshmukh, "Sentiment analysis on product reviews using machine learning techniques," in *Cognitive Informatics and Soft Computing*, pp. 639–647. Springer, 2019.

[32] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[33] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[35] Patrawut Ruangkanokmas, Tiranee Achalakul, and Khajonpong Akkarajitsakul, "Deep belief networks with feature selection for sentiment classification," in *2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*. IEEE, 2016, pp. 9–14.

[36] Haixia Long, Bo Liao, Xingyu Xu, and Jialiang Yang, "A hybrid deep learning model for predicting protein hydroxylation sites," *International journal of molecular sciences*, vol. 19, no. 9, pp. 2817, 2018.

[37] Rahul Ghosh, Kumar Ravi, and Vadlamani Ravi, "A novel deep learning architecture for sentiment classification," in *2016 3rd international conference on recent advances in information technology (RAIT)*. IEEE, 2016, pp. 511–516.

[38] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Mohammed Ali Al-Garadi, Retnagowri Rajandram, and Khairunisa Shaikh, "Hierarchical text classification of autopsy reports to determine mod and cod through term-based and concepts-based features," in *Advances in Data Mining. Applications and Theoretical Aspects*, Petra Perner, Ed., Cham, 2017, pp. 209–222, Springer International Publishing.

[39] Samuel Danso, Eric Atwell, and Owen Johnson, "A comparative study of machine learning methods for verbal autopsy text classification," *arXiv preprint arXiv:1402.4380*, 2014.

[40] Wei Liang Yeow, Rohana Mahmud, and Ram Gopal Raj, "An application of case-based reasoning with machine learning for forensic autopsy," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3497–3505, 2014.

[41] Muhammad Rehman Shahid, Asim Munir, Layeba Ifraheem, Hamza Aldabbas, Abdul Wadood, and Tariq Alwada'n, "Machine learning for autopsy reports forensic using text classification techniques," in *2022 2nd International Conference on Computing and Information Technology (ICCIT)*. IEEE, 2022, pp. 148–153.

[42] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, and Khairunisa Shaikh, "Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study," *Journal of forensic and legal medicine*, vol. 57, pp. 41–50, 2018.

[43] Francisco Duarte, Bruno Martins, Cátia Sousa Pinto, and Mário J Silva, "Deep neural models for icd-10 coding of death certificates and autopsy reports in free-text," *Journal of biomedical informatics*, vol. 80, pp. 64–77, 2018.

[44] Esteban Guillen, Trilce Estrada, and Matthew Cain, "Deepmanner: Automatically determining manner of death," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1361–1366.

[45] Zhaodong Yan, Serena Jeblee, and Graeme Hirst, "Can character embeddings improve cause-of-death classification for verbal autopsy narratives?," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, 2019, pp. 234–239.

[46] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, KDD 16, p. 785794, Association for Computing Machinery.

[47] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*, p. 318362, MIT Press, Cambridge, MA, USA, 1986.

[48] Thorsten Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[49] Muhammad Rehman Zafar and Naimul Mefraz Khan, "Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," *arXiv preprint arXiv:1906.10263*, 2019.

[50] Ranjan Satapathy, Erik Cambria, and Amir Hussain, "Sentiment analysis in the bio-medical domain," *Springer Interntional Publishing AG, Gwerbestrasse*, vol. 11, pp. 6630, 2017.

[51] Bayu Yudha Pratama and Riyanarto Sarno, "Personality classification based on twitter text using naive bayes, knn and svm," in *2015 International Conference on Data and Software Engineering (ICoDSE)*. IEEE, 2015, pp. 170–174.

[52] Céline Battaïa, ''Lanalyse de lémotion dans les forums de santé (analysis of emotion in health fora)[in french],'' in *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 3: RECITAL*, 2012, pp. 267–280.

[53] Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya, ''Medical sentiment analysis using social media: towards building a patient assisted system,'' in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[54] Kerstin Denecke and Yihan Deng, ''Sentiment analysis in medical settings: New opportunities and challenges,'' *Artificial intelligence in medicine*, vol. 64, no. 1, pp. 17–27, 2015.

[55] Frederick C Mish, *Merriam-Webster's collegiate dictionary*, vol. 1, Merriam-Webster, 2004.

[56] Penubaka Balaji, O Nagaraju, and D Haritha, ''Levels of sentiment analysis and its challenges: A literature review,'' in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE, 2017, pp. 436–439.

[57] Soroosh Afyouni, Ahmed E Fetit, and Theodoros N Arvanitis, ''Perspectives through social media analysis,'' *Enabling Health Informatics Applications*, vol. 213, pp. 243, 2015.

[58] Tanveer Ali, David Schramm, Marina Sokolova, and Diana Inkpen, ''Can i hear you? sentiment analysis on medical forums,'' in *Proceedings of the sixth international joint conference on natural language processing*, 2013, pp. 667–673.

[59] Peter D Turney, ''Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,'' in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 417–424.

[60] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, ''Efficient estimation of word representations in vector space,'' *arXiv preprint arXiv:1301.3781*, 2013.

[61] Kushal Dave, Steve Lawrence, and David M Pennock, ''Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,'' in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 519–528.

[62] Lorraine Goeuriot, Jin-Cheon Na, Wai Yan Min Kyaing, Christopher Khoo, Yun-Ke Chang, Yin-Leng Theng, and Jung-Jae Kim, ''Sentiment lexicons for health-related opinion mining,'' in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 2012, pp. 219–226.

[63] Bo Pang, Lillian Lee, et al., ''Opinion mining and sentiment analysis,'' *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[64] Tejaswini M Untawale and G Choudhari, ''Implementation of sentiment classification of movie reviews by supervised machine learning approaches,'' in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2019, pp. 1197–1200.

[65] Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath, "Classification of sentimental reviews using machine learning techniques," *Procedia Computer Science*, vol. 57, pp. 821–829, 2015.

[66] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*. Springer, 2003, pp. 986–996.

[67] Larry R Medsker and LC Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, pp. 64–67, 2001.

[68] Ozan Irsoy and Claire Cardie, "Opinion mining with deep recurrent neural networks," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 720–728.

[69] Soujanya Poria, Erik Cambria, and Alexander Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowledge-Based Systems*, vol. 108, pp. 42–49, 2016.

[70] Jagannatha AN and Hong Y., "Bidirectional rnn for medical event detection in electronic health records.," *Proceedings of NAACL-HLT*, vol. pp. 473-82, 2016,.

[71] Andrew L Beam, Benjamin Kompa, Inbar Fried, Nathan P Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane, "Clinical concept embeddings learned from massive sources of multimodal medical data," *arXiv preprint arXiv:1804.01486*, 2018.

[72] Alper Kürşat Uysal, "Comparative performance analysis of techniques for automatic drug review classification," *Celal Bayar Üniversitesi Fen Bilimleri Dergisi*, vol. 14, no. 4, pp. 485–490.

[73] Tianhua Chen, Pan Su, Changjing Shang, Richard Hill, Hengshan Zhang, and Qiang Shen, "Sentiment classification of drug reviews using fuzzy-rough feature selection," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2019, pp. 1–6.

[74] Shweta Yadav, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya, "Medical sentiment analysis using social media: towards building a patient assisted system," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[75] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez, "Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks," in *Proceedings of the 2010 workshop on biomedical natural language processing*. Association for Computational Linguistics, 2010, pp. 117–125.

[76] Vinodhini Gopalakrishnan and Chandrasekaran Ramaswamy, "Patient opinion mining to analyze drugs satisfaction using supervised learning," *Journal of applied research and technology*, vol. 15, no. 4, pp. 311–319, 2017.

[77] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang, ''Drug-drug interaction extraction via convolutional neural networks,'' *Computational and mathematical methods in medicine*, vol. 2016, 2016.

[78] Sisi Liu and Ickjai Lee, ''Sentiment classification with medical word embeddings and sequence representation for drug reviews,'' in *International Conference on Health Information Science*. Springer, 2018, pp. 75–86.

[79] Boshu Ru, Dingcheng Li, Yueqi Hu, and Lixia Yao, ''Serendipitya machine-learning application for mining serendipitous drug usage from social media,'' *IEEE Transactions on NanoBioscience*, vol. 18, no. 3, pp. 324–334, 2019.

[80] Zhang Min, ''Drugs reviews sentiment analysis using weakly supervised model,'' in *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. IEEE, 2019, pp. 332–336.

[81] Anne Cocos, Alexander G Fiks, and Aaron J Masino, ''Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts,'' *Journal of the American Medical Informatics Association*, vol. 24, no. 4, pp. 813–821, 2017.

[82] Theres Bemila, Isha Kadam, Anushka Sidana, and Shivani Zemse, ''An approach to sentimental analysis of drug reviews using rnn-bilstm model,'' in *Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST)*, 2020.

[83] Felix Graber, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder., ''Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning,'' Accessed 10/20/2020.

[84] Pimwadee Chaovalit and Lina Zhou, ''Movie review mining: A comparison between supervised and unsupervised classification approaches,'' in *Proceedings of the 38th annual Hawaii international conference on system sciences*. IEEE, 2005, pp. 112c–112c.

[85] S Loria, P Keen, and M Honnibal, ''Textblob: simplified text processing. secondary textblob: simplified text processing,'' 2014.

[86] Vipul Kumar Chauhan, Ashish Bansal, and Dr Amita Goel, ''Twitter sentiment analysis using vader,'' *International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT)*, vol. 4, no. 1, pp. 485–489, 2018.

[87] Chen Li, Runqing Song, Maria Liakata, Andreas Vlachos, Stephanie Seneff, and Xiangrong Zhang, ''Using word embedding for bio-event extraction,'' in *Proceedings of BioNLP 15*, 2015, pp. 121–126.

[88] Abdelhafid Zeroual, Fouzi Harrou, Abdelkader Dairi, and Ying Sun, ''Deep learning methods for forecasting covid-19 time-series data: A comparative study,'' *Chaos, Solitons & Fractals*, vol. 140, pp. 110121, 2020.

[89] Manoj Sethi, Sarthak Pandey, Prashant Trar, and Prateek Soni, ''Sentiment identification in covid-19 specific tweets,'' in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020, pp. 509–516.

[90] Xingyou Wang, Weijie Jiang, and Zhiyong Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," in *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 2016, pp. 2428–2437.

[91] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush, "Character-aware neural language models," *arXiv preprint arXiv:1508.06615*, 2015.

[92] Abdalraouf Hassan and Ausif Mahmood, "Convolutional recurrent deep learning model for sentence classification," *Ieee Access*, vol. 6, pp. 13949–13957, 2018.

[93] Mohd Usama, Wenjing Xiao, Belal Ahmad, Jiafu Wan, Mohammad Mehedi Hassan, and Abdulhameed Alelaiwi, "Deep learning based weighted feature fusion approach for sentiment analysis," *IEEE Access*, vol. 7, pp. 140252–140260, 2019.

[94] Mohd Usama, Belal Ahmad, Enmin Song, M Shamim Hossain, Mubarak Alrashoud, and Ghulam Muhammad, "Attention-based sentiment analysis using convolutional and recurrent neural network," *Future Generation Computer Systems*, vol. 113, pp. 571–578, 2020.

[95] Anna Kruspe, Matthias Häberle, Iona Kuhn, and Xiao Xiang Zhu, "Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic," *arXiv preprint arXiv:2008.12172*, 2020.

[96] László Nemes and Attila Kiss, "Social media sentiment analysis based on covid-19," *Journal of Information and Telecommunication*, pp. 1–15, 2020.

[97] Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien, "Sentiment analysis of covid-19 tweets by deep learning classifiersa study to show how popularity is affecting accuracy in social media," *Applied Soft Computing*, vol. 97, pp. 106754, 2020.

[98] Martin Müller, Marcel Salathé, and Per E Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," *arXiv preprint arXiv:2005.07503*, 2020.

[99] Ekasari Nugraheni, Purnomo Husnul Khotimah, Andria Arisal, Andri Fachrur Rozie, Dianadewi Riswantini, and Ayu Purwarianti, "Classifying aggravation status of covid-19 event from short-text using cnn," in *2020 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*. IEEE, 2020, pp. 240–245.

[100] Mudasir Ahmad Wani, Nancy Agarwal, and Patrick Bours, "Impact of unreliable content on social media users during covid-19 and stance detection system," *Electronics*, vol. 10, no. 1, pp. 5, 2021.

[101] Akib Mohi Ud Din Khanday, Syed Tanzeel Rabani, Qamar Rayees Khan, Nusrat Rouf, and Masarat Mohi Ud Din, "Machine learning based approaches for detecting covid-19 using clinical text data," *International Journal of Information Technology*, vol. 12, no. 3, pp. 731–739, 2020.

[102] Jim Samuel, GG Ali, Md Rahman, Ek Esawi, Yana Samuel, et al., "Covid-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, pp. 314, 2020.

[103] Xiaoling Xiang, Xuan Lu, Alex Halavanau, Jia Xue, Yihang Sun, Patrick Ho Lam Lai, and Zhenke Wu, "Modern senicide in the face of a pandemic: an examination of public discourse and sentiment about older adults and covid-19 using machine learning," *The Journals of Gerontology: Series B*, 2020.

[104] Charu C Aggarwal and ChengXiang Zhai, "A survey of text classification algorithms," in *Mining text data*, pp. 163–222. Springer, 2012.

[105] Vipul Kumar Chauhan, Ashish Bansal, and Dr Amita Goel, "Twitter sentiment analysis using vader," *International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT)*, vol. 4, no. 1, pp. 485–489, 2018.

[106] Rabindra Lamsal, "Coronavirus(covid-19)tweets dataset," 2020.

[107] Pierre Boisard, *Camembert, mythe français (Le)*, Odile Jacob, 2007.